A to Z
Revista Eletrônica

JUL/DEZ 2014
VOLUME 03 | NÚMERO 02

NOVAS PRÁTICAS EM

INFORMAÇÃO E CONHECIMENTO

Intelectus Ágil (capa: Priscila Piccolo Pagnoncelli)

UFPR
PKP PUBLIC KNOWLEDGE PROJECT
Gestão da Informação UFPR
ISSN: 2237-826X

AtoZ: novas práticas em informação e conhecimento
www.atoz.ufpr.br

Universidade Federal do Paraná
Setor de Ciências Sociais Aplicadas
Curso de Gestão da Informação
Av. Prefeito Lothário Meissner, 632 - Campus III
Jardim Botânico
Curitiba – PR, Brasil
80210-170
Fone: +55(41)3360-4389
Fax: +55(41)3336-4471
E-mail: revistaatoz@ufpr.br
URL: http://www.atoz.ufpr.br

Periodicidade: Semestral
ISSN: 2237-826X
Qualis/Capes: B4 - Interdisciplinar / B5 - Ciências Sociais Aplicadas I / B4 - Engenharias III
Indexada/registrada em: Directory of Open Access Journals (DOAJ); Sumários.org; Google Acadêmico; LivRe! Portal para periódicos de livre acesso na Internet; InfoBCI; Latindex Catálogo
Diretrizes para autores: http://www.atoz.ufpr.br/index.php/atoz/about/submissions#authorGuidelines

AtoZ
NOVAS PRÁTICAS EM INFORMAÇÃO E CONHECIMENTO

AtoZ : Novas Práticas em Informação e Conhecimento. — Vol. 3, n. 2 (jul./dez. 2014)- . — Curitiba : Universidade Federal do Paraná, Curso de Gestão da Informação, 2014- .
v.

Semestral.
Publicação online: <http://www.atoz.ufpr.br>
ISSN 2237-826X

1. Comunicação científica – Periódico. 2. Informação – Periódico. 3. Conhecimento – Periódico.
I. Curso de Gestão da Informação. II. Universidade Federal do Paraná.

CDD 001(8162)

"Onde está o conhecimento que perdemos na informação?"[1] Explorando a mineração de dados (e de textos) e metodologias de design como apoio à gestão da informação

"Where is the knowledge we have lost in information?"[1] Exploring data mining (and texts) and design methodologies to support information management

Patricia Zeni Marchiori[1]

[1] *Universidade Federal do Paraná (UFPR), Curitiba, Paraná, Brasil*

Autor para correspondência/*Corresponding author*: Patricia Zeni Marchiori [pzeni@ufpr.br]

A tarefa de coleta, análise e busca de significados derivados de dados e de ações de usuários frente a distintas tecnologias são a tônica deste v3n2 da "**AtoZ: novas práticas em informação e conhecimento**". A produção e disponibilidade de volumes crescentes de dados provenientes de distintas fontes estimulam um novo arcabouço de soluções, enfoques revisitados de metodologias de investigação já consolidadas, e possibilidades concretas de intervenção em realidades educativas e sociais; as quais temos o prazer de publicar.

No contexto da mineração de dados, esta edição apresenta entrevista com especialistas ligados e formados pela Pontifícia Universidade Católica do Paraná (PUC-PR), que explicam conceitos centrais e periféricos relacionados à mineração de textos, sua aplicação em *business intelligence*, as condições para o aproveitamento das técnicas e algoritmos, entre outros destaques.

Corroborando um dos aspectos levantados pelos entrevistados, Costa, Bernardino e Viterbo – ao analisarem os boletins de ocorrências da Polícia Rodoviária Federal – destacam a necessidade de se melhorar a qualidade dos dados já na etapa de coleta/alimentação da base, como o desafio para os gestores de repositórios estruturados e não estruturados de dados. Este requisito é essencial para o pré-processamento e o emprego de distintos algoritmos em busca de associações confiáveis.

Quanto à mineração de textos,Trucolo e Digiampietri exploram a extração automática de termos presentes na produção científica de pesquisadores, aliadas a técnicas de análise de tendências, como potenciais recursos na averiguação/crescimento de assuntos ou áreas de pesquisa (ainda) não influentes/prevalentes. Novos estudos que utilizem a metodologia proposta pelos autores podem, futuramente, validar e ajustar tais estratégias de análise, ampliando os recursos de investigação disponíveis para estudos quantitativos em informação e conhecimento.

O foco nos usuários de tecnologia e informação – mais especialmente em jogos/*games* – igualmente tem se fortalecido como um direcionamento de pesquisa na área de informação, conhecimento e tecnologia. Japppur, Forcellini e Spanhol exploram um modelo conceitual para jogos educativos digitais derivado do *Design Science Research Methodology*, o qual foi testado – com resultados positivos – em duas turmas do Programa Jovem Aprendiz do SENAC/SC. Ainda na linha de jogos educativos digitais, o *Object-Oriented Hypermedia Design Method* (OOHDM) e técnicas de inteligência artificial foram utilizados por Villacis e colegas tanto na concepção conceitual como na aplicação em interfaces interativas do "jogo da velha", o qual foi disponibilizado para crianças entre 7 e 11 anos. O estudo comprovou a eficácia do uso das interfaces como estímulo cognitivo nesta faixa etária.

A participação ativa de usuários de sistemas de informação, especialmente aqueles com acesso a plataformas móveis (e estimulados a participar de soluções), demonstra uma tendência de pesquisa cujo foco extrapola a participação pontual do usuário e investe no seu efetivo engajamento. Neste particular, o estudo apresenta-

[1] T.S. Eliot - Chorus I - "The Rock" (1934). Disponível em: http://www3.dbu.edu/mitchell/eliotroc.htm.

do por Lima, Alves, Costa e Sales utiliza a metodologia *Design Thinking* aplicada a uma proposta de solução de *software* voltada à mobilidade urbana; enquanto Lopez-Faicán e Chambas-Eras – com o objetivo oferecer alternativas à intervenção de professores em atividades de ensino – implementam e testam um modelo de incerteza (usando Redes Bayesianas e o modelo Felder-Silverman) voltado à previsão de estilos de aprendizagem em AVAs/Moodle.

Contudo, os autores são cautelosos e não se furtam – como reza a boa prática científica – a indicar limites às suas investigações, considerando-os como necessários para a crítica e revisão de práticas em suas áreas de interesse. Para eles, questionar "o que existe por detrás dos dados" e os contextos sociais, econômicos, e de macro (ou micro) políticas que afetam a criação de conhecimento são decorrência natural dos estudos quantitativos que envolvem – ou buscam explicitar – aspectos das relações sociais, nem sempre revelados pela informação registrada.

Obs: Apresentamos neste número a nova proposta de diagramação pdf, em atendimento às opiniões de nossos leitores e colaboradores quando da pesquisa de avaliação. Outras novidades estão a caminho para os números de 2015! Agradecemos o constante apoio e desejamos uma profícua leitura!

Profa. Dra. Patricia Zeni Marchiori
Editora-Chefe

ISSN: 2237-826X
3(2), jul./dez. 2014

Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas

Data Mining: Applications, tools, learning types and other subtopics

Deborah Ribeiro Carvalho[1], Marcelo Rosano Dallagassa[2]

[1] *Pontifícia Universidade Católica do Paraná (PUC-PR), Curitiba, PR, Brasil*
[2] *Unimed Paraná, Curitiba, PR, Brasil*

Correspondência para/*Correspondence to*: Deborah Ribeiro Carvalho [ribeiro.carvalho@pucpr.br]

Deborah Ribeiro Carvalho possui graduação em Processamento de Dados pela Universidade Federal do Paraná – UFPR (1979), mestrado em Informática Aplicada pela Pontifícia Universidade Católica do Paraná – PUC-PR (1999), doutorado em Informática Aplicada pela Pontifícia Universidade Católica do Paraná (2002) e doutorado em Computação de Alto Desempenho pela Universidade Federal Do Rio Janeiro (COPPE) (2005). Professor da Pontificia Universidade Católica do Paraná, Programa de Pós-Graduação em Tecnologia Aplicada em Saúde e Professor Colaborador do Mestrado Em Gestao da Informacao (UFPR). Tem experiência na área de Ciência da Computação, atuando principalmente nos seguintes temas: mineração de dados, aquisição de conhecimento, apoio a decisão, pós-processamento dos padrões descobertos, na Saúde.

Marcelo Rosano Dallagassa possui graduação em Engenharia Civil pela Universidade Federal do Paraná – UFPR (1988) e Mestrado em Tecnologia em Saúde pela Pontifícia Universidade Católica do Paraná – PUC-PR (2009). Desde 2001 atua como Analista de Negócios e Especialista da UNIMED Paraná, participando nos projetos; Data Warehouse, Portal BI Unimed PR e Informações Estratégicas. Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Apoio a Decisão, atuando principalmente nos temas: Descoberta do Conhecimento em Base de Dados, Data Warehouse, Avaliação de Tecnologias em Saúde, RES – Registro Eletrônico de Saúde e Mineração de Dados. Atua também com Professor de Pós-Graduação nas disciplinas Banco de Dados, Data Warehouse e Data Mining, nas seguintes instituições: FAE, PUC-PR, Universidade Positivo e FESP.

Resumo

Especialistas na área de mineração de dados apresentam conceitos, características, limites e potencialidades da mineração de dados, incluindo indicação de ferramentas disponíveis, relações com a inteligência artificial, e implicações de seu uso na área de business intelligence.

Palavras-chave: Mineração de dados. Ferramentas para mineração de dados. Mineração de dados - uso.

Abstract

Experts in the field of data mining present concepts, features, limitations and possibilities of the data mining process, including the indication of tools available, links to artificial intelligence, and the implications of it's use in business intelligence.

***Keywords**: Data mining. Data mining tools. Data mining use.*

1. Qual a aplicabilidade da mineração de dados na construção do conhecimento organizacional, em especial na área de *business intelligence*?

As organizações, de maneira em geral, acumulam grande volume de dados e informações, em distintos formatos e estrutura, que são constantemente demandados. Entre as estratégias de utilização mais frequentes estão a extração de informação e a de conhecimento. Vale destacar que estas duas formas se complementam, dependendo da situação de contexto do problema de gestão em questão. Se o contexto permitir o estabelecimento prévio de premissas que orientem a extração de informações e estas forem suficientes ao gestor, a mineração de dados teria pouco a contribuir, aliado ao fato de não justificar o custo despendido. Por outro lado, se as estratégias baseadas em premissas não atenderam plenamente às expectativas, a mineração de dados passa a ser uma alternativa a ser considerada.

Vale destacar que algumas vezes a análise das informações geradas a partir de premissas, pode demandar muito tempo e recursos. Este também pode ser uma situação na qual a mineração de dados pode ser considerada para agilizar esse processo.

A mineração de dados deve ser adotada para tornar mais eficiente o apoio à tomada de decisão, elemento essencial para o conceito de *business intelligence*.

São inúmeras as aplicações de mineração de dados utilizadas na área de *business intelligence*, as que identificam perfis e características de clientes conforme as ofertas de produtos, alertas de fraudes, agrupamento de regiões conforme características de vendas, associações de produtos e serviços vinculados aos hábitos de consumo, entre outras.

2. Quais as condições de uso da mineração de dados em organizações de pequeno porte?

Não difere muito das condições de uso da mineração de dados em empresas de grande porte. Em ambos os grupos de organizações é necessário modelar e popular as bases de dados, bem como integrá-las. Ou seja, devem ser garantidas condições de coleta, consistência e armazenamento dos dados para posterior extração de informações. Para estas atividades são necessários profissionais capacitados que façam parte do corpo de colaboradores ou que sejam contratados por demanda específica.

Agregar atividades de mineração de dados vai também exigir que os dados estejam disponíveis e de profissionais que dominem as respectivas técnicas. Profissionais estes que podem fazer parte da equipe funcional ou serem contratados especificamente para esta atividade.

Ou seja, as exigências são muito similares, independentemente do porte da empresa.

3. Como está a disponibilidade de ferramentas de mineração de dados (alternativas em *software* livre/gratuito e alternativas pagas)?

A partir da intensificação recente de pesquisas na área de desenvolvimento de algoritmos, há uma grande oferta de ferramentas para mineração de dados em ambiente livre e gratuita, com código fonte aberto (*General Public Licence*), entre elas: Weka (Witten & Frank, 2000), Mahout (Ingersoll, 2009), Orange data mining (Demsar, Zupan, Leban, & Curk, 2004), Rapid Miner (Hofmann & Klinkenberg, 2013), Tanagra (Rakotomalala, 2005), Keel (Alcala-Fdez et al., 2011) etc.

Em termos de aplicações proprietárias, existem várias soluções e algoritmos de *data mining* incorporadas em ferramentas de *business intelligence*, como por exemplo: Oracle (Tamayo, Berger, Campos, & Yarmus, 2005), Microsoft (Seidman, 2001), SAS (Fernandez, 2003), entre outras.

Porém um dos principais desafios ainda consiste em identificar qual estratégia algorítmica melhor se adequa ao contexto, problema e questão.

4. Como se pode saber se uma base de dados é adequada para uma mineração de dados?

Uma base de dados é adequada para a mineração de dados, a partir da respectiva preparação, conforme as etapas previstas no processo de KDD (*Knowledge Discovery in Database*) proposto por Fayyad, Piatetsky Shapiro, Smyth, e Uthurusam (1996), entre elas: seleção e integração dos dados em um repositório único padronizado, remoção de ruídos e *outliers* etc.

Da mesma forma que para qualquer processamento estatístico, o conjunto de dados deve representar, de forma confiável, o universo (mundo real) a ser investigado, possibilitando assim inferir a situação problema como um todo, seja pela perspectiva de completeza ou complexidade do problema. Um dos "mitos" criados, a partir da motivação inicial das discussões sobre a mineração de dados, era ser uma alternativa para "grandes" bases de dados. Este fato decorreu da própria dificuldade de processamento inerente à descoberta e identificação de informações oportunas ao processo decisório em grandes conjuntos. Mas, dispor de um grande conjunto não constitui requisito obrigatório desde que este represente o universo, por amostragem ou não.

Quando a mineração de dados resultar na descoberta de elementos potencialmente úteis para o apoio a tomada de decisão, pode-se afirmar que a base de dados atendeu às expectativas.

5. O que são "ruídos" nas bases de dados e como eles podem influenciar no resultado obtido em um processo de mineração de dados?

"Ruído" representa conteúdo nas bases de dados que pode prejudicar a qualidade da informação extraída, a partir de qualquer método, seja ele tradicional ou baseado em estratégias mais elaboradas. Destacam-se como ruídos: valores fora do domínio, ausência de valores, inconsistências etc. É importante lembrar que o mundo real é ruidoso, ou seja, se uma base de dados representa uma abstração deste mundo real, esta será ruidosa a despeito dos esforços despendidos para a sua modelagem e respectiva população. Cabe aos profissionais da área de tecnologia da informação minimizar o impacto negativo que estes ruídos possam representar nas informações extraídas e disponibilizadas aos gestores.

Por exemplo, todas as vezes que, ao informar os dados cadastrais se omite ou não se informa corretamente a renda, gera-se um ruído no conjunto de dados.

6. Quão próximos estão os estudos de mineração de dados e aprendizagem de máquina?

O processo KDD como um todo compreende uma série de áreas que se relacionam de forma multidisciplinar, a saber: estatística, banco de dados, inteligência artificial, aprendizado de máquina etc. Ou seja, o processo KDD é construído a partir de conceitos destas diversas áreas. Aprendizagem de máquina é a automação de um processo de aprendizagem.

7. O que é aprendizado preditivo e descritivo? Uma mesma tarefa de mineração de dados pode ser dos dois tipos ao mesmo tempo?

O Aprendizado envolve generalização a partir da experiência e no caso de Aprendizado de Máquina, este busca generalizar a experiência retratada em conjuntos de dados. O aprendizado descritivo analisa os dados e identifica similaridades (agrupamentos) ou associações (regras de associações).

O aprendizado preditivo analisa os dados que representam eventos passados buscando relações entre estes que permitam predizer situações em novos dados futuros, tais como: a classificação para predições de valores discretos e a regressão para predições de valores contínuos.

Como exemplo de aprendizado descritivo, pode-se citar a associação entre eventos demandados por pacientes (de hospitais), que podem indicar possíveis relações de causa-efeito ao ser complementada com a respectivo espaço de tempo entre estes eventos associados.

Como exemplo de aprendizado preditivo, pode-se identificar potenciais clientes interessados em algum novo produto a ser divulgado.

Em geral, a natureza do problema em questão define a natureza do aprendizado a ser adotado na experimentação de mineração de dados, ou seja, em geral os problemas são preditivos ou descritivos. Porém muitas vezes as duas estratégias podem ser adotadas de forma complementar. Por exemplo, o sistema aprende sobre as preferências históricas de compras e avaliações e a partir deste aprendizado realiza recomendações de novos produtos. O aprendizado para esta recomendação pode ser obtido pela hibridização destes dois tipos de aprendizados.

8. Quais as relações, se existentes, entre mineração de dados e inteligência artificial?

A inteligência artificial discute estratégias para emular o comportamento humano, da natureza etc., como estratégia de solução para problemas computacionais. Neste sentido, se a mineração de dados busca descobrir generalizações nos dados, como uma instancia de aprendizado, a inteligência artificial apoia na pesquisa e modelagem destas estratégias para emulação. Ou seja, a inteligência artificial contribui na modelagem do "coração" dos algoritmos que descobrem padrões a partir dos dados.

Como exemplo de aplicações de inteligência artificial que estejam diretamente relacionadas com a mineração de dados, pode-se citar o desenvolvimento da robótica móvel para apoiar o processo produtivo das indústrias.

E, como exemplo da mineração de dados, cita-se a possibilidade de avaliar se um futuro cliente é um potencial adimplente ou inadimplente.

9. Como podemos identificar, e justificar, a necessidade de um "cientista de dados" no mercado de trabalho com informação?

A partir da quantidade de dados e informações, nas mais diversas áreas do conhecimento, gestores e profissionais envolvidos em tomada de decisão demandam por novas possibilidades de otimizar este processo. A integração de dados oriundos de diversas fontes, como redes sociais e bases de dados próprias ou mesmo públicas, possibilita compreender melhor todas as variáveis envolvidas.

Nesse contexto, profissionais com competência para explorar estes dados e informações, desenvolver modelos matemáticos, identificar novas oportunidades que melhor atendem às necessidades de gestão do negócio, são essenciais.

Este profissional pode ser o cientista de dados, o qual trabalha em três espaços: área do problema, o da matemática e o de tecnologia da informação. Sua função é transformar os dados disponíveis em elementos de apoio a decisão.

REFERÊNCIAS

Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac J., Garcia, S., Sanchez, S., & Herrera F. (2011). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. of Mult.-Valued Logic & Soft Computing*, 17, 255–287. Retirado de http://sci2s.ugr.es/publications/ficheros/2010-JMVLSC-Alcala_Fdez-KEEL-dataset.pdf

Demsar, J., Zupan, B., Leban, G., & Curk, T. (2004). Orange: From experimental machine learning to interactive data mining. *8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 537-539. doi: 10.1007/978-3-540-30116-5_58

Fayyad, U. M., Piatetsky Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. California, USA: AAAI, MIT.

Fernandez, G. (2003). *Data mining using SAS application*. London: Chapman & Hall.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data mining use cases and business analytics applications*. Retirado de https://books.google.com/books?isbn=1482205491

Ingersoll, G. (2009). *Introducing Apache Mahout Scalable, commercial-friendly machine learning for building intelligent applications*. Retirado de http://www.ibm.com/developerworks/java/library/j-mahout/j-mahout-pdf.pdf

Rakotomalala, R. (2005). TANAGRA: a free software for research and academic purposes. *Proceedings of EGC RNTI-E-3*, 2th, 697-702. Retirado de http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

Seidman, C. (2001). *Data mining with Microsoft SQL Server 2000 technical reference*. Redmond: Microsoft.

Tamayo, P., Berger, C., Campos, M., Yarmus, J., Milenova, B., Mozes, A., ... , & Myczkowski, J. (2005). Oracle data mining. In Maimon, O., & Rokach, L. (Eds.). *Data Mining and Knowledge Discovery Handbook* (1315-1329). New York: Springer. doi: 10.1007/0-387-25465-X_63

Witten I. H., & Frank E. (2000). *Machine learning algorithms in Java*. Retirado de http://www.cs.waikato.ac.nz/ml/weka/

Como citar esta entrevista (ABNT):

CARVALHO, D. R.; DALLAGASSA, M. R. Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 82-86, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this interview* (APA):**

Carvalho, D. R., & Dallagassa, M. R. (2014). Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 82-86 Retrieved from http://www.atoz.ufpr.br

ISSN: 2237-826X
3(2), jul./dez. 2014

Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos

Trend analysis of the Brazilian scientific production in information science area: exploratory study of text mining

Caio Cesar Trucolo[1], Luciano Antonio Digiampietri[1]

[1] *Universidade de São Paulo (USP), São Paulo, SP, Brasil*

Autor para correspondência/*Corresponding author*: Caio Cesar Trucolo [trucolo@gmail.com]

Financiamento/*Funding*: O trabalho apresentado neste artigo foi parcialmente financiado pela CAPES (bolsa de mestrado) e pelo CNPq (Projeto Universal e bolsa de produtividade em pesquisa).

Recebido/*Submitted*: 17 Out. 2014
Aceito/*Approved*: 15 Nov. 2014

Resumo

Introdução: A análise de tendências pode ser utilizada como uma estratégia para identificar assuntos ou áreas de pesquisa com potencial de popularidade mas que ainda não são muito disseminados. Este trabalho consiste em identificar tendências por mineração de texto e análise histórica das produções científicas (artigos científicos) de doutores da área de Ciência da Informação.
Método: De natureza exploratória, este trabalho foi construído em três etapas. A primeira etapa foi a da obtenção dos dados dos currículos cadastrados na plataforma Lattes. A segunda etapa consistiu na extração automática dos termos mais importantes inseridos nos títulos das publicações e, na terceira etapa, foram feitas regressões lineares e não lineares dos índices de importância baseados em frequência dos termos extraídos.
Resultados: Informações gerais sobre as tendências identificadas para a área de Ciência de Informação para curto, médio e longo prazo são apresentadas.
Conclusão: Este trabalho apresenta e aplica uma metodologia de identificação de tendências que ainda pode ser considerada um primeiro passo ante ao potencial da análise de tendências para a produção científica nacional. Além disso, informações gerais sobre as tendências identificadas e os comportamentos dessas tendências ao longo do tempo foram discutidas.

Palavras-chave: Análise de tendências. Ciência da informação. Redes sociais.

Abstract

***Introduction**: Trend analysis can be used as a strategy to identify subjects or research areas with potential of popularity which are not very widespread. This work consists of trend identification by text mining and historic analysis of the scientific productions (scientific papers) of the Information Science area PhDs.*
***Method**: This work, having an exploratory basis, was built in three steps. The first step was the data gathering of the curricula registered in Lattes platform. The second one consisted of automatic extraction of the most important terms inside the publications titles and, in the third step, linear and non linear regression of the frequency based importance index of the extracted terms were executed.*
***Results**: Identified trends from the Information Science area for short, medium and long time were presented.*
***Conclusion**: This work presents and applies a trend identification method that can be seen as a first step considering all the potential of the national scientific production trend analysis. Moreover, trend analysis general information and the trends behavior over time were discussed.*

***Keywords**: Trend analysis. Information science. Social networks.*

INTRODUÇÃO

Estratégias e políticas públicas têm sido inseridas no país para melhorar a qualidade e aumentar a produtividade da pesquisa científica. Muitas vezes essas políticas são escolhidas de acordo com áreas de pesquisa já consolidadas e populares, nas quais se sabe que haverá retorno, ou ainda, identificadas como tendências mundiais. Um país com dimensões continentais como o Brasil – tanto em extensão geográfica quanto em diversidade cultural – poderia investir em áreas e temas com potencial de crescimento, ampliando o potencial de retorno da investigação científica.

A produção científica no Brasil vem crescendo exponencialmente nas últimas décadas (Digiampietri et al., 2012a) o que só faz crescer o interesse em entender as características da pesquisa no País. Tal análise pode beneficiar da utilização da mineração de texto em tais produções com o objetivo de tentar identificar áreas e temas de pesquisa nas quais os pesquisadores de determinada área trabalham (Miyata, Kano, & Digiampietri, 2013). Assim, analisar tendências a partir das produções científicas para áreas específicas se configura como uma estratégia para encontrar temas de pesquisa com potencial de impacto (Trucolo & Digiampietri, 2014b).

O Brasil possui uma base de dados ímpar de cadastro de currículos de pesquisadores chamada Plataforma Lattes. Nessa base existem mais de três milhões de currículos cadastrados com informações importantes para análise de pesquisadores e redes acadêmicas.

Este artigo consiste em um trabalho de metaestudo para a área de Ciência da Informação com o intuito de identificar tendências de termos de pesquisa para curto, médio e longo prazo. O objetivo é desenvolver e aplicar uma metodologia para identificar tendências de assuntos e ramos de pesquisa a partir das informações dos currículos cadastrados na Plataforma Lattes dos doutores que atuam em Ciência da Informação. É de interesse deste trabalho explorar a metodologia de forma a identificar e prever tendências a partir de análises históricas.

O restante deste artigo está organizado da seguinte forma: a seção 2 sumariza os trabalhos correlatos, a seção 3 apresenta a plataforma Lattes e sua importância informativa sobre a produção científica nacional, a metodologia é descrita na seção 4, na seção 5 os resultados são apresentados e, a seção 6, que contém as considerações finais.

Trabalhos correlatos

Análises de tendência vêm sendo estudadas ao longo dos últimos anos e diversas são as áreas de aplicação. Os trabalhos que mais se aproximam desta proposta são aqueles que assumem como método a mineração de texto e, como aplicação, documentos textuais históricos.

No trabalho de Bolelli, Ertekin, Zhou e Giles (2009), o *Lattent Dirichilet Allocation*, que é um modelo generativo probabilístico e a *Gibbs Sampling*, algoritmo gerador de amostras que se aproximam de distribuições de probabilidade específicas, foram utilizados conjuntamente com a ordem temporal dos documentos para a criação de um modelo generativo que aprende as distribuições de autor, tópico e palavra. Em uma aplicação sintética, houve uma taxa de acerto de aproximadamente 72%.

O trabalho de Kawamae e Higashinaka (2010) consistiu em tentar predizer as distribuições dos tópicos em artigos científicos levando o tempo em consideração. Com a mesma ideia, e em um novo trabalho, Kawamae (2012) estabeleceu uma diferença entre tópicos estáveis (que não possuem variação significativa ao longo do tempo) e tópicos dinâmicos, tentando rebater outros modelos que apenas levam em consideração as explosões de tópicos (aumentos súbitos de aparecimento de tópicos em determinados períodos). Para a avaliação, o autor compara o modelo proposto com outros dois modelos utilizando medida de perplexidade e medida de erro L1, que são meios de medir e comparar a acurácia de predição de modelos em relação as amostras utilizadas. Por fim, o modelo apresenta a medida de perplexidade e a taxa de erro L1 menores que a dos outros dois modelos, sendo 2,44 a média de L1.

Jayashri e Chitra (2012) propuseram um modelo com rede *Adaptative Resonance Theory* (ART), que é um tipo de rede neural que propõem resolver o problema de "esquecimento" de aprendizado conforme a apresentação de novas informações, para identificar tópicos em documentos científicos de diferentes bases e detectar tendências considerando os picos de frequência desses tópicos extraídos. A abordagem utilizada foi capaz de detectar os tópicos em alta para diferentes bases de dados.

Park et al. (2011) utilizaram uma abordagem de detecção de tendências usando seleção de características baseada em IG-I (*Improved Gini – Index*). Para cada tópico dado como entrada, subtópicos foram extraídos para então se analisar o comportamento temporal de cada subtópico e identificá-los como em alta ou em baixa. Para os quatro tópicos dados como entrada para o modelo, foram realizados testes de medida F1 para avaliação dos subtópicos com SVM (*Support Vector Machine*), técnica supervisionada de inteligência artificial, e kNN (*k – nearest neighbours*), técnica não supervisionada de inteligência artificial. O resultado de F1 para SVM foi de 0,982 enquanto para kNN foi de 0,916. Uma limitação do método utilizado é que o processo não é totalmente automatizado.

Abe e Tsumoto (2009) selecionaram termos de importância por TF-IDF (*Term Frequency – Inverse Document Frequency*), que é um índice que mede a importância de determinada palavra para um documento dentro de

uma coleção de documentos, e coeficiente de Jaccard, coeficiente que mede a similaridade entre conjuntos de amostras, utilizando regressão linear *a posteriori* para detectar tendências emergentes. Todas as tendências detectadas foram confirmadas como tendências reais por especialistas dos domínios.

Trucolo e Digiampietri (2014a) analisaram tendências de termos extraídos automaticamente a partir de métodos de regressões lineares e não lineares para a rede de doutores inseridos em programas de pós graduação *strictu sensu* avaliados pela CAPES em Ciências da Computação. Além disso, foi realizada uma análise da rede social baseada em coautorias entre os professores permanentes de cada programa para verificar se existia correlação entre as principais tendências identificadas e as coautorias entre os programas. Nesse trabalho, os autores não conseguiram identificar correlações significativas.

Além dos documentos textuais utilizados como base para a análise de tendências, nos últimos anos as redes sociais começaram a ser também utilizadas para auxiliar nesta análise. Cimenler, Reeves e Skvoretz (2014) analisam algumas métricas de redes sociais, em sua maioria métricas de centralidade, para entender quão significativas são essas métricas para prever o desempenho de pesquisadores baseando-se em índices derivados de citação como o índice h. O método de regressão de Poisson é utilizado para analisar a importância das métricas para alguns tipos de redes formadas por pesquisadores de uma faculdade de engenharia.

Uma revisão sistemática sobre técnicas de identificação e análise de tendências para outras aplicações além de documentos textuais históricos pode ser encontrada em Trucolo e Digiampietri (2014b).

Uma das principais contribuições do presente trabalho é o aprofundamento relevante para com a aplicação porque os dados extraídos da plataforma Lattes proporcionam uma análise bastante rica sobre a condição científica nacional. Pode-se dizer, portanto, que esse trabalho se diferencia dos demais citados pela característica de aproximação da análise à realidade científica do país. Outra característica importante é a capacidade de a identificação e predição de tendências utilizada aqui não necessitar de esforço humano, ou seja, ela é realizada a partir de termos e expressões extraídos automaticamente da base de dados sem a necessidade de manualmente se estabelecer a importância dos termos ou valores limítrofes para as tendências.

MÉTODOS

Todo o processo de análise de tendências foi realizado em três fases: obtenção dos dados; extração automática dos termos; e análise de tendência dos termos extraídos.

Obtenção dos dados

Para a obtenção dos dados, inicialmente foram identificados todos os doutores que atuam na área de Ciência da Informação e que possuem currículos cadastrados na plataforma Lattes. As informações dos currículos destes pesquisadores foram tabuladas e armazenadas em um banco de dados seguindo a metodologia apresentada em Digiampietri et al. (2012a), para a realização dos testes, foram extraídos os termos de 34.289 títulos de artigos publicados entre os anos de 1991 e 2012.

Extração automática de termos

A técnica de extração automática de termos utilizada consiste em determinar os termos mais importantes de um conjunto de documentos pela frequência adjacente das palavras que compõem esses termos. A fórmula (1) utilizada para o cálculo dos pesos de cada termo candidato é a seguinte:

$$FED(TC) = f(TC) \times \left(\prod_{i=1}^{T} (FE(N_i) + 1)\,(FD(N_i) + 1)\right)^{1/T} > 1.0 \quad (1)$$

Em que f(TC) é a frequência do termo candidato TC, e FE(Ni) e FD(Ni) indicam a frequência dos candidatos da esquerda e da direita, respectivamente. Esta fórmula é detalhadamente descrita por Nakagawa e Mori (2002).

Observou-se que os termos compostos tinham mais significado do que os termos simples em relação aos assuntos abordados pelas publicações. Desta forma, os termos utilizados na fase de análise de tendências foram os termos compostos e de maiores pesos.

Análise de tendências

Com base nos termos extraídos, foram calculados os índices de importância dos termos para cada ano. O índice de importância utilizado nesse trabalho é o TF-IDF (*Term Frequency divided by Inverse Document Frequency*), que é um dos índices mais utilizados para aferir a importância de termos. Com esses índices calculados, foram utilizadas análises de regressão do tipo linear e não linear. Análises de regressão seguem o formato (2)

$$Y \approx f(X, \beta) \quad (2)$$

em que a variável dependente Y pode ser aproximada pelas variáveis independentes X e seus respectivos parâmetros β para determinada função *f*(). Nas análises de regressão desse trabalho a variável dependente é o índice TF-IDF do termo e variável independente é o tempo (os anos de publicações dos artigos).

O método de mínimos quadrados foi utilizado para determinar as curvas de tendência que mais se adequavam as séries temporais de cada termo. Os tipos de regressão utilizados foram linear, exponencial, logarítmica, *power law* e polinomial de grau 2 a 5. Posteriormente, foi calculado o erro quadrático para cada curva de tendência gerada para se determinar a curva mais adequada à série temporal de cada termo.

A classificação dos termos como tendências foi baseada na previsão, a partir da curva de tendência mais adequada (isto é, com menor erro quadrático) para alguns anos específicos que indicassem curto, médio e longo prazo. A análise histórica dos termos foi realizada entre os anos de 1991 e 2012, com isso, a análise de curto prazo foi realizada para o ano de 2013, a de médio prazo para o ano de 2015 e a de longo prazo para o ano de 2020. Optou-se por utilizar dados das publicações apenas até o final de 2012 pois a partir de 2013 muitos currículos não estão atualizados (Digiampietri et al., 2014).

RESULTADOS

A partir da extração automática de termos, conforme explicado na seção anterior, foram selecionados três termos entre os mais importantes para exemplificar o comportamento temporal de cada um. O gráfico 1 ilustra esses comportamentos e é possível visualizar comportamentos bem diferentes entre eles. Os termos **ciência da informação** e **gestão do conhecimento** vinham em uma crescente parecida até o ano de 2006, quando se separaram. **Ciência da informação** continuou crescendo até 2009 quando se estabilizou e teve uma queda significativa em 2012. Já **gestão do conhecimento** começou a decrescer em 2006 e se estabilizou a partir de 2008. O termo **meio ambiente**, diferentemente dos outros dois termos, tem um comportamento estável com altos e baixos não muito significativos durante o período.

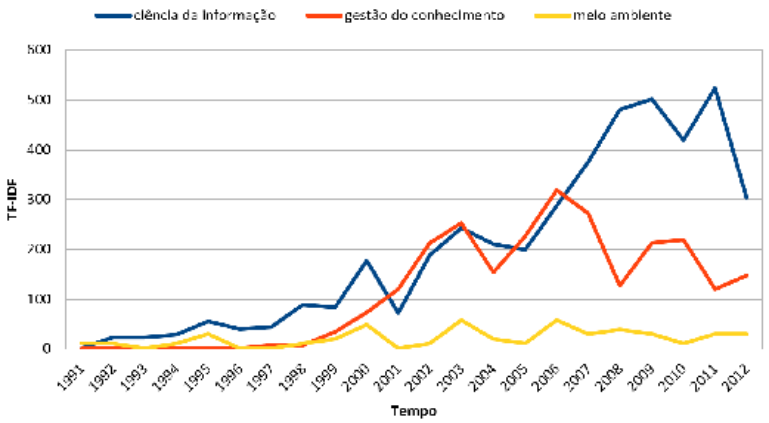


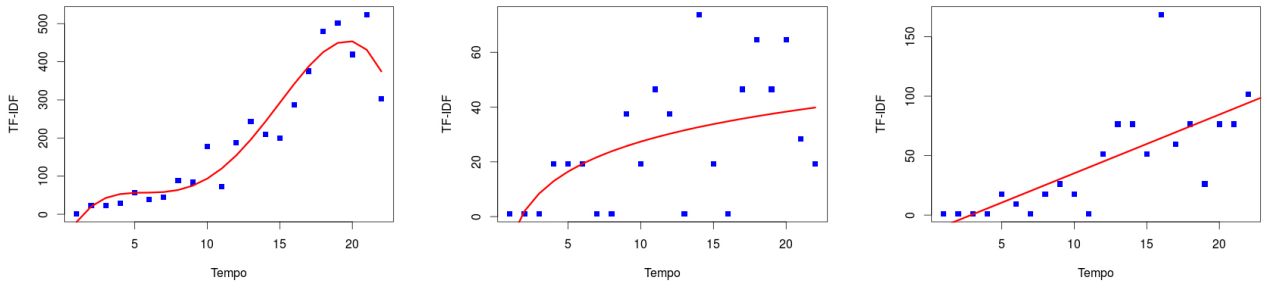
Gráfico 1. Comportamento temporal de três termos
Fonte: os autores.

Como explicado anteriormente, foram calculadas as curvas de tendência das séries temporais de cada termo extraído. As figuras 2, 3 e 4 mostram as curvas de tendência baseadas nas regressões polinomial de grau 4, logarítmica e linear, respectivamente, para os termos **ciência da informação**, **latin america** e **comunicação científica**. A forma algébrica das curvas de tendência são, respectivamente (3)

$$Y = -84{,}76 - 0{,}3 \times X^4 + 1{,}28 \times X^3 - 15{,}89 \times X^2 + 79{,}48 \times X$$
$$Y = -8{,}99 + 15{,}80 \times log(X)$$
$$Y = -14{,}24 + 4{,}94 \times X \quad (3)$$

Pelos gráficos 2, 3 e 4 é possível notar uma tendência de aumento do índice TF-IDF para os termos **latin america** e **comunicação científica** enquanto que **ciência da informação**, que teve um aumento substancial no meio da série temporal, está em uma fase de queda. O comportamento temporal do termo **latin america** é interessante pelo motivo de que mesmo tendo grandes variações entre os anos, a curva de tendência aponta um aumento baixo do índice para os anos seguintes.

[2] [3] [4]



Gráficos 2, 3 e 4. [2] Curva de tendência gerada pela regressão não linear polinomial de grau 4 para os termo ciência da informação. [3] latin america. [4] comunicação científica.
Fonte: os autores.

A tabela 1 apresenta as vinte principais tendências a curto, médio e longo prazo em ordem decrescente. Nesta tabela já é possível notar comportamentos de alguns dos termos. **Information retrieval**, por exemplo, não aparece entre as principais tendências a curto prazo, mas em 2015 já aparece entre as duas principais. Para uma visão melhor desses comportamentos, a tabela 2 mostra o deslocamento de posições para médio e longo prazo entre as vinte principais tendências de 2013. A tabela 2 igualmente confirma a queda do índice TF-IDF para médio e longo prazo do termo **ciência da informação** (verificado na curva de tendência da figura 2). O termo **comunicação científica** apesar de ter uma tendência positiva, como visto na figura 4, teve um deslocamento negativo das posições para médio e longo prazo.

A fim de se avaliar a acurácia do modelo proposto, foram comparados os resultados previstos para os 20 termos com maiores tendências de popularidade e os resultados reais TF-IDF dos mesmos para o ano de 2011. Observou-se um erro padrão médio de aproximadamente 38,6% e um grau de correlação positivo entre os valores previstos e os valores reais em torno de 0,55.

As oscilações de tendências ao longo do tempo se devem aos diferentes modelos de regressão determinados automaticamente, ou seja, o modelo mais adequado baseado no método dos mínimos quadrados. O gráfico 5 exemplifica o comportamento das curvas de regressão para os termos **information retrieval** e **comunicação científica**. Analisando-se o gráfico 5 juntamente com a tabela 2, nota-se que comunicação científica está tendo um crescimento, porém, esse crescimento é baixo em relação a termos como **information retrieval**, que é a principal tendência para 2020.

Tendências em 2013	Tendências em 2015	Tendências em 2020
ciência da informação	sustentabilidade ambiental	information retrieval
santa catarina	information retrieval	sustentabilidade ambiental
comunicação científica	santa catarina	marc21 bibliográfico
minas gerais	marc21 bibliográfico	multiwords expressions
sustentabilidade ambiental	comunicação científica	bioactive glass
comportamento informacional	multiwords expressions	doenças negligenciadas
américa latina	doenças negligenciadas	médio porte
information retrieval	médio porte	direitos autorais
biblioteca universitária	betim paes	buenos aires
marc21 bibliográfico	betim paes leme	diretrizes curriculares
mato grosso	diretrizes curriculares	betim paes
região metropolitana	minas gerais	betim paes leme
doenças negligenciadas	direitos autorais	febre amarela
inteligência competitiva	classificação facetada	classificação facetada
são paulo	febre amarela	suggestion box
multiwords expressions	suggestion box	ivan illich
neural network	ivan illich	claims derived
médio porte	claims derived	crenicichla menezesi
língua portuguesa	crenicichla menezesi	rheumatoid arthritis study
betim paes	rheumatoid arthritis study	rheumatoid arthritis

Tabela 1. Vinte principais tendências para curto, médio e longo prazo
Fonte: os autores.

Posição	2013	2015		2020	
1°	ciência da informação	sustentabilidade ambiental	4	information retrieval	1
2°	santa catarina	information retrieval	6	sustentabilidade ambiental	-1
3°	comunicação científica	santa catarina	-1	marc21 bibliográfico	1
4°	minas gerais	marc21 bibliográfico	6	multiwords expressions	2
5°	sustentabilidade ambiental	comunicação científica	-2	doenças negligenciadas	2
6°	comportamento informacional	multiwords expressions	10	médio porte	2
7°	américa latina	doenças negligenciadas	6	betim paes	2
8°	information retrieval	médio porte	10	inteligência competitiva	8
9°	biblioteca universitária	betim paes	11	santa catarina	-6
10°	marc21 bibliográfico	minas gerais	-6	comunicação científica	-5
11°	mato grosso	américa latina	-4	língua portuguesa	2
12°	região metropolitana	mato grosso	-1	mato grosso	0
13°	doenças negligenciadas	língua portuguesa	6	américa latina	-2
14°	inteligência competitiva	comportamento informacional	-8	região metropolitana	1
15°	são paulo	região metropolitana	-3	minas gerais	-5
16°	multiwords expressions	inteligência competitiva	-2	biblioteca universitária	1
17°	neural network	biblioteca universitária	-8	neural network	1
18°	médio porte	neural network	-1	comportamento informacional	-4
19°	língua portuguesa	ciência da informação	-18	são paulo	1
20°	betim paes	são paulo	-5	ciência da informação	-1

Tabela 2. Alteração das posições a médio e longo prazo das vinte principais tendências de curto prazo
Fonte: os autores.

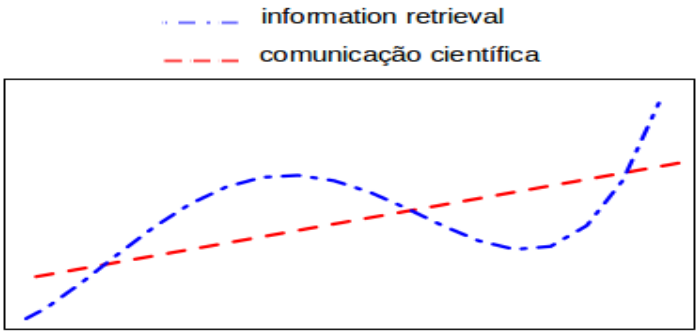


Gráfico 5. Comportamento das curvas de regressão dos termos **information retrieval** e **comunicação científica**
Fonte: os autores.

CONSIDERAÇÕES FINAIS

A aplicação de estratégias e políticas públicas mais acuradas para o aumento da qualidade e produtividade da ciência no país dependem de análises mais profundas em relação a realidade acadêmica brasileira. O Brasil, por ser um país com dimensões continentais tanto em extensão geográfica quanto em diversidade cultural, necessita de estudos específicos para identificar áreas e assuntos com grande potencial de impacto científico e social.

Com caráter informacional, este trabalho apresentou informações gerais sobre as tendências de assuntos e termos para a área de Ciência da Informação. Foram mostradas as principais tendências de curto, médio e longo prazo e o comportamento de alguns desses termos ao longo desse intervalo. Dessa forma, foi possível vislumbrar quais assuntos estarão em alta para curto, médio e longo prazo.

Este trabalho não contempla o agrupamento dos termos extraídos em tópicos para se, então, analisar a tendência desses tópicos mais gerais que podem ser relevantes para a área de Ciência da Informação. Neste trabalho também não foi realizada uma análise de correlação entre as principais tendências entre programas de pós-graduação e as principais coautorias inter programas como feito por Trucolo e Digiampietri (2014a) pelo fato de nem todos os doutores da análise estarem inseridos em programas de pós-graduação.

Os resultados deste trabalho ainda podem ser considerados iniciais, considerando-se todo o potencial da análise de tendências da produção científica nacional. Em trabalhos futuros, a estrutura das fontes de informação, ou seja, as características das redes sociais, serão agregadas. Métricas das redes serão inseridas de forma que auxiliem na explicação do comportamento das séries temporais. Com isso, objetiva-se aumentar o poder de acurácia do modelo de predição de tendências.

REFERÊNCIAS

Abe, H., Tsumoto, S. (2009). Evaluating a method to detect temporal trends of phrases in research documents. *8th IEEE International Conference on Cognitive Informatics*, 378-383. doi:10.1109/COGINF.2009.5250711

Bolelli, L., Ertekin, S., Zhou, D., & Giles, C. L. (2009). Finding topic trends in digital libraries. *9th ACM/IEEE-CS Joint Conference on Digital Libraries*, 69-72. doi:10.1145/1555400.1555411

Cimenler, O., Reeves, K. A., & Skvorets, J. (2014). A regression analysis of researchers' social network metrics on their citation performance in a college of engineering. *Journal of Informetrics*, 8(3), 667-682. doi:10.1016/j.joi.2014.06.004

Digiampietri, L. A., Mena-Chalco, J. P., Pérez-Alcázar, J. J., Tuesta, E. F., Delgado, K. V., Mugnaini, R., & Silva, G. S. (2012a). Dinâmica das relações de coautoria nos programas de pós-graduação em computação no Brasil. *2012 Brazilian Workshop on Social Network Analysis and Mining*.

Digiampietri, L. A., Mena-Chalco, J. P., Pérez-Alcázar, J. J., Tuesta, E. F., Delgado, K. V., Mugnaini, R., & Silva, G. S. (2012b). Minerando e caracterizando dados de currículos Lattes. *2012 Brazilian Workshop on Social Network Analysis and Mining*. Retirado de http://www.imago.ufpr.br/csbc2012/anais_csbc/eventos/brasnam/artigos/BRASNAM%20-%20Minerando%20e%20Caracterizando%20Dados%20de%20Curriculos%20Lattes.pdf

Digiampietri, L., Mugnaini, R., Mena-Chalco, J., Delgado, K., & Pérez-Alcázar, J. (2014). Análise da atualização dos Currículos Lattes. *IV Encontro Brasileiro de Bibliometria e Cientometria*. Retirado de http://www.uspleste.usp.br/digiampietri/bibtex/DigiampietriEtAl_EBBC2014.pdf

Kawamae, N. (2012). Theme chronicle model: chronicle consists of timestamp and topical words over each theme. *21st ACM International Conference on Information and Knowledge Management*, 2065-2069. doi:10.1145/2396761.2398573

Kawamae, N., & Higashinaka, R. (2010). Trend detection model. *19th International Conference on World Wide Web*, 1129-1130. doi:10.1145/1772690.1772838

Jayashri, M., & Chitra, P. (2012). Topic clustering and topic evolution based on temporal parameters. *2012 International Conference on Recent Trends in Information Technology*, 559-564. doi:10.1109/ICRTIT.2012.6206816

Miyata, B. K. O., Kano, V. Y., & Digiampietri, L. A. (2013). Combinando mineração de textos e análise de redes sociais para a identificação das áreas de atuação de pesquisadores. *Second Brazilian Workshop on Social Network Analysis and Mining*. Retirado de https://drive.google.com/viewerng/viewer?a=...

Nakagawa, H., & Mori, T. (2002). A simple but powerful automatic term extraction method. *Second International Workshop on Computational Terminology*. doi:10.3115/1118771.1118778

Park, H., Kim, E., Bae, K., Hahn, H., Sung, T., & Kwon, H. (2011). Detection and analysis of trend topics for global scientific literature using feature selection based on Gini-Index. *23rd IEEE International Conference on Tools with Artificial Intelligence*, 965–969. doi:10.1109/ICTAI.2011.166

Trucolo, C. C., & Digiampietri, L. A. (2014a). Análise de tendências da produção científica nacional da área de Ciência da Computação. *Revista de Sistemas de Informação da FSMA*, 14, 2-9. Retirado de http://www.fsma.edu.br/si/edicao14/FSMA_SI_2014_2_Estudantil_1.pdf

Trucolo, C. C., & Digiampietri, L. A. (2014b). Uma revisão sistemática acerca das técnicas de identificação de análise de tendências. *X Simpósio Brasileiro de Sistemas de Informação*, 639-650. Retirado de http://www.uspleste.usp.br/digiampietri/bibtex/TrucoloEDigiampietri2014a.pdf

Como citar este artigo (ABNT):

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 87-94, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this article* (APA):**

Trucolo, C. C., & Digiampietri, L. A. (2014). Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 87-94. Retrieved from http://www.atoz.ufpr.br

ISSN: 2237-826X
3(2), jul./dez. 2014

Optimización del juego tres en raya con niveles de dificultad utilizando heurísticas de inteligencia artificial

Optimizing the stages of difficulty of the Tic-Tac-Toe game using artificial intelligence heuristics

César Javier Villacís[1], Walter Marcelo Fuertes[1], Christian Andrés Bustamante[2], Margarita Elizabeth Zambrano[1], Edgar Porfirio Torres[1], Hernán Mauricio Aules[1], Ana Gladys Tacuri[1], Mario Oswaldo Basurto[3]

[1] *Universidad de las Fuerzas Armadas (ESPE), Sangolquí, Ecuador*
[2] *Secretaría de Inteligencia del Gobierno Ecuatoriano, Quito, Ecuador*
[3] *Universidad Tecnológica Israel, Quito, Ecuador*

Correspondência para/*Correspondence to*: {cjvillacis, wmfuertes, mezambrano, eptorres3, hmaules, agtacuri}@espe.edu.ec, obasurto@uisrael.edu.ec, christian.bustamante@sin.gob.ec

Recebido/*Submitted*: 30 Out. 2014
Aceito/*Approved*: 15 Nov. 2014

Resumen

Introducción: Los videojuegos educativos, además de proporcionar distracción y diversión, estimulan el desarrollo del pensamiento de los niños en las áreas lógica y espacial. Esta investigación presenta la optimización de un videojuego educativo relacionado con los juegos de lógica, conocido como el Tres en Raya o Tic-Tac-Toe (en inglés), enfocado a niños entre 7 y 11 años.
Método: Para llevarlo a cabo, se ha empleado el Método de Diseño Hipermedia Orientada a Objetos para el diseño conceptual y navegacional de la aplicación, lo cual permite crear interfaces de usuario interactivas y amigables con el usuario. Además, se utilizó técnicas de Inteligencia Artificial para que el jugador, simulado y controlado por la computadora, pueda emular la toma de decisiones por sí mismo, y esté en capacidad de enfrentar al usuario. Las técnicas de Inteligencia Artificial utilizadas son del tipo heurísticas, implementadas mediante un método numérico inédito, basado en series numéricas finitas, lo cual difiere de otros del mismo tipo, donde lo común es utilizar el Algoritmo del MIN-MAX. Para la ejecución del juego se ha recurrido a la incorporación de un agente virtual que brindará soporte al usuario en los diferentes niveles de dificultad del juego.
Resultados: Los resultados muestran que existen diferencias cognitivas por edad, género y nivel de dificultad en el juego de una muestra representativa de niños de una unidad educativa.
Conclusión: Eses tipo de programa estimula el desarrollo cognitivo de los niños que se encuentran en su etapa primaria de formación.

Palabras-clave: Videojuegos educativos. Diseño Hipermedia Orientado a Objetos. Inteligencia Artificial. Tres en Raya.

Abstract

***Introduction**: Educational games, in addition to providing distraction and fun, stimulate the development of children's thinking in logic and spatial areas. This research presents the optimization of an educational video game related games logic, known as Noughts and Crosses or Tic-Tac-Toe, focused on children between 7 and 11 years.*
***Method**: To carry this out, we used the Object Oriented Hypermedia Design method for conceptual design and navigational application, which allows creating interactive and user friendly interfaces. In addition, Artificial Intelligence techniques used for the player simulated as well as controlled by computer, can emulate decisions for itself and be able to face the user. These Artificial Intelligence techniques are of heuristic type, implemented through an unprecedented numerical method based on finite numerical series, which differs from others of the same type, where it is common to use the algorithm of MIN-MAX. A virtual agent, that provides support to the users at different levels of difficulty of the game, was incorporated.*
***Results**: The results obtained show that there are cognitive differences by age, gender and level of difficulty in the game from a representative sample of children of an educational unit.*
***Conclusion**: This kind of program stimulates cognitive development in children who are in their primary stage of educational training.*

***Keywords**: Educational Games. Object-Oriented Hypermedia Design Method. Artificial Intelligence. Tic-Tac-Toe.*

INTRODUCCIÓN

El juego Tres en Raya estimula la cognición de los niños. Probablemente es el juego más difundido, y sencillo en su concepción, en el cual un jugador gana si consigue tener una línea recta de tres de sus símbolos del mismo tipo. La línea puede ser horizontal, vertical o diagonal. Es uno de los juegos clásicos que fueron creados en el Medio Oriente, para el desarrollo de los niños, motivándoles su destreza y habilidad mental que coadyuva a un mejor desarrollo. Normalmente son los niños pequeños los que juegan al Tres en Raya. La misma simplicidad del juego lo hace ideal como herramienta pedagógica para enseñar los conceptos de teoría de juegos.

Existen varios métodos para resolver el problema del Tres en Raya tales como el algoritmo del Minimax (Russell & Norvig, 2002), el algoritmo de búsqueda Alpha-Beta (Watson, 2008), algoritmos genéticos (Bhatt, Varshney, & Deb, 2008), redes neuronales (Grim, Somol, & Pudil, 2005), Inteligencia Artificial basada en estrategias (Chakraborty, 2009), entre los más conocidos. La mayoría de estos métodos tienen un alto grado de dificultad en la resolución de este juego. Ante este escenario, esta investigación intenta encontrar una solución menos compleja basada en un sistema de reglas combinadas con técnicas heurísticas, aleatoriedad y una máquina de estados finitos representada por una lista enlazada.

Diferentes estudios han demostrado que el juego de Tres en Raya requiere pensamiento creativo, actitud para solucionar problemas, capacidad para adquirir nuevas destrezas y habilidad para usar herramientas de software. Desde este enfoque surge el siguiente cuestionamiento ¿El juego de Tres en Raya, con diferentes niveles de dificultad, incrementa la estimulación del pensamiento cognitivo en el área lógica y espacial de los niños entre 7 y 11 años? Las preguntas orientadoras de la investigación en cambio son las siguientes: ¿Cuáles son las principales teorías educativas que fundamentan el uso de videojuegos educativos para el desarrollo cognitivo de los niños? ¿Cuál es la metodología más adecuada para diseñar, implementar y poner en funcionamiento videojuegos educativos? ¿Qué modelo matemático se puede aplicar para resolver el juego del Tres en Raya? ¿Cómo se puede evaluar y validar el videojuego educativo propuesto?

Como alternativa de solución, esta investigación presenta la implementación optimizada del video juego de Tres en Raya, enfocado a niños entre 7 y 11 años. Para llevarlo a cabo se ha empleado el Método de Diseño Hipermedia Orientada a Objetos (*Object-Oriented Hypermedia Design Method* – OOHDM) según Schwabe, Rossi, & Barbosa (1996) para el diseño conceptual y navegacional de la aplicación, lo cual ha permitido crear interfaces de usuario interactivas y amigables con el usuario. Además se utilizó técnicas de Inteligencia Artificial (IA) para que el jugador simulado y controlado por la computadora pueda tomar decisiones por sí mismo y pueda enfrentarse al usuario. Las técnicas de IA utilizadas son del tipo heurísticas implementadas mediante un método numérico inédito basado en series numéricas finitas, lo cual difiere de otros del mismo tipo, donde lo común es utilizar el Algoritmo del MIN-MAX. Para su ejecución se ha recurrido a la incorporación de un agente virtual que brindará soporte al usuario en los diferentes niveles de dificultad del juego. Los resultados muestran que este tipo de programas estimulan el desarrollo cognitivo de los niños que se encuentran en su etapa primaria de formación.

Entre las principales contribuciones de este estudio se puede mencionar: (1) obtener un sistema de reglas combinadas de razonamiento con técnicas heurísticas, aleatoriedad y una máquina de estados finitos representada por una lista enlazada; (2) diseño y construcción de una aplicación con una interface gráfica de usuario (GUI) con Programación Orientada a Objetos (POO); (3) Implementación de una librería de clases que permiten representar el ambiente y las reglas del juego del Tres en Raya.

El resto del artículo ha sido organizado como sigue: La sección Marco Referencial describe las teorías que sustentan esta investigación. La sección Diseño e Implementación detalla el modelo de casos, la especificación de requerimientos, la implementación de algoritmos del juego y la implementación de su interfaz gráfica. En la siguiente sección se muestran y se discuten los resultados obtenidos. En la sección Trabajos Relacionados se analiza el estado del arte. Finalmente, se presentan las conclusiones y posibles trabajos futuros sobre la base de los resultados obtenidos.

MARCO REFERENCIAL

El juego del Tres en Raya es un juego de estrategia en el que participan dos jugadores que se enfrentan entre sí, en un tablero de seis fichas, cada uno con tres fichas de un mismo tipo, pero diferentes de su oponente. Consiste, por un lado, en lograr alinear las tres fichas formando una raya horizontal, vertical o diagonal, y por otro, en tratar de entorpecer los movimientos del contrario para evitar que el usuario o la máquina consigan alinear sus fichas antes.

Teorías de Aprendizaje

Entre las principales teorías que sustentan el hecho de que el juego de Tres en Raya estimula la cognición de los niños, podemos citar el estudio de Feuerstein, Klein, & Tannenbaum (1991), en relación al desarrollo de habilidades cognitivas, quien considera que el juego sin lugar a dudas es un medio de aprendizaje que ocurre de forma vivencial. Sostiene que durante su ejecución se producen situaciones de percepción, atención, memoria y razonamiento, generando cambios en el pensamiento y el comportamiento. Estas capacidades que se producen en el cerebro se denominan modificabilidades cognitivas. Esta teoría pretende desarrollar la capacidad humana modificándola a través de la exposición directa a los estímulos y a la experiencia, a través del aprendizaje formal e informal, destacando el papel especial del mediador. Por tanto la mediación para que produzca modificabilidad cognitiva, ha de ser intencionada, con significado, que genere pensamiento positivo,

que logre controlar y regular sus metas, animando a compartir, fomentando empatía pero al mismo tiempo consciente de la individualidad, causando en el niño desafío, confrontación, con respuestas divergentes para la solución de problemas.

En este mismo contexto, la perspectiva de Vygotsky (2000), radicaliza que el desarrollo cognitivo ocurre por la intervención de otra persona, siendo capaz de producirla a través de una buena mediación. Introduce el concepto, de Zona de Desarrollo Próximo (ZDP), que se define como la distancia que hay entre los resultados del aprendizaje autónomo del niño (nivel actual de desarrollo) y los resultados posibles con intervención pedagógica, entendida como la guía de un adulto o en colaboración con otro compañero más capaz (nivel de desarrollo potencial). Esta zona es diferente según la persona y, en este sentido, considerando esta diversidad, pretende que el alumnado alcance los mayores y mejores resultados posibles dentro de su ZDP.

Finalmente, Lipman (1998), basa su teoría en el lenguaje a través de técnicas de diálogo. Su postulado es muy cercano a la propuesta de Feuerstein, y resulta complementaria y potenciadora de la teoría de la modificabilidad cognitiva. El Diálogo estimula la reflexión, a través de preguntas y de respuestas elaboradas nuevamente en forma de pregunta, proceso que conduce al desarrollo de habilidades de razonamiento, clarifica significados, analiza conceptos, descubre supuestos implícitos, es decir creando un escenario intencionado para pensar bien y pensar autónomamente. Al asociar los supuestos teóricos de la modificabilidad cognitiva y el juego de Tres en Raya, se deduce que durante sus niveles de complejidad los niños expuestos a la vivencia del juego presumiblemente fortalecerán una serie de habilidades cognitivas, tales como la navegación espacial, el razonamiento, la memoria y la percepción tridimensional.

Inteligencia Artificial para Videojuegos

La Inteligencia Artificial (IA) es un campo de las ciencias computacionales que se puede aplicar en el proceso de enseñanza aprendizaje, que fomenta el razonamiento lógico, para resolver problemas complejos y encontrar soluciones de una manera más rápida y segura. La IA otorga a la computadora la capacidad de aparentar un raciocinio humano, y logra el dominio del aprendizaje por el reforzamiento y ejercitación, favorece procesos de construcción de conocimiento, reconoce una extensa gama de errores de razonamiento, provee conjuntos de problemas distintos y gradúa la dificultad relativa (Bello, 2002), (Rich & Knight, 1994), (Torres Vinueza, Fuertes, Villacís Silva, Zambrano Rivera, & Prócel Silva, 2013), (Villacis, Fuertes, Bustamante, Almachi, Procel, Fuertes, & Toulkeridis, 2014). En este proyecto se ha seleccionado la combinación de técnicas heurísticas que hacen posible resolver más rápidamente problemas conocidos o similares a otros conocidos.

Máquina de Estados Finitos

Una máquina de estados finitos es un dispositivo o un modelo de un dispositivo, el cual tiene un número de estados que pueden ser activados en cualquier momento dado. Esta puede operar sobre entradas que toman cualquier transición desde un estado a otro, o pueden causar una salida o acción que puede tomar. Una máquina de estados finitos puede solamente estar en un estado en un determinado tiempo (Buckland, 2005).

Series Numéricas

Una serie numérica es la suma de una sucesión ordenada de elementos. Una serie es finita cuando se conoce el número de elementos o cuando el número de sumandos termina (De Burgos, 2013). En forma general una serie infinita se representa por $\sum_{n=1}^{\infty} u_n$ donde u_n representa la ley de desarrollo y n = {1, 2, 3, 4,...}, entonces:

$$\sum_{n=1}^{\infty} u_n = u_1 + u_2 + u_3 + \cdots + u_n + \cdots \quad (1)$$

DISEÑO E IMPLEMENTACIÓN

En esta sección se muestra cómo se diseñó e implementó el videojuego didáctico “Tres en Raya”. El proyecto consistió de cuatro niveles de dificultad, un nivel para jugar entre dos usuarios y tres niveles para jugar contra el jugador controlado por la computadora (JCC) con IA.

Modelo IA, caso Tres en Raya

Para el modelo de IA de la aplicación se han empleado técnicas heurísticas tanto débiles como fuertes, en el cual se utiliza un método numérico basado en series numéricas, que son representadas por listas enlazadas y arreglos. Estas se encargan de almacenar los diferentes movimientos hechos por la misma aplicación que viene a ser el jugador controlado por la computadora (JCC) y el usuario, donde cada movimiento se opera en base a una máquina de estados finitos. En la Tabla 1, se indica el estado inicial de todo el arreglo (i.e., cero) y corresponde a un espacio vacío o a un casillero libre:

Objeto	Estado
Usuario	1
Jugador controlado por la computadora (JCC)	3
Espacio vacío	0

Tabla 1. Estados finitos del juego.
Fuente: Elaboración propia.

Así, dado:

$$V_k \in \mathbb{E}, desde\ k = 0, hasta\ k = 8$$
$$V_k \in \mathbb{E}, para\ 0 \leq k \leq 8 \quad (2)$$

El vector V_k está representado en memoria RAM por un arreglo unidimensional llamado mArregloJugadas, cuyos valores iniciales corresponden a cero, como se puede ver en la Figura 1:

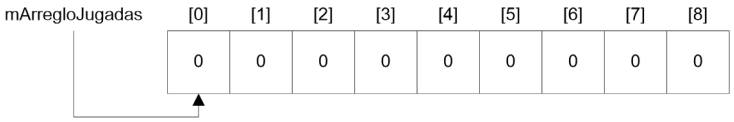


Figura 1. Representación del vector V_k a través de un arreglo unidimensional.
Fuente: Elaboración propia.

El método numérico basado en series finitas se indica en lo Cuadro 1, donde cada serie finita ha sido obtenida en base a una sumatoria que representa a un valor acumulado en una determinada fila, columna o diagonal del juego del Tres en Raya:

Filas	$\sum_{i=0}^{n=2} f_i = a$	$\sum_{i=3}^{n=5} f_i = b$	$\sum_{i=6}^{n=8} f_i = c$
Columnas	$\sum_{i=0}^{n=6} c_i = d$ $Step\ 3$	$\sum_{i=1}^{n=7} c_i = e$ $Step\ 3$	$\sum_{i=2}^{n=8} c_i = f$ $Step\ 3$
Diagonales	$\sum_{i=0}^{n=8} d_i = g$ $Step\ 4$	$\sum_{i=2}^{n=6} d_i = h$ $Step\ 2$	
Diagonales (Caso Trivial)	$\sum_{i=0}^{n=8} d_i = x$ $Step\ 4$	$\sum_{i=2}^{n=6} d_i = y$ $Step\ 2$	

Cuadro 1. Método numérico basado en series finitas.
Fuente: Elaboración propia.

Caso 1: Bloquea el Jugador controlado por la computadora al usuario. En este caso de debe considerar lo siguiente:

$$si\ a = 2 \vee b = 2 \vee c = 2 \vee d = 2 \vee e = 2 \vee f = 2 \vee g = 2 \vee h = 2 \quad (3)$$

Entonces se va a generar:

$$si\ v_{[k]} = 0 \ \longrightarrow\ v_{[k]} := 3 \ \ \wedge Bloquea\ JCC \tag{4}$$

A continuación en la Figura 2 se muestra un ejemplo para el Caso 1:

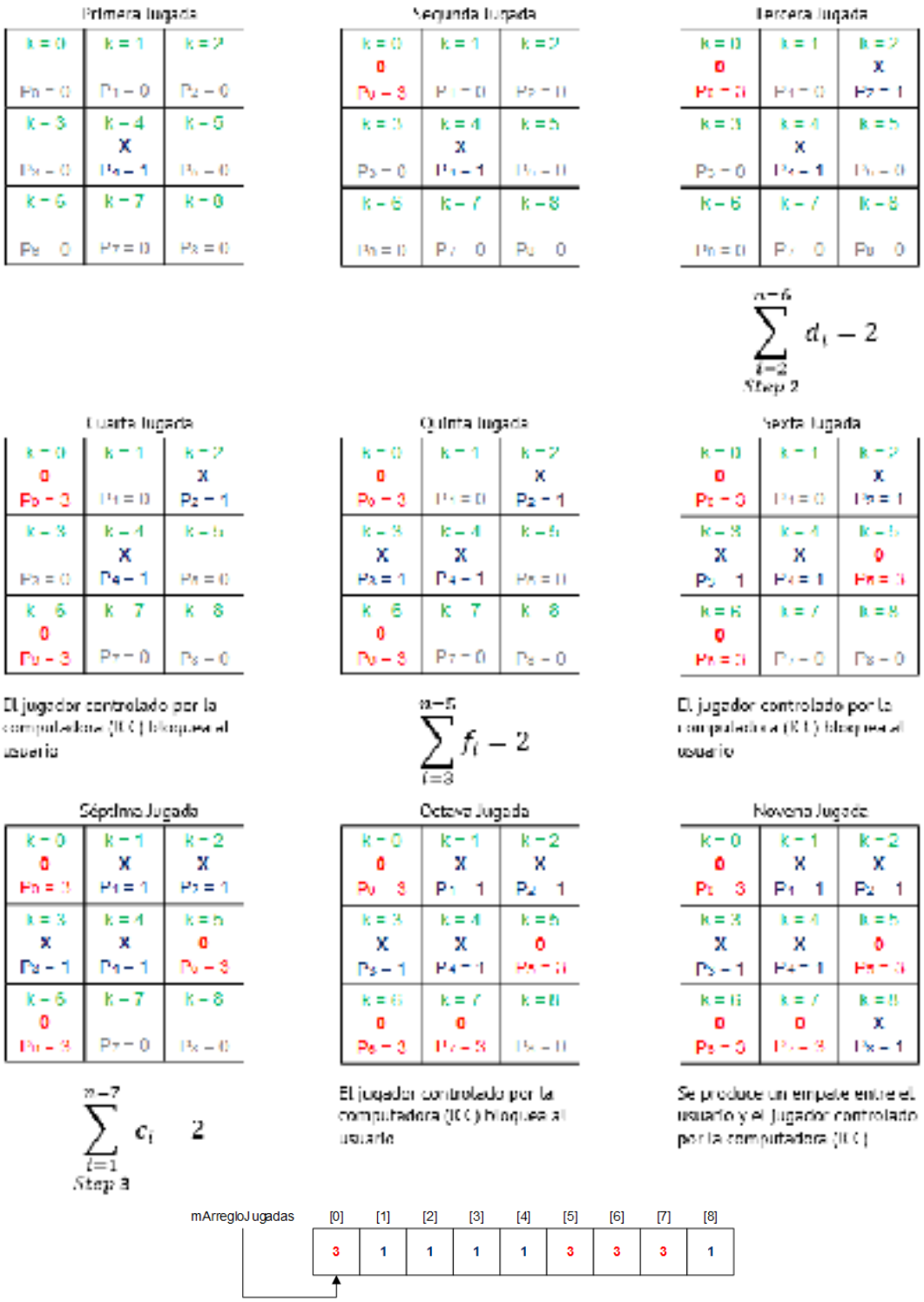


Figura 2. Ejemplo para el Caso 1.
Fuente: Elaboración propia.

Caso 2: Gana el Jugador controlado por la computadora (JCC) al usuario. En este caso de debe considerar lo siguiente:

$$si\ a = 6\ \vee\ b = 6\ \vee c = 6\ \vee d = 6\ \vee e = 6\ \vee f = 6\ \vee g = 6\ \vee h = 6 \tag{5}$$

Entonces se obtendrá:

$$si\ v_{[k]} = 0 \ \longrightarrow\ v_{[k]} := 3 \ \ \wedge Gana\ JCC \tag{4}$$

A continuación en la Figura 3 se muestra un ejemplo para el Caso 2:

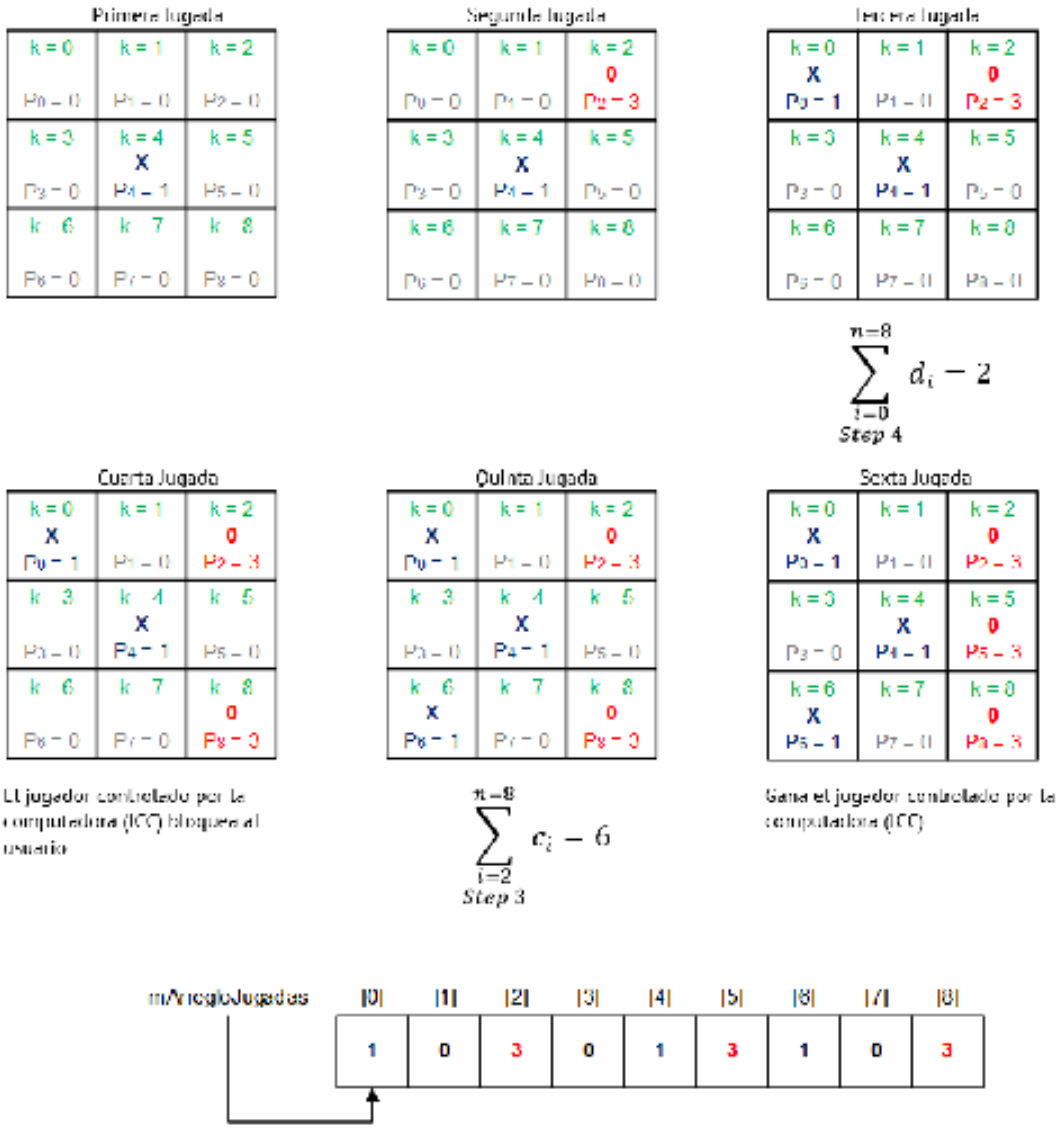


Figura 3. Ejemplo para el Caso 2.
Fuente: Elaboración propia.

Caso Trivial: Bloqueo en las diagonales. En este caso trivial de debe considerar lo siguiente:

$$si\ x = 5\ \vee\ y = 5$$

Este caso trivial genera:

$$si\ v_{[k]} = 0 \ \longrightarrow\ v_{[k]} := 3 \ \ \wedge Bloquea\ JCC$$

A continuación en la Figura 4 se muestra un ejemplo para el Caso Trivial:

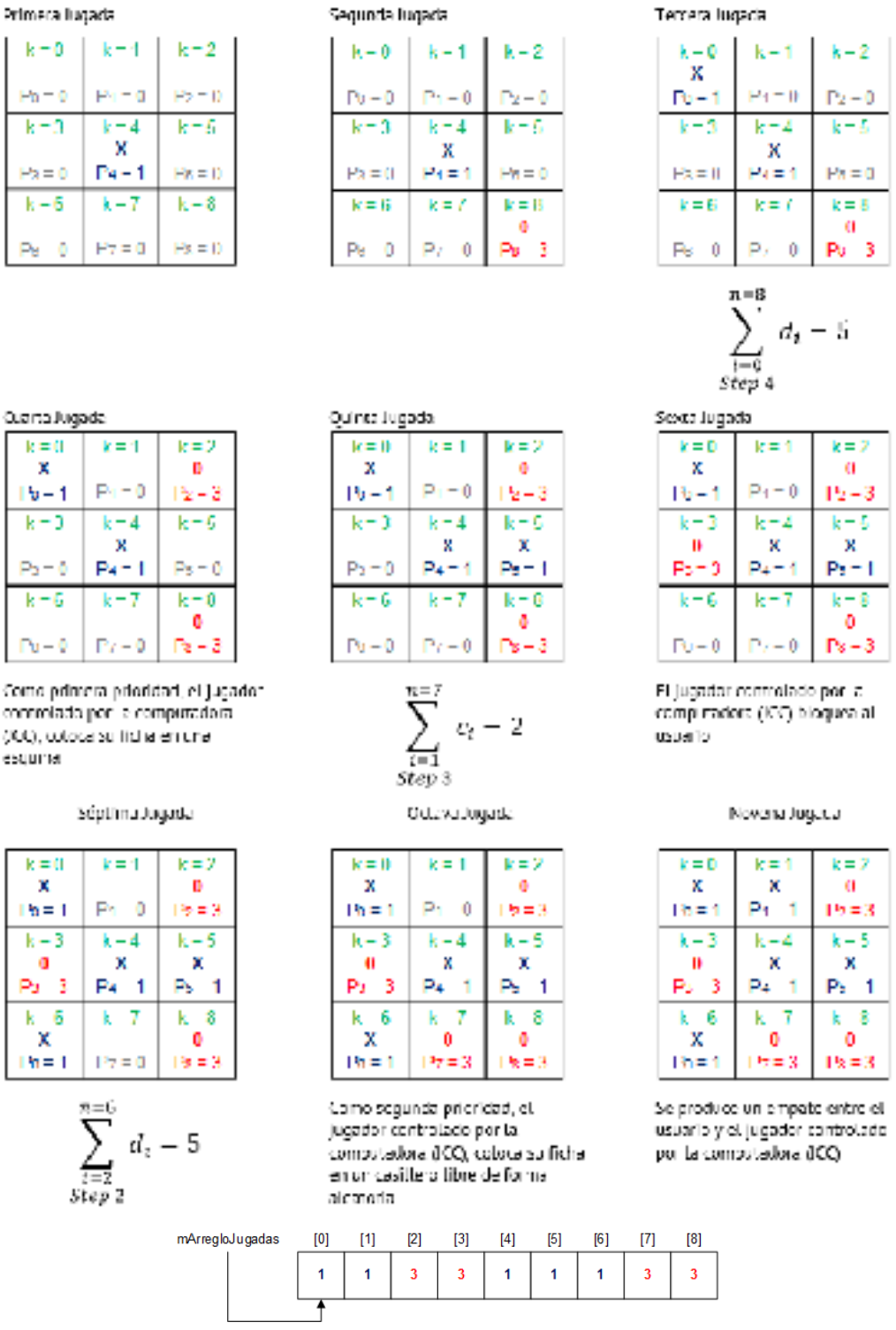


Figura 4. Ejemplo para el Caso Trivial.
Fuente: Elaboración propia.

Especificación de Requerimientos

El sistema tuvo que cumplir las siguientes tareas y roles: (1) Dos jugadores; (2) nivel principiante; (3) nivel intermedio; (4) nivel avanzado; (5) opción archivo; (6) opción ayuda; (7) opción ver. Las Figuras 5 y 6, muestran los diagramas de casos de uso de la aplicación del Juego del Tres en Raya y el diagrama de secuencias de ¿cómo cargar el juego?:

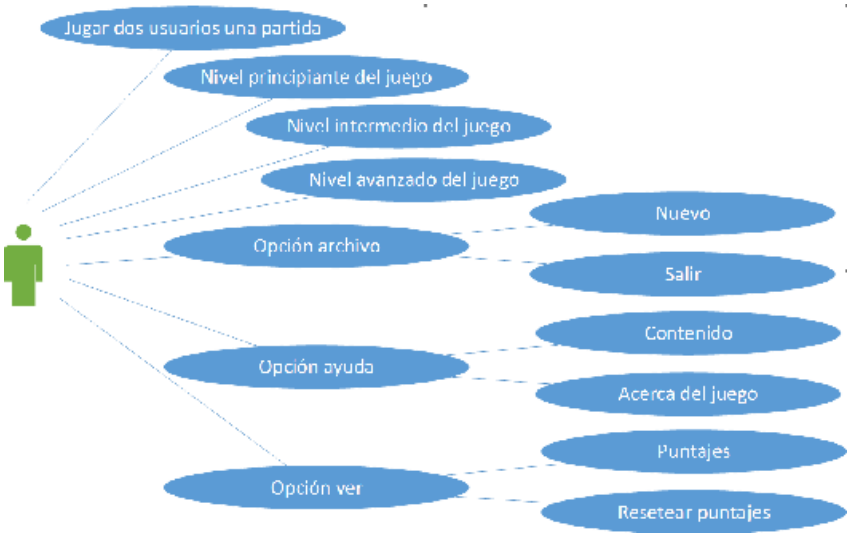


Figura 5. Caso de uso del Juego del Tres en Raya.
Fuente: Elaboración propia.

Diseño Conceptual

En esta fase, se diseñaron e implementaron archivos planos para la manipulación y almacenamiento de la información, como son puntajes, datos referenciales, configuraciones y usuarios del sistema. Así mismo, se determinó que la arquitectura a utilizar sería Cliente Servidor de 2-Capas, el cual se complementa con OO-HDM, separando el diseño de interfaz con las reglas del negocio y los controladores del juego. El sistema posee cuatro controladores del juego los cuales son: (1) controlador de dos usuarios; (2) controlador de jugadas aleatorias, (3) controlador de jugadas inteligentes débiles, y (4) controlador de jugadas inteligentes fuertes. Además posee tres clases para el manejo de animaciones, manejo del sonido y el manejo de archivos planos del sistema.

Modelo Navegacional

En esta fase, el desarrollo de las interfaces estuvo marcado por el uso de formularios (los controles *Form*), los que permiten una adecuada forma para desarrollar vistas, además de ser estéticamente acertadas. Los objetos Navegacionales son: formulario del juego del Tres en Raya, formulario acerca del juego, formulario del contenido del juego y formulario de los puntajes del juego. Los contextos navegacionales son: seleccionar nivel de dificultad del juego, seleccionar las piezas del tablero del juego, menú archivo, menú ayuda del juego, menú ver puntajes.

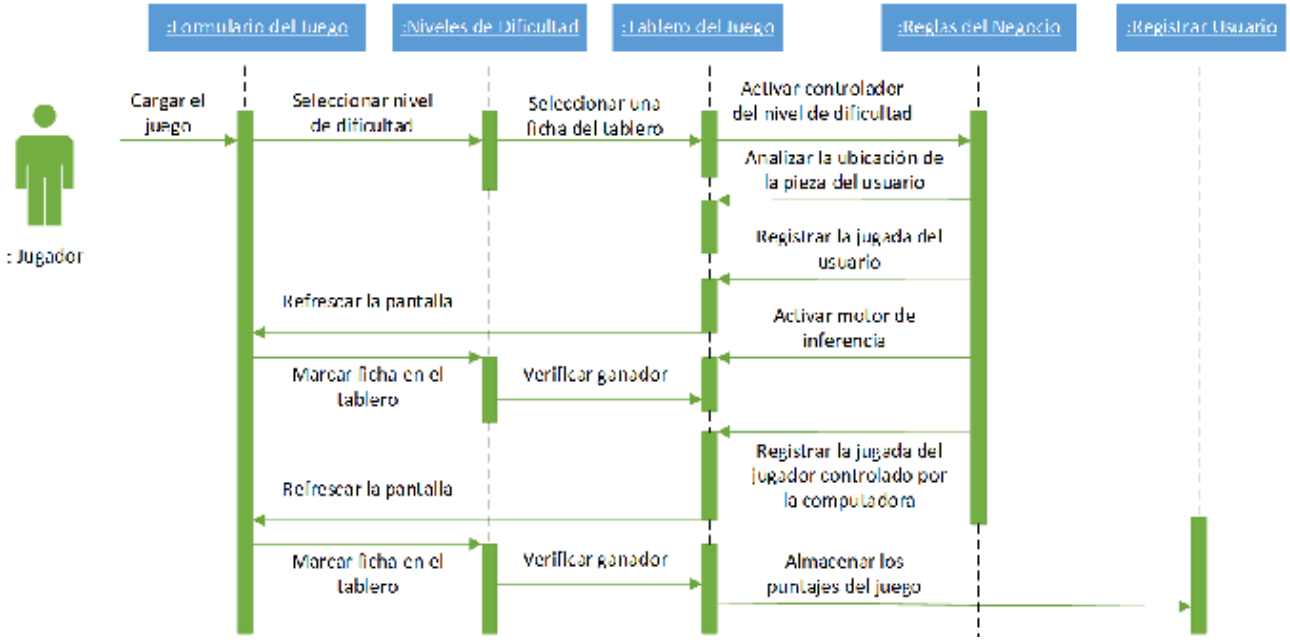


Figura 6. Diagrama de secuencias "Cargar el juego".
Fuente: Elaboración propia.

Construcción de la Interfaz Gráfica del Juego del Tres en Raya

La interfaz del juego se divide en la clase *Programa*, que se encarga de mostrar la ventana principal de la aplicación representada por la clase frmTresEnRaya, que presenta los tres niveles de dificultad del juego (nivel principiante, nivel intermedio y nivel avanzado), y que contiene la opción de jugar dos usuarios entre sí. El juego además tiene tres formularios adicionales los cuales son: a) Formulario Acerca del juego (frmAcercaDe) que presenta la información de los creadores del juego sin fines de lucro y donde se indica el uso del logotipo animado de *Pocoyo*[1]; b) Formulario Contenido (frmContenido) que presenta la información del contenido y funcionamiento del juego desarrollado en la herramienta de autor Articulate Studio; c) Formulario Puntajes (frmPuntajes) que presenta la información acerca de los puntajes obtenido por el usuario del juego. El juego del Tres en Raya tiene derechos reservados en la empresa Virtual Learning Solutions. La Figura 7 muestra la interfaz gráfica del videojuego didáctico y su ejecución en los tres niveles de dificultad:

[a] [b] [c]

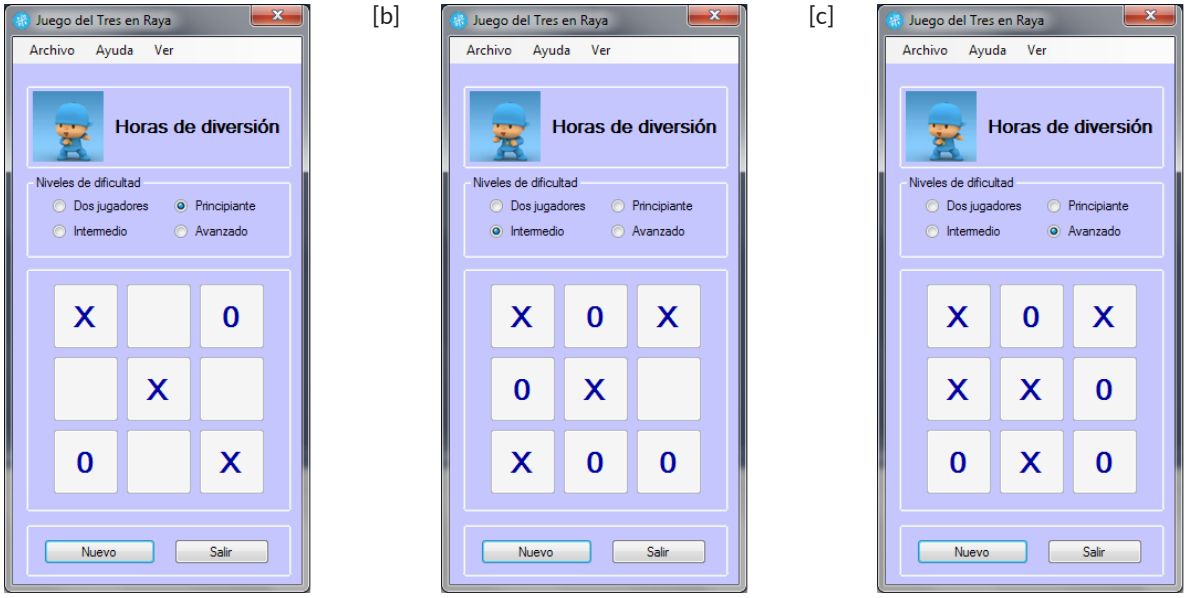


Figura 7. **(a)** Nivel principiante. **(b)** Nivel intermedio. **(c)** Nivel avanzado.
Fuente: Elaboración propia.

EVALUACIÓN DE RESULTADOS

Para la evaluación del juego de Tres en Raya, se receptó una muestra aleatoria a 30 niños/as comprendidos entre las edades de 7 a 11 años de una institución educativa, con la finalidad de conocer el nivel de aceptación del juego, como un generador desafiante al proceso mental cognitivo de los aprendizajes lúdicos. Mediante el juego que está enfocado a la creatividad, al razonamiento lógico; el desafío es ganarle al juego, el mismo que consta de tres niveles principiante, intermedio y avanzado. Para esto se aplicó un instrumento de medición estadístico de observación directa en varias computadoras, en un mismo tiempo determinado. Luego de su procesamiento estadístico se obtuvo los siguientes resultados:

En relación a la media aritmética, el 47.9% gana al juego en el nivel 1, mientras que el 77.2 % lo hace en el nivel 2. Esto se da cuando el niño/a ya está más familiarizado con el mismo. Sin embargo el 0 % no logra ganar en el último nivel. Esto debido a que el grado de dificultad es mayor y por lo tanto tiene que desarrollar un mayor número de destrezas cognitivas y fisiológicas (concentración). En relación con la desviación estándar se presenta un fenómeno casi parecido, es decir; el nivel 1 tiene: 8.424, el dos: 9.326 y el tercero se ha desestimado, debido que en el nivel 1 nadie pierde, mientras que en el tercero nadie puede lograr su objetivo (i.e. ganarle al juego Tres en Raya.

A continuación nos apoyaremos también de manera experimental y teórica en el análisis de los gráficos siguientes en relación a los niveles de dificultad: El Gráfico 1(a) muestra que el 80% de los/as niños /as comprendidos/as entre las edades de 7 a 9 años logran ganarle al juego Tres en Raya en el primer nivel, mientras

[1] Images © Granada/Zinkia, 2015. Pocoyo © Zinkia S. L. Bajo Licencia de G.V. La Serie, los logotipos y los personajes de la Serie Pocoyó son marcas registradas de Zinkia entertainment S. L. y se utilizan bajo licencia.

que el 6.7% entre las edades de 10 a11 años logran hacerlo. Esto indica que mientras menor edad tiene el/a niño/a, posee mayor interés por los juegos educativos, por lo que su aprendizaje se desarrolla de manera significativa.

El Gráfico 1(b) muestra que el 36.7% de los/as niños/as comprendidos entre las edades de 7 a 10 años logran ganarle y también empatan con el juego, en cambio que sólo el 10.0% logran ganar el juego. Es decir se observa que el 53,3% no logran su objetivo de alcanzar el siguiente nivel. Esto constituye un indicador objetivo, que indica que el juego presenta mayor grado de dificultad, por lo que los/as niños/as tendrán que desarrollar nuevas destrezas cognitivas, haciéndole al juego más desafiante e interesante para el/a niño/a.

El Gráfico 1(c) muestra que el 0% de los/as niños no alcanzan su objetivo de ganarle en este nivel al juego. Se observa que el 53,3% entre las edades de 7-9 años solo logran empatar, más bien el 36.7% pierden. Se observa además claramente que los /as niños/as entre los de 10 y 11 años, logran empatar en un 10%. Por tanto cuando se incrementa el nivel del juego, este genera mayor dificultad para quienes desean alcanzar la meta de ganarle al juego.

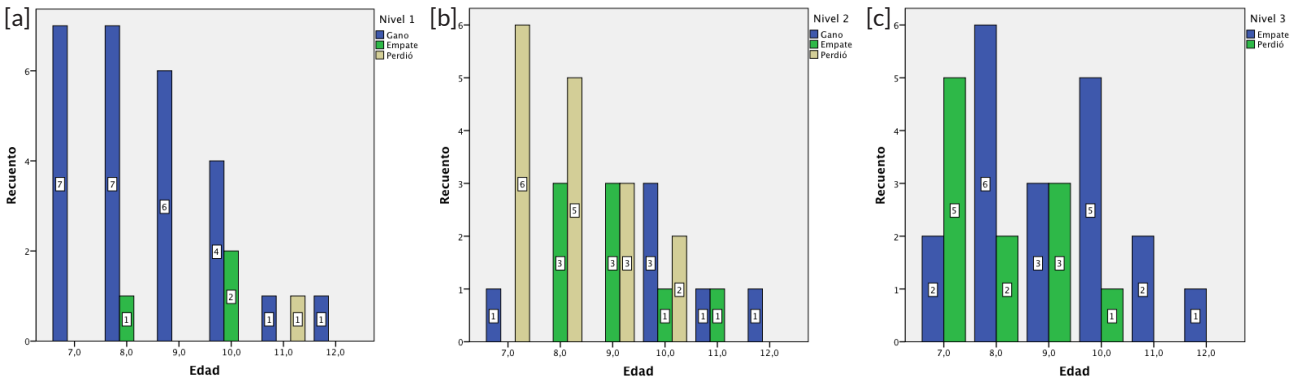


Gráfico 1. **(a)** Resultados de la Evaluación del juego Nivel 1. **(b)** Resultados de la Evaluación del juego Nivel 2. **(c)** Resultados de la Evaluación del juego Nivel 3.
Fuente: Elaboración propia.

Además se obtuvo una evaluación pedagógica en base a la matriz de evaluación de la Ficha Simplificada Catalogación y Evaluación de Programas Educativos del Dr. Pere Marqués de la Universidad Autónoma de Barcelona (2002). La Tabla 3 muestra los resultados obtenidos al justipreciar cada criterio de evaluación. Como se puede apreciar, existe una aceptación en todos los criterios evaluados por los docentes (Rango de aceptación de 0:2).

Criterio	F Test
Aspectos pedagógicos y funcionales	1,54
Aspectos técnicos y estéticos	1,75
Recursos didácticos que utiliza	1
Esfuerzo cognitivo requerido	1,43

Tabla 2. Resultados del Test de Análisis de Confiabilidad.
Fuente: Elaboración propia.

Discusión

El reto de esta investigación, está dado para que los/as niñas se interesen por las actividades lúdicas de aprendizaje de orden significativos, desarrollando la meta-cognición en juegos virtuales y educativos. Con estos juegos se busca estimular el pensamiento lógico, desarrollando sus capacidades. Lo lúdico es altamente motivador, creativo, atractivo, divertido y muy cercano a la realidad. Si a esto le adicionamos la parte del aprendizaje matemático-lógico, se torna más eficaz. Esto conduce a pensar en los juegos educativos no solo como un entretenimiento o una diversión. Hay que pensar, más bien que en la actualidad se debe concienciar del potencial educacional de los juegos, los mismos que deben tener una estrategia metodológica y dinámica que influya positivamente en los/as niños/as.

TRABAJOS RELACIONADOS

Durante la investigación se han encontrado una importante cantidad de trabajos relacionados con los videojuegos tipo Tres en Raya. A continuación se resume los revisados en este estudio:

Russell y Norvig (2002) proponen una solución basada en estrategias óptimas utilizando el algoritmo del MiniMax, donde Min tiene algo que decir y Max por tanto debe encontrar una estrategia contingente, que especifica el movimiento de Max en el estado inicial. Después los movimientos de Max en los estados que resultan de cada respuesta posible de Min de los anteriores movimientos, etc. Otros trabajos con diversas soluciones del juego de Tres en Raya son el algoritmo de búsqueda Alpha-Beta (Watson, 2008), algoritmos genéticos (Bhatt, 2008), redes neuronales (Grim et al., 2005), Inteligencia Artificial basada en estrategias (Chakraborty, 2009), entre los más conocidos. Trabajos más contemporáneos como el de Xu, Min, Zhu, & Chen (2014), que proponen implementar un algoritmo genético para resolver el problema de reducción de atributos enfocado en múltiples objetivos, que involucren pruebas de costos de múltiples tipos. Para evaluar el rendimiento del algoritmo, se definen tres métricas, desde el punto de vista estadístico. Este trabajo se diferencia del nuestro, ya que en este se realizan análisis estadísticos sobre varios juegos usando un algoritmo genético, mientras que en el nuestro se utilizan diferentes tipos de heurísticas. En otro contexto, Bondre, Kapgate, & Ponkshe (2014), proponen la implementación del juego del Tres en Raya, utilizando para el efecto conceptos avanzados de Procesamiento de Imágenes y Técnicas de Programación de Juegos. Se propone que el desarrollo del juego se base en la simplicidad, de modo que sea fácil de entenderlo y utilizarlo. Con este propósito se determina utilizar un Láser de luz roja y un sistema de visión por computadores y procesamiento de imágenes combinado con una cámara Web, para controlar la selección y posicionamiento de movimiento escogido por el usuario. Comparado con el nuestro, el desarrollado en este proyecto utiliza un rayo láser y una cámara Web para llevar el control del juego entre el usuario y la máquina. Gronli, Hansen, Ghinea, & Younas (2014), proponen una implementación de una versión móvil del juego de Tres en Raya para realizar una comparación de aspectos claves de estudio de tres plataformas líderes en el mercado que son Androide (basada en Linux de Google), Windows Phone (de Microsoft), e iOS (de Apple). Nuestro trabajo se enfoca en la implementación del juego con fines didácticos.

CONCLUSIONES Y TRABAJO FUTURO

En esta investigación se optimizó el video juego de Tres en Raya. Para lograrlo se ha empleado OOHDM, para el diseño conceptual y navegacional de la aplicación, con lo cual se creó una interfaz de usuario interactiva y amigable. Para la implementación de algoritmos se aplicó técnicas heurísticas de Inteligencia Artificial para que el jugador simulado y controlado por la computadora, pueda emular la toma de decisiones por sí mismo, y esté en capacidad de enfrentar al usuario. El método numérico es inédito, y se basó en series numéricas finitas, con lo cual se redujo la dificultad de los algoritmos. Para la ejecución del juego se incorporó un agente virtual que brindó soporte al usuario en los diferentes niveles de dificultad. Las pruebas de evaluación del video juego, fueron realizados en una muestra estratificada y representativa de niños y niñas de 7 a 11 años. Los resultados muestran que este tipo de programas estimulan el desarrollo cognitivo de los niños que se encuentran en su etapa primaria de formación.

Como trabajo futuro se plantea el diseño y desarrollo del juego de Tres para ambientes 3D distribuidos enfocados a la educación y al entretenimiento, multi plataforma, que permiten integrar servicios de reconocimiento de voz o faciales (*Speech or Facial Recognition*).

REFERENCIAS

Bello, R. (2002) *Aplicaciones de la Inteligencia Artificial*. Guadalajara: Universidad de Guadalajara, México.

Bhatt, A., Varshney, P., & Deb, K. (2008). In search of no-loss strategies for the game of tic-tac-toe using a customized genetic algorithm. *Proceedings of Genetic and Evolutionary Computation conference, Atlanta*, USA, 889-896.

Bondre, C., Kapgate, D., & Ponkshe, R. (2014). *Laser Guided Tic Tac Toe*. image, 6, 7.

Buckland, M. (2005). *Programming game AI by example*. Jones & Bartlett Learning. Worware Game Developers Library. (1st ed.) USA.

Chakraborty, P. (2009) Artificial Intelligence Based Strategies to Play the Tic-Tac-Toe Game. *Journal of Technology and Engineering Sciences*. l 1, (1).

De Burgos J. (2013). *Series Numéricas y de Potencias*. García Maroto Editores.

Feuerstein, R., Klein, P. S., & Tannenbaum, A. J. (Eds.). (1991). *Mediated Learning Experience (MLE): Theoretical, psychosocial and learning implications*. Freund Publishing House Ltd. England.

Grim, J., Somol, P., & Pudil, P. (2005). Probabilistic neural network playing and learning Tic–Tac–Toe. *Pattern recognition letters*, 26(12), 1866-1873.

Gronli, T. M., Hansen, J., Ghinea, G., & Younas, M. (2014, May). Mobile Application Platform Heterogeneity: Android vs Windows Phone vs iOS vs Firefox OS. *IEEE International Conference on Advanced Information Networking and Applications*, 2014, 28th, 635-641.

Lipman, M. (1998). *Pensamiento complejo y educación* (Vol. 43). Ediciones de la Torre.

Rich, E., & Knight, K. (1994). *Inteligencia artificial*. (2nd ed.) McGraw Hill: México, 1994.

Russell S., & Norvig, P., (2002). *Artificial Intelligence: A modern approach*. Pearson Education International. (2nd ed.), New Jersey – USA.

Schwabe, D., Rossi, G., & Barbosa, S. D. (1996). Systematic hypermedia application design with OOHDM. *Proceedings of the seventh ACM conference on Hypertext*, 116-128.

Torres Vinueza, M. D., Fuertes, W., Villacís Silva, C. X., Zambrano Rivera, M. E., & Prócel Silva, C. T. (2013). Puzzlemote: videojuego controlado con el mando de la Wii para niños de 6 a 10 años. *AtoZ: novas práticas em informação e conhecimento*, 2(2), 94-105.

Villacís, C., Fuertes, W., Bustamante, A., Almachi, D., Procel, C., Fuertes, S., & Toulkeridis, T. (2014, October). Multi-player Educational Video Game over Cloud to Stimulate Logical Reasoning of Children. *IEEE/ACM International Symposium Distributed Simulation and Real Time Applications*, 2014, 18th, 129-137.

Vygotsky, L. S. (2000). *El desarrollo de los procesos psicológicos superiores*. Crítica. Grijalbo, Barcelona.

Xu, B., Min, F., Zhu, W., & Chen, H. (2014). A Genetic Algorithm to Multi-objective Cost-sensitive Attribute Reduction. *Journal of Computational Information Systems*, 10(7), 3011-3022.

Watson M. (2008). *Practical artificial intelligence programming with java*. The MIT Press. (3rd ed.). Massachusetts, USA.

Como citar este artigo (ABNT):

VILLACÍS, C. J.; FUERTES, W. M.; BUSTAMANTE, C. A.; ZAMBRANO, M. E.; TORRES, E. P.; AULES, H. M.; TACURI, A. G.; BASURTO, M. O. Optimización del juego tres en raya con niveles de dificultad utilizando heurísticas de inteligencia artificial. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 95-106, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this article* (APA):**

Villacís, C. J., Fuertes, W. M., Bustamante, C. A., Zambrano, M. E., Torres, E. P., Aules, H. M., Tacuri, A. G., & Basurto, M. O. (2014). Optimización del juego tres en raya con niveles de dificultad utilizando heurísticas de inteligencia artificial. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 95-106. Retrieved from http://www.atoz.ufpr.br

ISSN: 2237-826X
3(2), jul./dez. 2014

Redes bayesianas para predecir el estilo de aprendizaje de estudiantes en entornos virtuales

Bayesian networks to predict the learning style of student in virtual environments

Lissette Geoconda López-Faican[1], Luis Antonio Chamba-Eras[2]

[1] *Universidad Nacional de Loja (UNL), Loja, Ecuador*
[2] *Universidad Internacional del Ecuador (UIE), Quito, Ecuador*

Autor para correspondência/*Corresponding author*: Luis Antonio Chamba-Eras [luchambaer@internacional.edu.ec]

Recebido/*Submitted*: 30 Out. 2014
Aceito/*Approved*: 26 Nov. 2014

Resumen

Introducción: Describe la utilización de las Redes Bayesianas para implementar un modelo de incertidumbre que permita predecir el estilo de aprendizaje de los estudiantes mediante la interacción en un entorno virtual de aprendizaje basado en el modelo de Felder-Silverman.
Método: El modelo de incertidumbre se lo diseño y desarrolló para el funcionamiento en el LMS Moodle. Para validar el modelo propuesto se planteó un escenario educativo real conformado por dos grupos experimentales pertenecientes a la Universidad Nacional de Loja y Universidad Internacional del Ecuador.
Resultados: El bloque "Estilo de Aprendizaje" (EA) permitió a los estudiantes visualizar las probabilidades de cada dimensión de su EA observando que, de acuerdo a su interacción, cambiaban dichas probabilidades. De igual forma el docente pudo visualizar las probabilidades del EA que obtuvo cada estudiante al interactuar en un curso virtual alojado en el Entorno Virtual de Aprendizaje.
Conclusión: La propuesta podrá servir como apoyo al docente que desee identificar los estilos de aprendizaje predominantes de los estudiantes y, en base a ello, preparar actividades y recursos en su aula virtual.

Palavras-chave: Inteligencia artificial. Modelo bayesiano. Modelo Felder-Silverman.

Abstract

***Introduction**: It describes the use of Bayesian Networks to implement a model of uncertainty to predict the learning style of students through their interaction in a virtual learning environment based on the Felder-Silverman model.*
***Method**: The model uncertainty was designed and developed to be integrated in the LMS Moodle. In order to validate the proposed model, an actual educational scenario was built and two groups - one from the National University of Loja and other from the International University of Ecuador - were exposed to the experiment.*
***Results**: The block "Learning Style" (EA) allowed students to visualize the probabilities of each dimension of their EA by observing that, according to their interactions, these probabilities changed. Likewise, the teachers could visualize the probabilities of EA obtained by each student when these interactions were done in the hosted virtual course enclosed in the Virtual Learning Environment.*
***Conclusion**: The proposal may serve as support for teachers who want to identify predominant learning styles of their students and, based on that, prepare activities and resources in the courses under their responsibilities.*

***Keywords**: Artificial intelligence. Bayesian model. Felder-Silverman model.*

INTRODUCCIÓN

En la actualidad se ha adquirido un gran interés en determinar cómo los estudiantes aprenden y adquieren el conocimiento en los Entornos Virtuales de Aprendizaje (EVA) (Cerrillo, 2004; Mestre, Fonseca, & Valsés, 2007; Chía, & Muñoz, [s.d.]; Mejía, 2009; Zapata-Ros, 2012; Veytia, 2013). Para cumplir con ello, sobre las plataformas virtuales se han diseñado y utilizado cuestionarios que permiten identificar el estilo de aprendizaje. Sin embargo, éste método ha demostrado no ser el adecuado debido que además de consumir tiempo, es un método poco fiable, puesto que los estudiantes tienden a elegir respuestas arbitrariamente siendo inconscientes de los usos futuros que se les puede dar a los resultados. Por lo tanto, la información obtenida puede ser inexacta y puede no reflejar los estilos de aprendizaje reales (Yannibelli, Godoy, & Amand, 2006; González, 2009).

El concepto de estilos de aprendizaje (EA) surge en los diseños instruccionales de cursos virtuales que contienen información predeterminada referente al tema que se está abordando, pero muchas de las veces no es la adecuada, ni relevante y sobre todo innecesaria para cada estudiante. Esto se debe a que por medio del EVA no se aplica una evaluación previa, que sea confiable para detectar el EA y con ello poder conocer las necesidades de formación de cada estudiante, neutralizando así, las oportunidades de mejora en la enseñanza, como la desmotivación del estudiante para aprender de acuerdo a sus preferencias subjetivas.

Los EA se definen como la forma en la que las personas recopilan, procesan y organizan la información. Para su identificación existen instrumentos psicométricos útiles para averiguar el EA, tal es el caso del modelo de

Felder-Silverman (FSLSM) ya que es uno de los modelos con mayor reputación y ha sido implementado con éxito en muchos sistemas de e-learning (Garcia, Amandi, Schiaffino, & Campo, 2007; Carmona, Castillo, & Millán, 2009; Sarango, 2012). El modelo de Felder-Silverman clasifica a los estudiantes en 4 dimensiones: procesamiento, percepción, entrada, comprensión; donde cada dimensión tiene un conjunto de estrategias que dirigen sus preferencias a ciertos recursos académicos tales como: videos, foros, chats, texto, imágenes, entre otros.

En el ámbito de la educación, particularmente en los EVA, las Redes Bayesianas (RB), conocidas como modelos probabilísticos, modelo bayesiano o red de creencia han sido objeto de investigación y de un creciente interés en cuanto a predecir los EA. Esto radica en que un modelo bayesiano está circunscrito, como técnica de pronóstico, cuya principal característica es la valoración o cualificación a hechos o datos observados. Su rol como instrumento de pronóstico es muy importante ya que permite hacer inferencias sobre la probabilidad de ocurrencia de una situación dada sobre la base de las evidencias observadas. Por ello, es un instrumento extraordinario para el monitoreo o seguimiento de situaciones de interés (Jesús, 2000).

Existen investigaciones que proponen a los modelos probabilísticos como una alternativa de solución innovadora en los entornos de educación virtual. Una investigación que se puede tomar de referencia es la Evaluación de RB (Garcia et al, 2007), siendo su objetivo el de utilizar la técnica para detectar el EA de acuerdo a los diferentes comportamientos que tiene el estudiante en el entorno virtual. Otra investigación, siendo relevante para el caso de estudio es el Modelo Bayesiano del Alumno basado en el EA y las Preferencias (Carmona et al, 2009), la misma que da a conocer un modelo de EA y un modelo de decisión para cada alumno, diseñado de acuerdo a las preferencias e interacciones del usuario con el sistema.

Bajo éste panorama, es importante que los EVA brinden información confiable acerca de la forma en que aprenden los estudiantes, siendo esto, información base para diseñar estrategias de enseñanza a fin de maximizar el proceso de aprendizaje en los entornos virtuales. Para este propósito, el presente trabajo da a conocer la implementación de un bloque basado en un modelo de red de creencia funcional para el LMS Moodle 2.5.4, el mismo que provee a los docentes, estudiantes y demás usuarios, un estimado de la probabilidad relacionada a cada dimensión de su EA, resultados que son generados de acuerdo a la interacción que mantiene el estudiante con los recursos y actividades disponibles en el EVA.

El articulo está estructurado de la siguiente manera: Introducción, presenta el objeto de estudio, estado del arte y trabajos relacionados; Metodología, detalla el modelo de la red bayesiana, su implementación y validación del modelo en el LMS Moodle; Resultados y discusión, presenta un análisis del experimento realizado, así como la discusión de los resultados obtenidos; Conclusiones, establece los logros alcanzados y las líneas futuras que se generó tras la culminación del trabajo.

METODOLOGÍA

Dentro de la metodología de la investigación se combinó el estudio de casos para la argumentación teórica del trabajo, la observación activa y el método de experimento que se utilizó para el control y seguimiento de las actividades de evaluación de la propuesta. Para poder implementar/validar la herramienta de apoyo al EVA que permita predecir para cada estudiante las probabilidades relacionadas a cada dimensión de su EA en base al modelo de Felder-Silverman se utilizó los procesos de la Ingeniería de Software diseñando una RB que fue implementada como un bloque "Estilo de Aprendizaje" para el LMS Moodle 2.5.4, que fue validado en un escenario educativo real mediante un grupo experimental conformado por dos grupos de estudiantes de la Universidad Nacional de Loja y Universidad Internacional del Ecuador.

Teniendo las bases teóricas fundamentadas y estudiadas (estudio de casos), la primera instancia fue modelar la red bayesiana, de acuerdo a los requerimientos obtenidos para la investigación (observación activa), luego se procedió a utilizar la metodología de desarrollo de *software* clásica (Ingeniería de *Software*) para el diseño e implementación del modelo probabilístico, descritas en las secciones siguientes.

Modelo de la Red Bayesiana

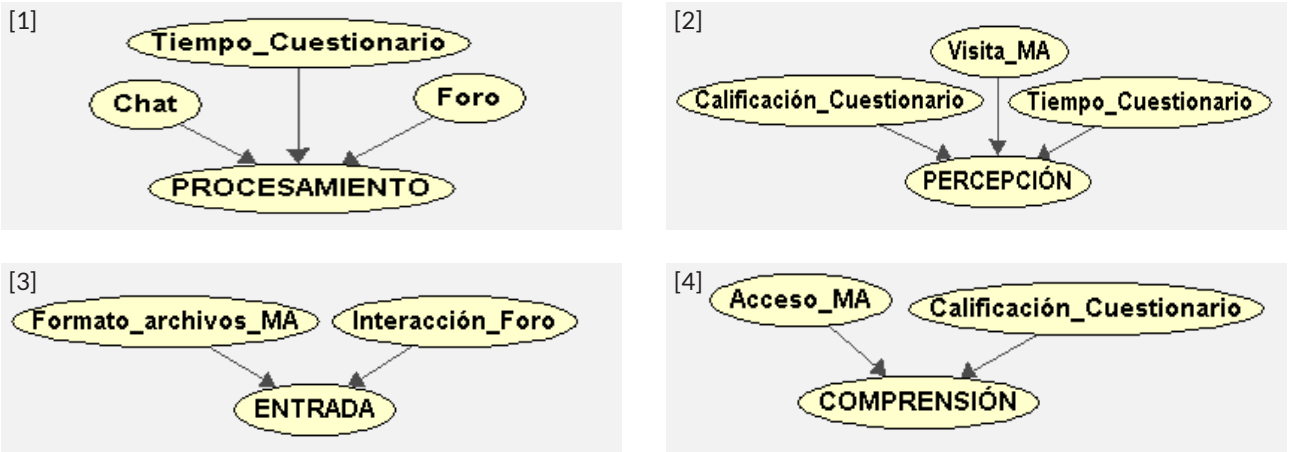
El modelo de RB representa la relación existente entre las dimensiones del modelo del EA de Felder-Silverman y los factores que lo determinan, siendo estos la interacción del estudiante con los recursos y actividades disponibles en el EVA (Cuadro1):

a) recursos: material de aprendizaje (archivo, carpeta, página, libro);
b) actividades: *chat*, foro, cuestionario.

Procesamiento (Activo, Reflexivo) Gráfico 1	
Chat	Escribe, lee mensajes, sin participación
Foro	Participa foros, Lee foros, Sin participación
Tiempo cuestionario	Bajo, Normal, Alto
Percepción (Sensitivo, Intuitivo) Gráfico 2	
Visita material de aprendizaje	Visita, No visita.
Calificación cuestionario	Bajo, Normal, Alto
Tiempo cuestionario	Bajo, Normal, Alto
Entrada (Visual, Verbal) Gráfico 3	
Formato archivos material de aprendizaje	Visual (video, imágenes) Verbal (audio, texto)
Interacción foro	Interactúa, No interactúa.
Comprensión (Secuencial, Global) Gráfico 4	
Acceso material de aprendizaje	Continuo , A saltos
Calificación cuestionario	Bajo, Normal, Alto

Cuadro 1. Relación entre EA y recursos/actividades del EVA
Fuente: Elaboración propia.

En base a las variables definidas en el Cuadro 1, se diseñó la estructura de la RB para cada dimensión del EA, la misma que está conformada por los Nodos padres (Interacción EVA), siendo nodos independientes, y los Nodos hijos (Dimensión EA).



Figuras 1-4. [1] Dimensión procesamiento. [2] Dimensión percepción. [3] Dimensión entrada. [4] Dimensión comprensión.
Fuente: Elaboración propia.

Las 4 RB (Figuras 1-4) en su conjunto integran un modelo de RB final como se observa en el figura 5, permitiendo con ello estimar las probabilidades en cada dimensión del EA.

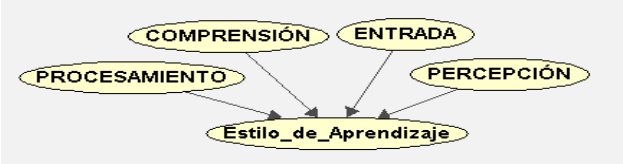


Figura 5. Red bayesiana final.
Fuente: Elaboración propia.

La predicción del EA en la RB final, se realiza mediante el proceso de inferencia, donde para ello es necesario definir las tablas de probabilidad asociadas a cada nodo. Por ello, la información útil para el proceso de inferencia está dada por:

a) la probabilidad a priori de los nodos padres: los valores para las tablas de probabilidad de los nodos **Interacción EVA**, se extrae de las evidencias que corresponde a la interacción que realiza el estudiante con el EVA;
b) la probabilidad condicionada de los nodos hijos: Para los nodos **Dimensiones del EA**, se determina una Tabla de Probabilidad Condicional (TCP), siendo esto base para realizar el proceso de inferencia en la RB. Se presenta a continuación el TCP del nodo comprensión (Tabla 1) y entrada (Tabla 2) siendo de referencia para los nodos hijos restantes. Los valores necesarios para la inferencia fueron estimados mediante datos recolectados y fuentes bibliográficas (Yu, Chen, 2006; Garcia et al., 2007; Graf, Kinshuk, & Tzu-Chien, 2009).

Acceso al MA	Continuo			Saltos		
Calificación cuestionario	Bajo	Normal	Alto	Bajo	Normal	Alto
Secuencial	0.60	0.80	1	0.40	0.20	0
Global	0.40	0.20	0	0.60	0.80	1

Tabla 1. TCP Nodo Comprensión.
Fuente: Elaboración propia.

Formato Archivo MA	Visual		Verbal	
Interacción Foro	No Interactúa	Interactúa	No Interactúa	Interactúa
Visual	1	0.75	0.25	0
Verbal	0	0.25	0.75	1

Tabla 2. TCP Nodo Entrada.
Fuente: Elaboración propia.

Implementación del modelo en LMS Moodle

El diseño de la RB se integró y codificó en un bloque llamado "Estilo de Aprendizaje" para el LMS Moodle versión 2.5.4.

En la figura 6 se observa el bloque cuya arquitectura final, altamente modular, contiene una estructura de directorios, siendo el directorio principal *ea*.

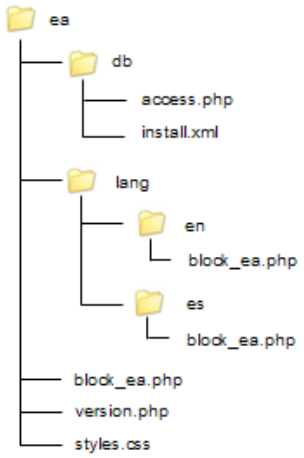


Figura 6. Directorio bloque *ea*.
Fuente: Elaboración propia.

El directorio ea contiene subdirectorios y ficheros que cumple con una funcionalidad específica, los mismos que se detallan a continuación:

a) db: directorio que contiene el fichero con los permisos particulares del bloque (access.php) y el fichero para crear las tablas del bloque EA en la base de datos Moodle (install.xml). Se crea una tabla

para cada nodo **Interacción EVA** que almacena las evidencia para cada estudiante, y una tabla para cada nodo **Dimisiones del EA**, que almacena las probabilidades condiciones (TCP) útiles para el proceso de inferencia;

b) lang: directorio que contiene todos los ficheros de idioma, para ello se crea una carpeta y un fichero por cada idioma, en el caso del bloque EA está desarrollado para el idioma Inglés (en) y Español (es);
c) block_ea.php (archivo principal del bloque): fichero que integra el funcionamiento de la RB, así como funciones complementarias;
d) version.php: hace referencia a la versión del bloque;
e) styles.css: fichero que se usa para controlar la forma en que se ven los elementos visuales (diseño) que forman parte del bloque "Estilo de Aprendizaje".

Método	Descripción
init()	Método nativo de Moodle usado para inicializar el bloque.
get_content()	Método nativo de Moodle usado para mostrar el contenido del bloque. El método genera la llamada a todos los métodos creados por el desarrollador para su respectiva ejecución, y de esta forma predecir el EA asociado a cada estudiante. • cargar_tcp() • extraer_actualizar_evidencias($userid) • inferenciaRB($userid) • listarEstudiantes() • estrategias()
cargar_tcp()	Función que carga los datos que son base para el proceso de inferencia. El método carga en las tablas de los nodos Dimensión EA, los datos definidas en las tablas de probabilidad condicional.
extraer_actualizar_evidencias ($userid)	El método obtiene las evidencias asociadas a cada estudiante, siendo almacenadas en las tablas de los nodos Interacción EVA.
inferenciaRB ($userid)	El método de acuerdo a la información almacenada en las tablas de los nodos Interacción Eva y de los nodos Dimensión EA, procede a realizar la inferencia utilizando como motor principal el Teorema de Bayes. La inferencia se realiza para las cuatro dimensiones del EA: • Procesamiento • Percepción • Comprensión • Entrada A más de ello, almacena en la tabla block_ea de la Base de Datos de Moodle los resultados obtenidos de la inferencia.
listarEstudiantes()	Método creado para listar los estudiantes de un curso con su EA. Este método es llamado cuando el usuario está en el curso con el rol de administrador, profesor y profesor sin permisos de edición.
estrategias()	Método creado para mostrar información sobre como aprenden los estudiantes según las 4 dimensiones del EA.

Cuadro 2. Métodos implementados bloque *ea*.
Fuente: Elaboración propia.

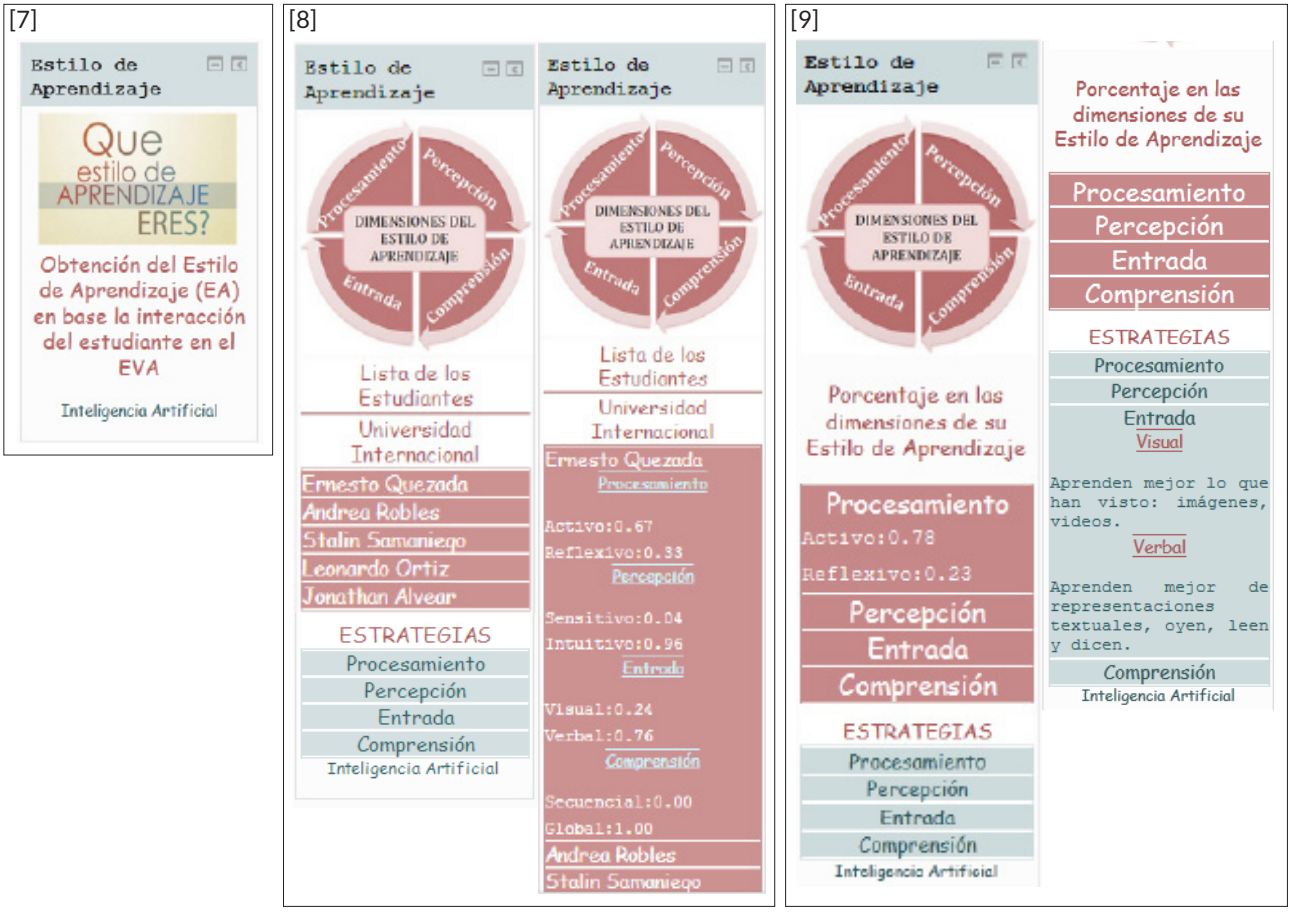
Los métodos mencionados en el Cuadro 2, relacionados al funcionamiento de la RB, se ejecutan cuando el usuario ingresa al entorno e interactúa con los recursos y actividades disponibles en el EVA.

Validación del bloque estilo de aprendizaje

Ya concluido el proceso de codificación de la RB, se procedió a implantar el bloque en el EVA basado en Moodle en el URL [http://www.estiloaprendizaje.com] a fin de monitorear y validar su funcionamiento. Para ello se estableció y creó un diseño instruccional de un curso virtual sobre "Introducción a Redes Bayesianas", a fin de que dos grupos experimentales conformado por estudiantes universitarios interactúen en el mismo y en base a ello observar el comportamiento del modelo de incertidumbre.

El bloque desarrollado para Moodle 2.5.4 genera como resultado para cada estudiante las probabilidades estimadas de su EA. El mismo que muestra la información dependiendo del rol que tiene asignado el usuario en el contexto del curso:

a) cuando el estudiante no ingresa sus credenciales en la página principal del EVA o ingreso con el rol de invitado, el bloque "Estilo de Aprendizaje" muestra su portada principal (figura 7).
b) cuando el usuario tiene asignado el rol de administrador, creador de cursos, profesor o profesor sin permisos de edición, el bloque muestra la lista de los estudiantes que pertenecen al curso, con las probabilidades de las dimensiones de su EA (figura 8).
c) cuando el usuario tiene asignado el rol de estudiante, el bloque muestra las probabilidades que tiene en cada una de las dimensiones del EA (figura 9).



Figuras 7-9. [7] Bloque estilo de aprendizaje. [8] Bloque rol docente. [9] Bloque rol estudiante.
Fuente: Elaboración propia.

Adicional, cuando el usuario se encuentra en el rol de estudiante, administrador, creador de cursos, profesor o profesor sin permisos de edición, el bloque da a conocer información referente al proceso de aprendizaje según las 4 dimensiones del EA.

RESULTADOS Y DISCUSIÓN

El bloque "Estilo de Aprendizaje" implementado en el EVA URL [http://www.estiloaprendizaje.com] luego de la experimentación realizada con los 2 grupos experimentales, permitió estimar para cada estudiante las probabilidades relacionadas a las 4 dimensiones del EA siendo estas: procesamiento, percepción, entrada, comprensión.

Para contrastar lo mencionado, fue necesario que en el EVA, se realice un diseño instruccional de un curso virtual sobre la temática de "Introducción a las Redes Bayesianas" siendo éste implementado en el EVA (Cuadro 3).

Cada sección del curso de RB estuvo compuesto por recursos y actividades de aprendizaje, las cuales de detallan en el Cuadro 4.

La finalidad del diseño instruccional en el curso del EVA fue la de permitir que el grupo experimental conformado por estudiantes universitarios interactúen con los recursos y actividades disponibles en el curso y en base a ello el bloque "Estilos de Aprendizaje" genere resultados. La población del Grupo Experimental que interactuaron en este proceso correspondió a 27 estudiantes divididos en dos grupos, 22 estudiantes de décimo módulo paralelo "A" de la carrera de Ingeniería en Sistemas de la Universidad Nacional de Loja período marzo 2014 - julio 2014 y 5 estudiantes de la Universidad Internacional del Ecuador sede Loja (Cuadro 5). El curso virtual estuvo bajo la observación-dirección de un tutor/docente experto en contenidos de Inteligencia Artificial, teniendo una duración de 22 días de interacción en el curso, (lunes a viernes), con fecha de inicio: martes, 29 de abril del 2014 y fecha de culminación miércoles, 28 de mayo del 2014.

Curso de Introducción a las Redes Bayesianas	
Secciones	**Descripción**
Bloque: Bienvenida	Dedicado a proporcionar información del curso así como a obtener las expectativas que tiene los estudiantes del mismo.
Sección 1: Introducción a las Redes Bayesianas.	Se enfoca a proporcionar los recursos y actividades que permitan a los estudiantes adentrase a las RB conteniendo los conceptos básicos para que los estudiantes comprendan el funcionamiento de la temática.
Sección 2: Teorema de Bayes en las Redes Bayesianas.	Proporciona el material y actividades útiles para que el estudiante comprenda el funcionamiento del Teorema de Bayes en cuanto a realizar la inferencia en la RB.
Sección 3: Algoritmos o Técnicas para la Inferencia.	Da a conocer diferentes algoritmos que son útiles para realizar la inferencia en las RB.
Sección 4: Herramientas.	Tiene como objetivo dar a conocer a los estudiantes algunas herramientas que existen para la aplicación de las RB.

Cuadro 3. Diseño instruccional.
Fuente: Elaboración propia.

Recursos	Actividades
• Materiales de lectura: archivos en diferentes formatos (docx, pdf, ppt, jpeg, png, gif). • Materiales de lectura complementaria (carpeta) • Libro en el formato del EVA. • Página en el formato del EVA. • Chat.	Actividades individuales de aplicación de lo estudiado tales como. • Foros de discusión. • Evaluación.

Cuadro 4. Recursos y actividades del curso.
Fuente: Elaboración propia.

Descripción	Usuarios
Universidad Nacional de Loja	22 estudiantes
Universidad Internacional del Ecuador	5 estudiantes
Tutor/Docente de la Carrera de Ingeniería en Sistemas.	Docente

Cuadro 5. Usuarios para el proceso de validación.
Fuente: Elaboración propia.

El modelo de la RB implementado en el bloque "Estilo de Aprendizaje" para Moodle 2.5.4, generó resultados cuantitativos sobre los valores de los EA de todos los integrantes del grupo experimental observables luego de finalizar las temáticas del curso virtual alojado en el EVA.

El bloque "Estilo de Aprendizaje" permitió a los estudiantes visualizar las probabilidades de cada dimensión de su EA, observando que de acuerdo a su interacción cambiaban dichas probabilidades. De igual forma el docente pudo visualizar las probabilidades del EA que obtuvo cada estudiante al interactuar en el curso de "Introducción a las Redes Bayesianas".

Finamente el bloque, fue desarrollado para una función perfectamente definida, el mismo que recae sobre brindar información que sea de apoyo para mejorar el proceso enseñanza aprendizaje en los EVA. Esta información mostrada por el bloque, que son las probabilidades estimadas que tiene un estudiante en cada una de las dimensiones del EA, permite orientar a los usuarios a estar informados sobre el proceso de aprendizaje que se dan en los cursos virtuales mediados por EVA. Por otro lado, el docente al saber cómo los estudiantes aprenden constituye otra de las principales aportaciones del trabajo, ya que en base a la información suministrada por el bloque, el docente encargado de impartir el curso puede diseñar estrategias que permitan mejorar el proceso de enseñanza aprendizaje en la educación virtual.

CONCLUSIONES

Las RB diseñan modelos que permitan llevar acabo predicciones. Por ello, por medio de las RB y sus algoritmos de inferencia, se pudo diseñar un modelo para ser implementado en el bloque "Estilo de Aprendizaje" para Moodle 2.5.4, generando así el diagnóstico de la forma en que aprenden los estudiantes, siendo éste el de predecir las probabilidades asociadas a cada dimensión del EA del modelo de Felder-Silverman.

La información generada por el bloque puede ser utilizada para diferentes propósitos. Uno de ellos sería el diseñar estrategias relacionado a los EA a fin de maximizar el aprovechamiento del aprendizaje en los entornos de educación virtual. También, es necesario mencionar que la modularidad del bloque desarrollado para el LMS Moodle deja abierta la puerta para que nuevos desarrolladores e investigadores del tema de la educación virtual mejoren con nuevos aspectos el modelo propuesto.

Los valores obtenidos en la predicción por medio de las RB, pueden servir a los docentes que trabajan sobre entornos virtuales de aprendizaje a mejorar sus diseños instruccionales y de esta manera proponer actividades y recursos de acuerdo al estilo de aprendizaje ya identificados.

La RB para predecir el EA en el EVA puede ser mejorada en diferentes aspectos, como por ejemplo: a) redefinir el modelo de la RB, identificando para ello nuevas variables relacionadas a los recursos y actividades que dispone un EVA (las mismas que deben ser útiles y relevantes para la inferencia en la red); b) efectuar una actualización de los valores definidas en las tablas de probabilidad condicional de los nodos dimensión EA, con el fin de acrecentar la validez del proceso de inferencia en la RB, garantizando con ello mayor confianza en la estimación de las probabilidades del EA.

Al tratarse de un bloque que brinda información sobre el EA del estudiante, se puede incorporar nuevas funcionalidades, siendo una de ellas, adaptar los contenidos del EVA de acuerdo a las características individuales que posee cada estudiante.

Se puede combinar el modelo bayesiano descrito en la investigación, con otras técnicas de Inteligencia Artificial como Procesamiento de Lenguaje Natural para identificar por medio los foros los estilos de aprendizaje al procesar los aportes-escritos que han hecho en sus diferentes interacciones en los temas que participen los estudiantes/tutores.

REFERÊNCIAS

Cerrillo, Q. (2004). *Aprendizaje Colaborativo y Redes de Conocimiento*. Actas de las Jornadas Andaluzas de Organización y Dirección de Instituciones Educativas. Granada-España. Grupo Editorial Universitario, 9.

Chía, L., & Muñoz, A. (n.d.). *Adaptación de las plataformas e-learning a los estilos de aprendizaje utilizando sistemas multiagentes*. Universidad Libre Cali. Colombia.

Carmona, C., & Castillo, G., Millán, E. (2009). Modelo Bayesiano del alumno basado en el estilo de aprendizaje y las preferencias. *IEEE-RITA*, 4(2), 139-146. Retirado de http://rita.det.uvigo.es/200905/uploads/IEEE-RITA.2009.V4.N2.A8.pdf

González, H. (2009). *Modelo dinámico del estudiante en cursos virtuales adaptativos utilizando técnicas de inteligencia artificial*. (Tesis de Maestria – Magister en Ingeniería de Sistemas). Universidad Nacional de Colombia.

Garcia, P., Amandi, A., Schiaffino S., & Campo, M. (2007). Evaluating Bayesian networks precision for detecting students learning styles. *Computers & Education*, 49(3), 794–808. doi:10.1016/j.compedu.2005.11.017

Graf, S., Kinshuk, & Liu, T-C. (2009). Supporting Teachers in Identifying Students Learning Styles in Learning Management Systems. *Educational Technology & Society*, 12(4), 3–14. Retirado de http://www.ifets.info/journals/12_4/2.pdf

Jesús, E. A. (2000). *Manual de metodologías, tomo II: La Técnica Bayesiana. Naciones Unidas para el Desarrollo Industrial*, Programa de Prospectiva Tecnológica para Latinoamérica y el Caribe.

Mejía, C. (2009). *Proceso de adaptación para entregar contenido basado en Estilos de aprendizaje del usuario*. (Tesis de Maestria – Máster en Informática y Automática Industrial). Universidad de Girona.

Mestre, U., Fonseca, J., & Valsés, R. (2007). *Entornos virtuales de enseñanza aprendizaje*. Editorial Universitaria.

Sarango, M. (2012). *Aplicación de técnicas de minería de datos para identificar patrones de comportamientos relacionados con las acciones del estudiante con el EVA de*. (Tesis de Grado – Ingeniero en Sistemas Informáticos y Computación). Universidad Técnica Particular de Loja.

Veytia, M. (2013). *Cinco dimensiones para favorecer la apropiación tecnológica en estudiantes virtuales*. Actas del Encuentro Internacional de Educación a Distancia. Universidad de Guadalajara, 21.

Yannibelli, V., Godoy, D., & Amand, A. (2006). A genetic algorithm approach to recognize students' learning styles. *Interactive Learning Environment*, 14(1), 2006. doi:10.1080/10494820600733565

Yu, D., & Chen, X. (2006). Using bayesian networks to implement adaptivity in mobile learning. Proceedings of the International Conference on Semantics, Knowledge and Grid, 2. doi:10.1109/SKG.2006.107

Zapata-Ros, M. (2012). *Teorías y modelos sobre el aprendizaje en entornos conectados y ubicuos: bases para un nuevo modelo teórico a partir de una visión crítica del "conectivismo"*. Universidad de Alcalá.

Como citar este artigo (ABNT):

LÓPEZ-FAICAN, L. G.; CHAMBA-ERAS, L. A. Redes bayesianas para predecir el estilo de aprendizaje de estudiantes en entornos virtuales. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 107-115, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this article* (APA):**

López-Faican, L. G., & Chamba-Eras, L. A. (2014). Redes bayesianas para predecir el estilo de aprendizaje de estudiantes en entornos virtuales. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 107-115. Retrieved from http://www.atoz.ufpr.br

ISSN: 2237-826X
3(2), jul./dez. 2014

Modelo conceitual para jogos educativos digitais

A conceptual model for digital educational games

Rafael Feyh Jappur[1], Fernando Antonio Forcellini[2], Fernando Jose Spanhol[2]

[1] *Faculdade SENAC Florianópolis, Florianópolis, SC, Brasil*
[2] *Universidade Federal de Santa Catarina (UFSC), Florianópolis, SC, Brasil*

Autor para correspondência/*Corresponding author*: Rafael Feyh Jappur [rjappur@gmail.com]

Recebido/*Submitted*: 12 Nov. 2014
Aceito/*Approved*: 13 Dez. 2014

Resumo

Introdução: Apresenta o desenvolvimento de uma proposta de modelo conceitual para jogos educativos digitais no contexto da sala de aula, tendo como foco de análise a aprendizagem para hábitos sustentáveis de consumo e produção no ambiente residencial. Os modelos conceituais existentes estão mais ligados à criação e recentemente à avaliação dos jogos. Já no contexto da aplicação, a maioria dos modelos foca primordialmente na descrição dos processos de validação dos jogos do que na mediação da prática pedagógica em si. Portanto, a questão de pesquisa versa sobre em como criar, aplicar e avaliar jogos educativos digitais para o processo de ensino e aprendizagem em sala de aula, tendo como foco de aprendizagem a cultura da sustentabilidade no ambiente residencial.
Método: Realizou-se pesquisa bibliográfica e de campo como parte prática para a coleta de dados. Ressalta-se o uso do *Design Science Research Methodology* como procedimento metodológico para o desenvolvimento do modelo conceitual.
Resultados: O mesmo foi testado em um estudo piloto em duas turmas do programa jovem aprendiz do SENAC/SC para verificação de sua consistência.
Conclusão: Conclui-se com base nos resultados apresentados pela aplicação do piloto que o modelo conceitual proposto contribui para o processo de ensino e aprendizagem em sala de aula.

Palavras-chave: Jogos educativos digitais. Modelo conceitual. Processos de ensino e aprendizagem. Sustentabilidade.

Abstract

***Introduction**: It presents the development of a proposed conceptual model for digital educational games in the context of the classroom in which the analysis focused on learning for sustainable habits production and consumption in the residential environment. The existing conceptual models are more linked to the creation and recently with evaluation of the games. In the context of the implementation, most models focus more on the description of the games validation processes than in mediating the pedagogical practice itself. Therefore, the research question is about how to create, implement and evaluate digital educational games for teaching and learning process in the classroom, with the focus on learning the culture of sustainability in the residential environment.*
***Method**: As far as the methodological characterization is concerned, the investigation was supported by bibliographical and field research for data collection. It was emphasized the use of Design Science Research Methodology as methodological procedure for the development of the conceptual model.*
***Results**: The system has been tested in a pilot study in two groups of Young Apprentice Program at SENAC/SC to check its consistency.*
***Conclusion**: It was concluded based on the results produced by the application of the pilot that the proposed conceptual model contributes to the process of teaching and learning in the classroom.*

***Keywords**: Digital educational games.Conceptual model. Teaching and learning process. Sustainability.*

INTRODUÇÃO

O artigo aborda o contexto do uso de jogos educativos digitais no processo de ensino e aprendizagem em sala de aula, tendo com foco a aprendizagem para a cultura da sustentabilidade, especificamente para o ambiente residencial, caracterizado pelos hábitos de consumo e produção humana em edificações. Sendo estas vinculadas, segundo Jappur, Forcellini e Selig (2010), a vários aspectos ambientais gerados em edificações, tais como: a produção de resíduos sólidos e líquidos, e o consumo de energia e água.

Os jogos educativos digitais podem ser um agregador para o ensino de conteúdos e para o processo de aprendizagem em sala de aula, ligados à cultura da sustentabilidade. Todavia, constata-se – conforme apresentam Kirriemuir e Mcfarlane (2004); Balasubramanian e Wilson, (2006); Baek (2008), Echeverría et al. (2011), entre outros – que os professores ou mediadores possuem dificuldades para aplicar os jogos em sala de aula.

Balasubramanian e Wilson (2006) apontam que os jogos educativos digitais ainda são pouco utilizados na escola e, para muitos educadores, o desafio é encontrar e utilizar bons jogos como ferramenta de aprendizagem. Para os autores, os jogos digitais, no seu processo de criação têm sido usados de forma não articulada aos princípios e necessidades pedagógicas. Isto muitas vezes gera fragilidades ou dificuldades no que tange à seleção e aplicação pelos educadores, por não compreenderem o valor agregado à aprendizagem (Balasubramanian & Wilson, 2006).

Constata-se, ainda, que outras dificuldades para a aplicação dos jogos educativos digitais no contexto da sala de aula estão ligadas a falta dos princípios pedagógicos na criação dos jogos, seja pela falta de mediação na aplicação, ou até mesmo da avaliação da eficiência e eficácia do uso dos jogos em sala de aula (Baek, 2008; Balasubramanian & Wilson, 2006; Echeverría et al., 2011; Kirriemuir & Mcfarlane, 2004).

O problema de modelar jogos educativos digitais é salientado por alguns autores, tais como Hsiao (2007), Moreno-Ger, Burgos, Martínez-Ortiz, Sierra e Fernández-Manjóna (2008), Aleven, Myers, Easterday e Ogan, (2010), Echeverría et al. (2011), Villalta et al. (2011) e Savi (2011). Porém, verifica-se na literatura consultada que existem lacunas de como criar, aplicar e avaliar os jogos educativos digitais para o contexto do processo de ensino e aprendizagem em sala de aula. Também se constata que os modelos pesquisados estão mais ligados à criação e recentemente com a avaliação dos jogos. Já no contexto da aplicação, a maioria dos modelos pesquisados focam na descrição dos processos de validação dos jogos, do que sobre a mediação da prática pedagógica em si (Hsiao, 2007; Savi, 2011).

O presente artigo, portanto, pretende apresentar uma proposta de modelo conceitual que estruture de forma integrada os processos para a criação, aplicação e avaliação de jogos educativos digitais, no contexto da sala de aula, tendo como foco de análise a aprendizagem para hábitos sustentáveis de consumo e produção no ambiente residencial.

Na sequência, apresenta-se a metodologia utilizada e uma breve descrição teórica sobre os jogos educativos digitais, para em seguida apresentar a proposta de modelo conceitual e a aplicação do teste piloto, assim como as considerações finais.

JOGOS EDUCATIVOS DIGITAIS

De acordo com Huizinga (2007), o jogo é uma atividade realizada dentro de um limite de espaço e tempo, segundo regras livremente consentidas, dotado de um fim em si mesmo, e acompanhado de um sentimento de tensão e alegria com consciência da diferença da vida cotidiana.

Criar um jogo que seja divertido e educativo (digital ou não digital) é um desafio significativo. No entanto, muito pouco tem sido escrito sobre como projetar jogos educativos eficazes (Aleven, Myers, Easterday, & Ogan, 2010; Echeverría et al., 2012; Schell, 2008).

Os jogos digitais quando preparados para o contexto educacional podem receber diferentes nomenclaturas (Savi, 2011). Neste artigo se utilizará – preponderantemente – a denominação de jogos educativos digitais, ainda que não de forma limitada.

A produção científica sobre jogos educativos digitais vem crescendo nas últimas duas décadas com pesquisas sobre o potencial destes jogos para o processo de ensino e aprendizagem (Hsiao, 2007; Moreno-Ger et al., 2008; Villalta et al., 2011).

O uso destes jogos como ferramenta educacional está lentamente se tornando uma prática aceita em ambientes de aprendizagem. Eles apresentam diversas possibilidades para o desenvolvimento do conhecimento e também ajudam a melhorar o processo de ensino e aprendizagem nas escolas (Villalta et al., 2011).

Contudo, segundo Balasubramanian e Wilson (2006), os jogos educativos digitais ainda são pouco utilizados e para muitos educadores encontrar e utilizar bons jogos são um desafio. Isso ocorre, em boa parte, porque muitos dos jogos têm feito uso limitado dos princípios pedagógicos e acabam sendo ignorados pelos educadores por agregarem pouco valor às aulas.

Moreno-Ger et al. (2008) abordam que avaliação da aprendizagem dos alunos é outro impeditivo para o uso dos jogos educativos digitais pelos professores. Segundo os autores, é necessário verificar se os alunos estão atingindo os objetivos pedagógicos propostos e fornecer algum tipo de *feedback*, como por exemplo, relatórios informando ao aluno o seu desempenho, tempo, erros cometidos, entre outras informações.

Além disso, várias questões como a relevância para currículo, precisão de conteúdos e compatibilidade da duração dos jogos com o horário de uso dos laboratórios de informática, têm impedido que os jogos educativos digitais se tornem uma atividade predominante nas instituições de ensino (Kirriemuir & Mcfarlane, 2004).

Contudo, a proposta de modelo conceitual aqui apresentada foi desenvolvida, em especial, com base nos modelos de Klein, Nir-Gal, Darom (2000), Gomes (2001), Aleven et al. (2010), Echeverría et al. (2011), Villalta et al. (2011), Savi (2011), que se relacionam com a criação, aplicação e avaliação de jogos educativos digitais para o processo de ensino e aprendizagem em sala de aula (Quadro 1). Ressalta-se que os autores pesquisados não fazem uma distinção clara sobre a criação, aplicação e avaliação, como proposto neste modelo, porém, apresentam diretrizes a serem consideradas.

Modelos de Jogos Educativos Digitais

Com base na revisão bibliográfica e no estudo mais aprofundado dos modelos selecionados nesta pesquisa (Quadro 1) nos quais se constatou um consenso entre os autores – segundo os quais, para um bom jogo educativo digital é importante que haja uma proposta educacional e lúdica. Para Echeverría et al. (2011) e Villalta et al. (2011), a dimensão lúdica se refere aos elementos do jogo, respeitando, todavia, as questões impostas pela dimensão educativa ou pedagógica. Aleven et al. (2010) afirmam que um bom jogo educativo digital deve possuir um projeto pedagógico e um projeto de jogo divertido. Nesse sentido, elencaram-se as propostas pedagógicas e lúdicas, apresentadas pelos autores dos modelos identificados e selecionados nesta pesquisa (Quadro 1).

Modelos Conceituais	Propostas Pedagógicas	Propostas Lúdicas
Mayer, Moreno (2002)	Taxionomia de Bloom Revisada (Anderson & Krathwohl, 2001).	Não apresenta.
Echeverría et al. (2011)	Taxionomia de Bloom revisada (Anderson & Krathwohl, 2001).	Mecânica, história, estética, tecnologia (Schell, 2008).
Aleven et al. (2010)	Taxionomia de Bloom revisada /Princípios do design instrucional (Anderson & Sosniak, 1995; Mayer & Moreno, 2003).	Mecânica, dinâmica, estética (Hunicke, Leblanc, & Zubek, 2004).
Villalta et al. (2011)	Não apresentam proposta única, mas devem estar alinhadas com as estratégias instrucionais.	Apresenta um guia de avaliação, que inclui: mecânica, progressão, metodologia, colaboração, informação na tela, holismo.
Savi (2011)	Três primeiros níveis da taxionomia de Bloom (conhecimento, compreensão e aplicação) (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956); aprendizagem de curto e longo prazo (Moody & Sindre, 2003).	Percepção dos alunos – Nível 1 Kirkpatrick, 1994); Modelo A, R, C, S – Atenção, relevância, competência e satisfação (Keller, 1987); Imersão, interação social, desafio, divertimento e competência (Gámez, 2009; Poels, Kort, & Ijsselsteijn, 2007; Sweetser & Wyeth, 2005; Takatalo, Häkkinen, Kaistinen, & Nyman, 2010).
Klein et al. (2000)	Feuerstein – Seis critérios de mediação: intencionalidade e reciprocidade, transcendência, significado, competência e autorregulação (Feuerstein, 1980; Klein, 1996).	Não apresenta.
Campos e Macedo (2011)	Feuerstein – Três critérios de mediação: intencionalidade e reciprocidade, transcendência, significado (Feuerstein, 1980).	Não apresenta.
Gomes (2001)	Mediação – todos os critérios de mediação (12 critérios); Processo cognitivo; mecanismos de aprendizagem (Feuerstein, 1980).	Não apresenta.

Quadro 1. Propostas pedagógicas e lúdicas.
Fonte: Elaboração própria.

No quadro 1 pode-se verificar, no que se refere às propostas pedagógicas, que os trabalhos de Mayer e Moreno (2002), Echeverría et al. (2011) e Aleven et al. (2010) utilizam a taxionomia revisada de Bloom (Anderson & Krathwohl, 2001). Já Savi (2011) aplicou os três primeiros níveis da taxionomia de Bloom de 1954 na versão original (conhecimento, compreensão e aplicação) (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1977), e o modelo de aprendizagem de curto e longo prazo de Moody e Sindre (2003). Aleven et al. (2010) incorporaram, em seu estudo, os princípios do design instrucional (Anderson & Sosniak, 1995; Mayer & Moreno, 2002; Moreno-Ger, 2008) e os objetivos pedagógicos da taxionomia revisada de Bloom (Anderson & Krathwohl, 2001).

Villalta et al. (2011), assim como Aleven et al. (2010), também apresentam a necessidade de os jogos estarem alinhados com as estratégias instrucionais e sistematizam seis diretrizes para um projeto de jogo na modalidade Classroom Multiplayer Presential Game (CMPG). Contudo, os autores argumentam que outros métodos e diretrizes podem ser integrados, a exemplo da taxonomia revisada de Bloom (Anderson & Krathwohl, 2001), como uma estrutura que ajuda na definição dos objetivos da aprendizagem. Os autores citam o modelo de Echeverría et al. (2011) como exemplo.

Por sua vez, Klein et al. (2000), Campos e Macedo (2011) e Gomes (2001) apresentam como proposta pedagógica os critérios da Experiência da Aprendizagem Mediada (EAM) de Feuerstein (1980), principalmente os três critérios fundamentais para uma EAM, a saber: intencionalidade e reciprocidade, transcendência, significado. Os autores salientam que os outros critérios da proposta de Feuerstein (1980) também são importantes, pois agregam e motivam a realização ou efetivação dos critérios fundamentais ou principais, de modo a enriquecer o processo de mediação da aprendizagem.

Gomes (2001) descreve doze critérios para a aprendizagem mediada, estando estes em conformidade com os critérios de mediação de Feuerstein (1980). Além dos critérios para EAM, Gomes (2001) também descreve um processo cognitivo com sete parâmetros de análise, adaptados do mapa cognitivo de Feuerstein (1980), e os mecanismos de aprendizagem, compostos por oito estratégias para as análises pedagógicas. Campos e Macedo (2011) demonstram um conjunto de indicadores para os três critérios principais da EAM, com foco na mediação de jogos em sala de aula.

No que tange às propostas lúdicas, constatou-se que os modelos de Mayer e Moreno (2002), Klein et al. (2000), Campos e Macedo (2011) e Gomes (2001) não apresentaram propostas lúdicas específicas, muito embora tenham citado a importância dos aspectos lúdicos para os jogos.

Já Echeverría et al. (2011), baseados nos estudos de Schell (2008), descrevem quatro elementos lúdicos para um jogo educativo digital: a mecânica (procedimentos, regras, os objetivos etc.); a história (roteiro de eventos); a estética (design gráfico, cores, música, efeitos sonoros etc.), e a tecnologia (dispositivos, displays, plataformas etc.).

Aleven et al. (2010) utilizaram os estudos de Hunicke, Leblanc, e Zubek (2004) para descrever os três componentes básicos para o desenvolvimento de jogos: mecânica (materiais, regras, objetivos, movimentos básicos e opções de controle para os jogadores); dinâmica (comportamentos que resultam da aplicação da mecânica do jogo); e estética (sensação, fantasia, narrativa, desafio, camaradagem, descoberta, expressão e submissão).

Villalta et al. (2011) propõem seis categorias com diretrizes para superar problemas apresentados por jogos, sendo estes: mecânica (interatividade e orientação, mecânica ligada aos objetivos da aprendizagem); progressão do jogo (narrativa, aumento gradual da dificuldade); metodologia (o professor é o mediador); colaboração (interações, mecânica vinculada à colaboração); informação na tela (distribuição espacial, elementos reconhecíveis, linguagem acessível e cuidados com excesso de informações); e holismo (guia de ação que inclui os aspectos educativos e lúdicos).

Savi (2011) não aprofunda a questão relacionada aos elementos de criação de jogos educativos digitais, mas destaca o design instrucional como um guia mais utilizado pelos desenvolvedores. Em seu trabalho, o autor relaciona alguns requisitos para avaliar a percepção dos alunos em relação à qualidade dos jogos. Para tal, o autor utiliza o nível 1 (reação) do modelo de Kirkpatrick (1994). Sendo que, para avaliar a motivação dos usuários ou alunos em jogar, é utilizado o modelo A, R, C, S (atenção, relevância, confiança e satisfação) de Keller (1987); e para avaliar a experiência do usuário com os jogos, são utilizados os seguintes parâmetros: imersão, interação social, desafio, divertimento e competência (Gámez, 2009; Ijsselsteijn, 2007; Poels & Kort, 2007; Sweetser & Wyeth, 2005; Takatalo, Häkkinen, Kaistinen, & Nyman, 2010).

METODOLOGIA

A natureza desta pesquisa, em acordo com Silva e Menezes (2005), possui mais similaridades com a pesquisa aplicada, pois objetiva gerar conhecimentos para aplicação prática dirigida à solução de problemas vivenciados no ambiente residencial.

Quanto à forma de abordagem, o presente trabalho representa uma pesquisa qualitativa, pois suas características principais coincidem com as recomendações feitas por vários autores como Silva e Menezes (2005), Vergara (2009) e Gil (2010).

Quanto aos objetivos, situa-se em três categorias: exploratória, descritiva e explicativa. Segundo Vergara (2009) e Gil (2010), os tipos de pesquisa não são mutuamente exclusivos, o que permite classificá-lo nestes três tipos.

Os procedimentos metodológicos utilizados foram preponderantemente a pesquisa bibliográfica e de campo, com a utilização do *Design Science Research Methodology* (DSRM) para o desenvolvimento do modelo conceitual.

A Metodologia de Pesquisa da Ciência do Design (*Design Science Research Methodology* – DSRM) busca preencher a falta de uma metodologia para servir como modelo aceito e válido para o desenvolvimento de artefatos para a ciência da informação. A DSRM incorpora princípios, práticas, e procedimentos necessários para realizar tais pesquisas e apresenta seis etapas/atividades que seguem uma sequência nominal ou procedural (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007):

a) identificação do problema e motivação;
b) definição dos objetivos;
c) design e desenvolvimento;
d) demonstração;
e) avaliação; e
f) comunicação.

Embora as atividades sejam apresentadas de forma sequencial, não impõem uma ordem exata para o início da pesquisa. A metodologia apresenta como principal diferencial o fato de ter sido concebida de modo a possibilitar que o início da pesquisa possa ocorrer em diferentes etapas, considerando o foco que se pretenda dar à investigação. Segundo Peffers et al. (2007) existem quatro pontos distintos para o início de uma pesquisa, Sendo estes: início gerado por problema; início gerado por solução; início gerado por projeto/design e desenvolvimento; e início gerado por cliente/contexto. Todavia, se a pesquisa iniciar de forma não sequencial, as atividades anteriores não desenvolvidas devem ser identificadas e alinhadas com o tema da pesquisa (Peffers et al., 2007).

O início desta pesquisa, conforme orienta o DSRM, foi o projeto e desenvolvimento do jogo educativo digital denominado de Simulador Ambiental (SA). A partir da experiência realizada com a criação do jogo foi concebido o modelo conceitual pretendido. O modelo conceitual foi testado em um estudo piloto em duas turmas do programa jovem aprendiz do SENAC/SC para verificação de sua consistência.

PROPOSTA DE MODELO CONCEITUAL PARA JOGOS EDUCATIVOS DIGITAIS

A proposta de modelo conceitual visa estabelecer uma estrutura de processos para a criação, aplicação e avaliação de jogos educativos digitais para o processo de ensino e aprendizagem em sala de aula (Figura 1).

O processo de criação envolve uma proposta pedagógica e lúdica para que os desenvolvedores criem bons jogos educativos digitais. A estrutura pedagógica propõe a utilização da taxionomia Revisada de Bloom conforme descrevem Anderson, Krathwohl (2001), Krathwohl (2002), Aleven et al. (2010), Villalta et al. (2011) e Echeverría et al. (2011) para a definição dos objetivos educacionais do jogo e de um guia de recomendações com os critérios de mediação de Feuerstein (1980), em conformidade com The International Center for the

Enhancement of Learning Potential [ICELP] (2012) e Gomes (2001), adaptados para o contexto de criação dos jogos educativos digitais.

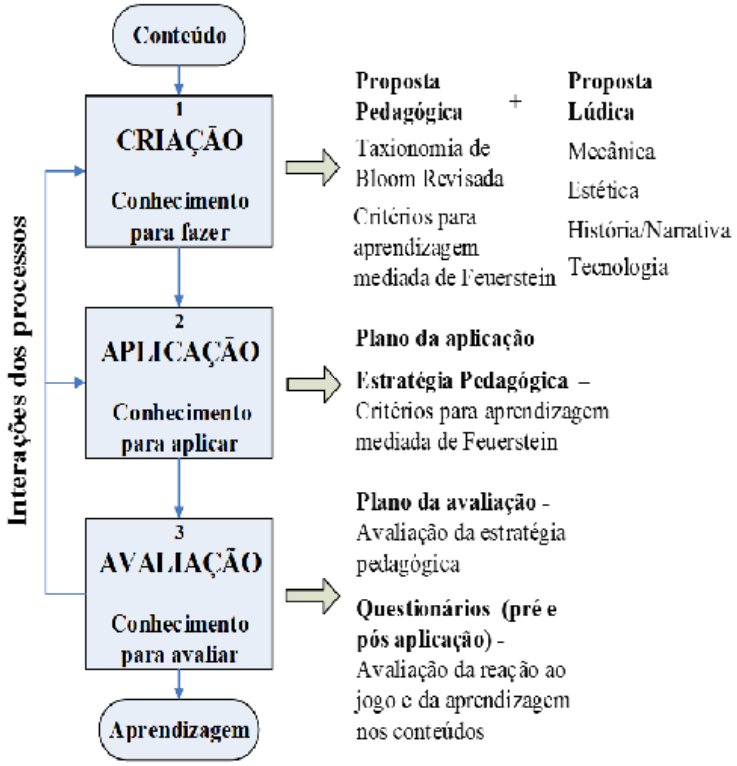


Figura 1. Modelo conceitual para jogos educativos digitais.
Fonte: Elaboração própria.

Trata-se de uma proposta que visa incrementar o conhecimento didático-pedagógico dos desenvolvedores, com a intenção de apoiar os processos do design instrucional do jogo, para que criem uma mecânica de jogo favorável à mediação da aprendizagem do jogador em relação a um conteúdo específico.

Já a proposta lúdica para a criação dos jogos educativos digitais busca estabelecer um guia de recomendações com os elementos lúdicos a serem incorporados nos jogos. O guia lúdico é composto por quatro elementos principais, que são: mecânica; estética; história/narrativa; e tecnologia. A seleção destes elementos se basea nos trabalhos de Hunicke et al. 2004, Schell (2008), Aleven et al. (2010), Echeverría et al. (2011), Villalta et al. (2011).

O processo de aplicação do modelo conceitual estabelece uma estrutura para a mediação do processo de ensino e aprendizagem de conteúdos. Para tanto, baseado em Schön (2000), é utilizado um plano para a aplicação do jogo, que apresenta o conhecimento do mediador na ação pedagógica, a reflexão para a ação pedagógica e a reflexão na ação pedagógica. As estratégias pedagógicas devem ser descritas em conformidade com critérios da aprendizagem mediada de Feuerstein (1980), em acordo com Klein, Nir-Gal e Darom (2000), Gomes (2001, 2002), Mentis (2011), Souza, Machado, e Depresbiteris (2004), Campos e Macedo (2011) e ICELP (2012); sendo estes os mesmos apresentados para a criação, porém voltados para mediação do aprendizado do conteúdo do jogo durante a prática pedagógica.

O processo de avaliação do modelo conceitual esta estruturado por um plano de avaliação e por dois tipos de avaliações. Descrições e narrações, por meio da observação participante, para avaliar as estratégias pedagógicas (critérios de mediação de Feuerstein) utilizadas na prática pedagógica e a aplicação de três questionários para dar suporte para a coleta de dados para avaliação das variáveis reação e aprendizagem. O primeiro questionário é aplicado antes do jogo (pré-teste) para capturar a percepção dos jogadores a respeito de seu conhecimento no conteúdo do jogo; já o segundo e o terceiro deverão ser aplicados após o uso do jogo (pós-testes). O segundo questionário visa conhecer a reação do jogador sobre a sua experiência e motivação com o jogo e se os objetivos educacionais apregoados para o mesmo foram atingidos; e o terceiro questionário é para verificar, em comparação com o primeiro, se o jogo possibilitou melhoria do aprendizado no conteúdo do jogo.

O quadro 2 apresenta o referencial teórico do modelo conceitual que fundamenta o processo de avaliação dos jogos educativos digitais e das estratégias pedagógicas a serem utilizadas na prática pedagógica.

Características	Referenciais teóricos
Avaliações baseadas na percepção dos alunos.	Nível 1 e 2 do modelo de avaliação de treinamentos de Kirkpatrick (1994).
Avaliação do nível de motivação dos jogos.	Modelo motivacional ARCS, desenvolvido por Keller (1987) e Savi (2011).
Avaliação da experiência do usuário com os jogos.	Modelos para avaliação da experiência do usuário em jogos (Gámez, 2009; Poels, Kort, & Ijsselsteijn, 2007; Savi, 2011, Sweetser & Wyeth, 2005; Takatalo et al., 2010).
Avaliação dos objetivos educacionais dos jogos.	Taxionomia revisada de Bloom (Anderson, Krathwohl, 2001; Bloom, 1956; Churches, 2009; Echeverría et al., 2011; Krathwohl, 2002).
Avaliação da aprendizagem mediada com a prática pedagógica.	Critérios para a aprendizagem mediada de Feuerstein (Campos & Macedo, 2011; Feuerstein, 1980; Gomes, 2001, 2002; Klein et al., 2000; Mentis , 2011; Souza et al., 2004; ICELP, 2012).

Quadro 2. Referenciais teóricos do processo de avaliação do modelo conceitual.
Fonte: Elaboração própria.

O quadro 3 apresenta a estrutura de avaliação do modelo conceitual. Sendo esta dividida em duas variáveis, cinco construtos e respectivas dimensões. Salienta-se que as dimensões do construto aprendizagem podem variar em função do conteúdo, temas e dos critérios da aprendizagem mediada utilizados na prática pedagógica.

Variáveis	Construtos	Dimensões
Reação	**Motivação**	Atenção
		Relevância
		Confiança
		Satisfação
	Experiência do usuário	Imersão
		Desafio
		Competência
		Divertimento
		Controle
		Interação social
	Objetivos educacionais do jogo	Lembrar
		Entender
		Aplicar
		Analisar
		Avaliar
		Criar
Aprendizagem	**Conhecimentos no conteúdo (pré e pós-teste)**	Depende do conteúdo e dos temas apresentados pelo jogo
	Estratégia pedagógica (Critérios para a aprendizagem mediada de Feuerstein)	Intencionalidade e reciprocidade
		Significado
		Transcendência
		Outros critérios da aprendizagem mediada de Feuerstein utilizados e observados na prática pedagógica (depende do uso dos outros critérios)

Quadro 3. Estrutura do modelo conceitual para o processo de avaliação.
Fonte: Elaboração própria.

Portanto, o quadro 3 apresenta o resumo da estrutura do modelo conceitual para a avaliação da prática pedagógica e do jogo educativo digital utilizado.

TESTE DA PROPOSTA DE MODELO CONCEITUAL PARA JOGOS EDUCATIVOS DIGITAIS

O teste piloto foi conduzido seguindo as seguintes etapas:

a) descrição do plano da aplicação e avaliação;
b) preparação dos materiais; apresentação do jogo e explicação dos questionários;
c) aplicação do questionário (pré-teste);
d) execução do jogo;
e) aplicação dos questionários (pós-testes); e
f) análise dos dados.

O teste piloto foi realizado na Faculdade SENAC Florianópolis, em parceria com o Projeto Recicle Ideias do SENAC Santa Catarina. A aplicação foi operacionalizada no laboratório de informática da Faculdade com duas turmas do Projeto Jovem Aprendiz (22 alunos – turma 804 e 16 alunos – turma 160), em períodos distintos. A média de idade dos alunos era entre 14 a 16 anos (Figura 2).

Figura 2. Aplicação turma 804.
Fonte: Elaboração própria.

O jogo educativo digital utilizado nas duas práticas pedagógicas foi o Simulador Ambiental (SA) (acesso http://www.meensina.org.br/site/simulador/) (Figura 3). Este jogo foi concebido junto ao desenvolvimento da proposta de modelo conceitual. O jogo visa à educação das pessoas para hábitos de consumo e produção mais sustentáveis no ambiente residencial. O contexto do SA está relacionado com a identificação e minimização dos impactos ambientais promovidos pelas ações humanas no ambiente residencial (Arbex, Jappur, Selig, & Varvakis, 2012).

Figura 3. Ambiente da área externa da residência.
Fonte: Me Ensina, (2013).

A proposta de jogo foi desenvolver alternativas para que o usuário pudesse navegar entre as telas dos ambientes de uma casa, facilitando a tomada de decisão em cada ambiente. O público alvo para qual o jogo foi projetado e desenvolvido é para o infanto-juvenil, não impossibilitando a utilização do jogo para outros públicos (Arbex et al., 2012).

Os itens dos questionários foram estabelecidos em acordo com o conteúdo do SA. O primeiro (pré-teste) e o terceiro (pós-teste) questionário foram elaborados com quatro questões para avaliar aprendizado dos jogadores no conteúdo. Já para o segundo questionário (pós-teste) elaboraram-se vinte questões para avaliar a reação dos alunos em relação à motivação, a experiência que o jogo proporcionou e se os objetivos pedagógicos foram alcançados.

Em referência aos resultados do aprendizado dos jogadores no conteúdo, constatou-se que houve melhoria no aprendizado dos alunos. Na primeira questão o aumento foi de aproximadamente em 13% e 27% nas turmas dos jovens aprendizes (804 e 106). Na segunda questão o aumento foi de aproximadamente em 12% e 18% nas turmas dos jovens aprendizes (804 e 106). A terceira questão apresentou aumento em 5% e igualmente em 5% nas turmas dos jovens aprendizes (804 e 106). A quarta questão apresentou aumento em 31% e 9% nas turmas dos jovens aprendizes (804 e 106).

O construto motivação, em praticamente todas as questões, obteve percentual acima de 70% de concordância, demonstrando que jogo teve um efeito positivo na motivação dos jogadores.

Os resultados do construto experiência do usuário no jogo proporcionaram reações positiva aos alunos, com destaque para as dimensões imersão, divertimento, competência e controle. Todavia, as dimensões interação social e desafio não foram tão bem avaliadas, demonstrando possibilidades de melhorias do jogo nestas dimensões. A dimensão interação social não foi tão bem avaliada, pois o jogo é individual, mesmo que os alunos compartilhem ideias durante o jogo. Já a dimensão desafio não teve boa pontuação, talvez pela linguagem e descrição das opções de escolha das questões arguidas nos ambientes do jogo. Porém, verifica-se uma possível contradição, pois 50% dos alunos da turma 804, e 59% dos alunos da turma 106 responderam que não concordam que obtiveram desempenho ótimo.

Os resultados apresentados dos objetivos educacionais atribuídos para o SA, segundo a percepção de mais de 70% dos alunos, demonstram que os mesmos concordam que o jogo ajuda a lembrar, entender, aplicar, analisar, avaliar e a criar conhecimento relacionado ao conteúdo do SA.

A estratégia pedagógica planejada para a aplicação do SA foi positiva, todos os critérios para a aprendizagem mediada foram trabalhados, conforme segue:

a) Intencionalidade e reciprocidade – As duas práticas pedagógicas sensibilizaram os participantes da importância dos hábitos sustentáveis de consumo e produção no ambiente residencial. Alguns problemas verificados como a gestão do tempo, velocidade da internet e o entendimento dos participantes quanto aos passos a serem seguidos não interferiram na reciprocidade das pessoas;
b) Significado – os participantes demonstraram terem entendido à razão da atividade de aprendizagem e a importância das tarefas. O planejamento das atividades valorizou as atitudes e as habilidades dos alunos, tais como: cooperação, argumentação, disciplina, respeito aos colegas, etc; e
c) Transcendência – A prática pedagógica com o uso do SA possibilitou aos participantes a fazerem analogias e debates de seus hábitos, desencadeando intenções para atitudes mais sustentáveis em suas residências.

Ressalta-se que os demais critérios da aprendizagem mediada também aconteceram, mesmo que de forma não planejada, foram utilizados em algum momento nas duas práticas pedagógicas.

CONCLUSÕES

O estudo piloto apresentou dados preliminares positivos para a adequabilidade da sistemática do modelo conceitual. O envolvimento dos alunos na prática pedagógica, com o uso do SA, permitiu que eles efetivamente aprendessem hábitos sustentáveis de consumo e produção para suas residências. Todavia, estudos futuros devem ser empreendidos para aprofundamento do mesmo.

Contudo, constataram-se – durante a realização do piloto – algumas limitações referentes à gestão do tempo dos jogadores e dos recursos tecnológicos. O cronograma da execução da prática não foi totalmente conforme, pois alguns participantes levaram muito tempo para realizar o cadastro, devido à lentidão dos sistemas (cadastro dos participantes e envio dos logins e senhas para os e-mails pessoais), ou por terem esquecido a senha do e-mail pessoal. Outros levaram mais de dez minutos para responder aos questionários, por dificuldades no entendimento de como proceder.

Todavia, o presente trabalho oferece uma variedade de temas a serem aprofundados; no entanto, sugere-se para pesquisas futuras a aplicação do modelo conceitual em outros contextos, com temáticas e conteúdos diferentes da trabalhada neste artigo. Como por exemplo, jogos para aprendizagem em geografia, matemática, português, entre outros.

Durante a pesquisa bibliográfica várias abordagens pedagógicas foram identificadas, sendo sugeridos para trabalhos futuros outros estilos e teorias da aprendizagem. Referente a proposta lúdica, sugere-se aprimorar suas características, como por exemplo, o elemento estético do processo de criação de jogos, em relação à persuasão que determinados tipos de cores e sons provocam nos jogadores, motivando-os a jogar ou a ficar atentos a uma determinada situação.

Considera-se que há um amplo leque de possibilidades de novas aprendizagens mediadas por jogos educativos digitais, mas, destaca-se também que a figura central do professor-mediador deve ser incluída nesta nova forma didática de ensino e aprendizagem.

REFERÊNCIAS

Aleven, V., Myers, E., Easterday, M., & Ogan, A. (2010). Toward a framework for the analysis and design of educational games. *IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*, 69-76. 2010. Retirado de http://delta.northwestern.edu/wordpress/wp-content/uploads/2013/11/Aleven-2010-Toward-a-framework-for-the-analysis-and-design-of-educational-games.pdf

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. New York: Longman.

Anderson, L. W., & Sosniak, L. A. (Eds.). (1995). *Bloom's taxonomy: A forty year retrospective*. Chicago: University of Chicago.

Arbex, D., Jappur, R., Selig, P., & Varvakis, G. (2012). Ergonomic aspects simulation digital online: An educational game proposal to promote environmental education. *Work: a journal of prevention, assessment and rehabilitation*, 41, 6011- 6015. doi:10.3233/WOR-2012-1052-6011

Baek, Y. (2008). What hinders teachers in using computer and video games in the classroom? Exploring factors inhibiting the uptake of computer and video games. *Cyberpsychology & Behavior*, 11(6), 665–671 doi:10.1089/cpb.2008.0127

Balasubramanian, N., & Wilson, B. G. (2006). Games and simulations. *Society for Information Technology and Teacher Education International Conference*. Retirado de http://site.aace.org/pubs/foresite/GamesAndSimulations1.pdf

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (Eds.). (1956). *Taxonomy of educational objectives* (V. 1). New York: David Mckay.

Bloom, B. S., Engelhart M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1977). *Taxionomia de objetivos educacionais: Domínio cognitivo* (6. Ed.). Porto Alegre: Globo.

Campos, M. C. R. M., & Macedo, L. de. (2011). Desenvolvimento da função mediadora do professor em oficinas de jogos. *Psicologia Escolar e Educacional*, 15(2). doi:10.1590/S1413-85572011000200003

Churches, A. (2009). *Bloom's digital taxonomy: Educational origami*. Retirado de http://edorigami.wikispaces.com/Bloom%27s+and+ICT+tools

Echeverría, A., García-Campo, C., Nussbaum, M., Gil, F., Villalta, M., Améstica, M., & Echeverría, S. (2011). A framework for the design and integration of collaborative classroom games. *Computers & Education*, 57(1), 1127–1136. doi:10.1016/j.compedu.2010.12.010

Feuerstein, R. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability*. Glenview: Scott, Foresman and Company.

Gámez, E. H. C. (2009). *On the core elements of the experience of playing video games: Studying the gaming experience*. LAP Lambert Academic Publishing, 2009.

Gil, A. (2010). *Como elaborar projetos de pesquisa* (5. Ed.). São Paulo: Atlas.

Gomes, C. M. A. (2001). *Em busca de um modelo psico-educativo para a avaliação de softwares educacionais*. Dissertação de Mestrado, Universidade Federal de Santa Catarina, Brasil. Retirado de http://repositorio.ufsc.br/xmlui/handle/123456789/79700

Gomes. C. M. A. (2002). *Feuerstein e a construção mediada do conhecimento*. Porto Alegre: Artmed.

Hsiao, H. C. (2007). A brief review of digital games and learning. *IEEE International Workshop on Digital Game and Intelligent Toy Enhanced Learning*, 124-129. doi:10.1109/DIGITEL.2007.3

Huizinga, J. (2007). *Homo ludens: O jogo como elemento da cultura*. São Paulo: Perspectiva.

Hunicke, R., Leblanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. *AAAI Workshop Technical Report WS-04-04*, 1-5. Retirado de http://www.cs.northwestern.edu/~hunicke/MDA.pdf

Jappur, R. F., Forcellini, F. A., & Selig, P. M. (2010). Indicadores de gestão do conhecimento para sustentabilidade em edificações. *9. Congresso Brasileiro de Gestão do Conhecimento*.

Keller, J. M. (1987). Development and use of the ARCS model of motivational design. *Journal of Instructional Development*, 10(3), 2–10.

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels*. Berrett-Koehler Publishers.

Kirriemuir, J., & Mcfarlane, A. (2004). *Literature review in games and learning*. Bristol: Futurelab. Retirado de http://www.mendeley.com/research/literature-review-in-games-and-learning/

Klein, S. P., Nir-Gal, O., & Darom, E. (2000). The use of computers in kindergarten, with or without adult mediation; effects on children's cognitive performance and behavior. *Computers in Human Behavior*, 16(6), 591-608. doi:10.1016/S0747-5632(00)00027-3

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory into Practice*, 41(4), 212-264. Retirado de http://www.celt.iastate.edu/teaching/RevisedBlooms1.html

Mayer, R. E., & Moreno, R. (2002). Aids to computer-based multimedia learning. *Learning and Instruction*, 12(1), 107-119. doi:10.1016/S0959-4752(01)00018-4

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia. *Learning Educational Psychologist*, 38(1). doi:10.1207/S15326985EP3801_6

Me Ensina. (2013). *Simulador ambiental*. Retirado de http://www.meensina.org.br/site/simulador/

Mentis, M. (2011). *Aprendizagem mediada dentro e fora de sala de aula*. São Paulo: SENAC.

Moody, D., & Sindre, G. (2003). Evaluating the effectiveness of learning interventions: an information systems case study. *Proceedings of ECIS*, Paper 80. Retirado de http://aisel.aisnet.org/ecis2003/80

Moreno-Ger, P., Burgos, D., Martínez-Ortiz, I., Sierra, J. L., & Fernández-Manjóna, B. (2008). Educational game design for online education. *Computers in Human Behavior*, 24(6), 2530–2540. doi:10.1016/j.chb.2008.03.012

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45-77. doi:10.2753/MIS0742-1222240302

Poels, K., Kort, Y. D., & Ijsselsteijn, W. (2007). "It is always a lot of fun!": exploring dimensions of digital game experience using focus group methodology. *ACM Conference on Future Play*, 83-89.

Savi, R. (2011). *Avaliação de jogos voltados para a disseminação do conhecimento*. Tese de Doutorado, Universidade Federal de Santa Catarina, Brasil. Retirado de http://www.gqs.ufsc.br/wp-content/uploads/2011/11/RafaelSavi_teseFinal_A5.pdf

Schell, J. (2008). *The art of game design: A book of lenses*. San Francisco: Morgan Kaufmann.

Schön, D. A. (2000). *Educando o professional reflexivo: um novo design para o ensino e a aprendizagem*. Artmed: Porto Alegre.

Silva E. L., & Meneses, E. M. (2005). *Metodologia da pesquisa e elaboração de dissertação* (4. Ed.). Florianópolis: PPGEP/UFSC.

Souza, A. M. M., & Machado, O. T. M., & Depresbiteris, L. (2004). *A mediação como princípio deducional: Bases teóricas das abordagens de Reuven Feuerstein*. São Paulo: Senac.

Sweetser, P., & Wyeth, P. (2005). GameFlow: a model for evaluating player enjoyment in games. *Computer Entertainment*, 3(3), 1-24.

Takatalo, J., Häkkinen, J., Kaistinen, J., & Nyman, G. (2010). Presence, involvement, and flow in digital games. In R. Bernhaupt (Ed.). *Evaluating user experience in games: Concepts and methods*. London: Springer.

The International Center for the Enhancement of Learning Potential (ICELP). (2012). *Research: Basic theory*. Retirado de http://www.icelp.org/asp/Basic_Theory.shtm

Vergara, S. C. (2008). *Projetos e relatórios de pesquisa em administração* (13. Ed.) São Paulo: Atlas.

Villalta. M., Gajardo, I., Nussbaum. M., Andreu, J. J., Echeverría, A., & Plass, J. L. (2011). Design guidelines for classroom multiplayer presential games (CMPG). *Computers & Education*, 57(3), 2039–2053. doi:10.1016/j.compedu.2011.05.003

Como citar este artigo (ABNT):

JAPPUR, R. F.; FORCELLINI, F. A.; SPANHOL, F. J. Modelo conceitual para jogos educativos digitais. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 116-127, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this article* (APA):**

Jappur, R. R., Forcellini, F. A., & Spanhol, F. J. (2014). Modelo conceitual para jogos educativos digitais. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 116-127. Retrieved from http://www.atoz.ufpr.br

Metodologia Design Thinking no projeto de software para mobilidade urbana: relato de aplicação

Applying design thinking in an urban mobility software project: an experience report

Adailton Magalhães Lima[1], Antonia Tamires Alves[2], Anderson Jorge Serra da Costa[1], Ernani de Oliveira Sales[1]

[1] *Universidade Federal do Pará (UFPA), Belém, PA, Brasil*
[2] *Universidade Federal da Bahia (UFBA), Salvador, BA, Brasil*

Autor para correspondência/*Corresponding author*: Adailton Magalhães Lima [adailton@ufpa.br]

Agradecimentos/*Acknowlegments*: Ao Coordenador e agentes da Secretaria Municipal de Transporte e Trânsito de Castanhal (SEMUTRAN), que colaboraram com insumos para a pesquisa realizada. À população entrevistada. A todos que auxiliaram o desenvolvimento deste projeto, direta ou indiretamente.

Recebido/*Submitted*: 30 Set. 2014
Aceito/*Approved*: 13 Dez. 2014

Resumo

Introdução: Apresenta um relato da condução de um estudo onde se aplicou a metodologia Design Thinking no contexto do projeto de uma solução de software ao contexto de mobilidade urbana.
Método: Dentre as técnicas sugeridas pela metodologia Design Thinking, este trabalho utilizou entrevistas, surveys, brainstorming e prototipação. O foco do estudo foram os transportes públicos e os processos intrínsecos a esse contexto.
Resultados: Tem-se um projeto de solução aderente aos problemas identificados nas pesquisas de campo, e também uma avaliação conduzida com os usuários alvo sobre o projeto de solução proposto.
Conclusão: A aplicação de Design Thinking pode ser eficiente no desenvolvimento de software, pois se tem nessa metodologia um direcionamento na elaboração de projetos centrada na construção de produtos voltados para atender as necessidades de seus usuários, ou seja, projeta-se a solução com base no quão usual o produto será para o usuário final.

Palavras-chave: Mobilidade Urbana. Gestão de transportes públicos. Design Thinking. Projeto de Software.

Abstract

***Introduction**: This paper presents a report about the conduction of a study where it was applied the Design Thinking methodology in the context of designing a software solution to urban mobility.*
***Method**: Among the techniques suggested by Design Thinking methodology, this study applied interviews, surveys, brainstorming and prototyping. The focus of the study was the public transportation context and the intrinsic processes in this context.*
***Results**: A software solution specification was developed based on the problems identified in field research, as well as an evaluation of the proposed solution conducted with the target users.*
***Conclusion**: Applying Design Thinking can be efficient in software development because it provides a focus on developing solutions and products targeted to meet the needs of its users, in other words, designs the solution based on how the product will be useful for the end user.*

***Keywords**: Urban Mobility. Public Transportation. Design Thinking. Software Project.*

INTRODUÇÃO

Aliado à constante busca das organizações por qualidade está o conceito de inovação, onde se tem o desenvolvimento de novas soluções para os mais diversos tipos de negócios e problemas. Diferentemente de outras metodologias focadas no processo de Engenharia de *Software* (como RUP, XP etc.), a metodologia *Design Thinking* surge com a finalidade de auxiliar este processo de inovação e negócios, sem especificidade de área de aplicação, e que não tem, intencionalmente, direcionamento algum para o desenvolvimento de *software* (Silva, Silva Filho, Adler, Lucena, & Russo, 2012).

Apesar de recentemente existirem vários casos de empresas *startups* desenvolvendo novos produtos de *software* e negócios, os relatos de experiência da aplicação da metodologia *Design Thinking* (DT) neste ramo ainda é incipiente. Como característica do *Design Thinking* tem-se a multidisciplinaridade, com técnicas e práticas que podem, com suas adequações, serem aplicadas a quase todo tipo de projeto, bem como é focada em atender as expectativas dos usuários do produto/serviço desenvolvido com base em sua estrutura (Silva, et al., 2012).

De acordo com Desconsi, 2012, o *Design Thinking* tem o poder de estimular, promover a inovação e transformar organizações e até mesmo sociedades através de seus métodos. Para isso, é necessário entender o papel do design e seu efeito através do pensamento multidisciplinar por meio da revisão da literatura, e identificar problemas com o conceito desenhado nas teorias de prática em sociologia, ciência e estudos de tecnologia e estudos de organização. Tais salvaguardas auxiliam no delineamento do campo do design e suas relações com os negócios, a gestão, a inovação e com isso tudo a cultura material do qual se inclui. O design parece ter

deixado de ser uma competência de profissões enraizadas em economias industrializadas, para se tornar algo que todos podem praticar.

Para Brown (2010), a missão do Design Thinking é traduzir observações em *insights*, e estes em produtos e serviços para melhorar a vida das pessoas. Com isso, dado que esta metodologia atenta para criação de soluções que têm a preocupação de atender as necessidades dos usuários – além de suas fases de aplicação assemelharem-se a algumas etapas pertinentes ao ciclo de vida de um *software* (engenharia de requisitos, por exemplo) – torna-se relevante a análise desta metodologia no universo de desenvolvimento de *software*. Ressalta-se que o Design Thinking não tem especificidade de áreas de aplicação, ou seja, pode ser adaptado e aplicado a diversos tipos de projetos.

Pesquisadores destacam o envolvimento do usuário final em engenharia de *software* como um conceito importante para o desenvolvimento de sistemas úteis e utilizáveis. No entanto, o envolvimento do usuário final ainda é uma questão delicada. Novos paradigmas, como a computação ubíqua e orientada a serviços fortalecem a necessidade de envolvimento do usuário final mais ativo, a fim de fornecer sistemas personalizados que são adaptados às necessidades de usuários finais individuais (Seyff, Ollmann, & Bortenschlager, 2011).

Neste contexto, o *Design Thinking* desponta como uma metodologia capaz de prover auxílio a essa necessidade exigida, dado que seu foco está no ser humano (usuário), além de se caracterizar como uma abordagem que vê na multidisciplinaridade, colaboração e percepção de pensamentos e processos de forma concreta, considerados caminhos que levam a soluções inovadoras para negócios (Silva, et al., 2012).

Neste trabalho foi escolhido o contexto de mobilidade urbana da cidade de Castanhal (PA), mais especificamente o âmbito dos transportes públicos. A motivação para esta escolha dá-se por consistir em um cenário real, bem como o serviço público de transporte é um dos principais utilizados pela sociedade. No estudo de campo realizado, através de abordagens aos funcionários da SEMUTRAN (Agentes e Coordenador) e aos usuários do serviço de transporte público do município, observou-se nos relatos que a população comumente evidencia situações de infrações e ilegalidades vivenciadas durante sua mobilidade ao utilizar os transportes públicos e, desta forma, sente-se no direito de impugnar tais ocorrências de maneira a denunciá-las. Visto isto, para mitigar tal problema, idealizou-se uma solução baseada em um aplicativo que viabilize a comunicação entre a população e a SEMUTRAN, de forma que os usuários de transporte público possam realizar denúncias e divulgar suas opiniões de insatisfação, no que se refere aos serviços prestados (ocorrência de infrações).

Para isto, pensou-se em aplicar DT como metodologia neste trabalho ao considerar que a mesma – com suas características referentes às técnicas intrínsecas às suas fases – possibilita que o objeto de estudo, no caso o contexto de mobilidade urbana da cidade, com foco nos transportes públicos, fosse analisado de maneira mais realista, possibilitando a imersão direta no cenário. Considera-se, ainda, a possibilidade, em médio prazo, de analisar os processos intrínsecos a esse tema, junto à instituição responsável pela gestão dos serviços públicos de transporte dispostos no município (SEMUTRAN), bem como com a população usuária e, através de observações de comportamentos dos *stakeholders* e do próprio processo motor do serviço público de transporte, identificar problemas reais e pertinentes. Além disto, dado que se trata de um trabalho que pretende atender diretamente uma significativa parcela da população, estima-se que DT possa prover o entendimento do que os envolvidos no universo do estudo "pensam" sobre o problema e "esperam" como solução. Isto consiste no pressuposto principal levantado neste estudo: a DT, aplicada como técnica auxiliadora ao processo de elicitação de requisitos, implica na criação de uma solução de *software* com maior proximidade ao que requisita as necessidades do usuário final.

O presente trabalho, portanto, apresenta o relato da experiência da aplicação de *Design Thinking* no projeto de uma solução de *software* voltada para a realização de denúncias de infrações evidenciadas durante a mobilidade de usuários de transportes públicos. Assim, apresentam-se na seção 2 os conceitos que acercam a metodologia *Design Thinking*; na seção 3 o relato do estudo de campo realizado; na seção 4 o protótipo da solução proposta ao problema identificado e sua validação e na seção 5 as considerações finais acerca do trabalho.

DESIGN THINKING

O DT pode ser definido como uma nova aplicação do conceito de Design idealizada pela empresa norte-americana de consultoria IDEO[1], que apresenta uma metodologia visando proporcionar um caminho mais fácil, rápido e assertivo para a inovação de negócios. Conforme Lookwood (2009, p. 11), *Design Thinking* é:

> Essencialmente um processo de inovação centrado no ser humano que enfatiza observação, colaboração, rápido aprendizado, visualização de ideias, construção rápida de protótipos de conceitos e análise de negócios dos concorrentes, para influenciar a inovação e a estratégia de negócio.

As fases do Design Thinking, como mostra a Figura 1, apesar de se apresentarem de forma linear, são aplicadas em ciclos de iteração não lineares e versáteis (Silva, et al., 2012). Isso porque suas fases podem, durante o processo de execução, serem ajustadas as necessidades do projeto, bem como ao contexto do problema. Tais realizações podem ocorrer de maneira independente, não sendo necessário esperar o término de uma fase para se iniciar outra. A seguir são apresentadas de maneira sucinta as principais fases que compreendem o ciclo de aplicação de *Design Thinking*:

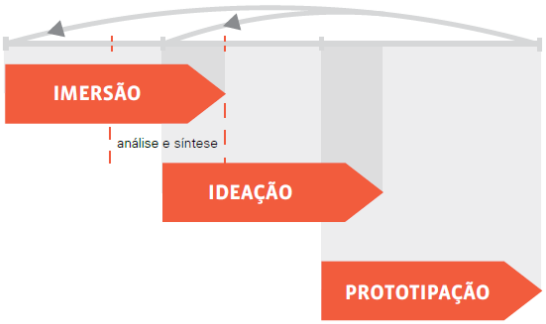


Figura 1. Esquema representativo das etapas do processo de *Design Thinking*.
Fonte: Vianna, 2012. p 18.

Fase de imersão

A fase de Imersão é dividida em preliminar e em profundidade. Na imersão preliminar a equipe busca entender o universo do problema e o tema a ser estudado a partir de diversas perspectivas, obtendo diferentes pontos de vistas acerca deste. São realizadas pesquisas exploratórias em campo e pesquisas *desk* onde se procura por referências e fontes de informações em livros, na Internet, dentre outras. Bem como são identificados, nesta fase, os principais envolvidos na esfera do projeto, além de serem definidos o escopo, limites e restrições para o planejamento e execução do projeto.

A Imersão em profundidade inicia com uma pesquisa baseada em entrevistas, estruturadas e não estruturadas, realizadas com os principais interessados (*stakeholders*) a fim de explorar as visões acerca do contexto do problema. Objetiva-se, com isto, identificar o que implica direta e/ou indiretamente na vida das pessoas, leva-se em consideração aspectos positivos e negativos. Desta forma, reflexões são geradas e registradas e, a partir destas, são extraídos *insights* e conclusões preliminares sobre o tema trabalhado.

Fase de Análise e Síntese

Na fase de Análise e Síntese, como o próprio nome sugere, são realizadas a análise e a síntese das informações coletadas na fase de Imersão. O objetivo é identificar os principais problemas e pessoas (denominadas de "personas") inerentes ao universo estudado. Além disto, determinam-se as primeiras ideias de solução a serem prototipadas. Em suma, se busca a compreensão, parcial ou total, do que é abordado no projeto.

Fase de Ideação

Na fase de Ideação acontece o *brainstorming* de ideias para gerar soluções inovadoras para o contexto estudado. Seções de *brainstorming*, por exemplo, são utilizadas nesta fase, dentre outras técnicas como *workshops* de co-criação e matrizes de posicionamento (Silva, et al., 2012).

[1] Disponível em: <http://www.ideo.com/>. Acesso em: 6 out. 2013.

Fase de Prototipação

Nesta fase é quando de fato materializa-se a abstração feita na fase de Ideação, sobre as hipóteses de soluções apontadas como satisfatórias para sanar ou mitigar os problemas dos usuários. Assim, devem ser construídos protótipos de artefatos que representem os produtos e serviços criados para atender os problemas identificados.

ESTUDO DE CAMPO

O estudo realizado neste projeto objetivou identificar os fatores que implicam, direta e/ou indiretamente, na mobilidade urbana das pessoas que necessitam utilizar os transportes públicos para sua locomoção diária. Para isto, utilizou-se o contexto dos usuários de transportes públicos circulantes no município de Castanhal e localidades adjacentes (chamadas de agrovilas), que são esses: ônibus, táxi, mototáxis, vans, ônibus escolares e os ônibus que transportam os fornecedores de insumos agrícolas das agrovilas vizinhas ao município.

O estudo iniciou com uma busca abrangente de informações sobre mobilidade urbana no município como forma de permitir uma visão holística do contexto analisado. Tais problemas, relatados na seção 3.2, foram identificados através de questionamentos feitos com base nas informações coletadas através da observação e exploração dos fenômenos ocorridos no ambiente do problema estudado (relações, comportamentos, expectativas dos indivíduos, dentre outros).

Desta forma, pôde-se compreender o universo do problema, analisar o impacto deste no ser humano e, com isto, propor uma solução que se adéqua às necessidades dos sujeitos e se encaixe ao problema: buscou-se a percepção e compreensão através do pensamento abdutivo, o qual permite que '*design thinkers*' explorem possibilidades olhando para o futuro, enquanto ainda analisam oportunidades olhando para o passado (Boer & Bonini, 2010).

A seguir são apresentados os principais resultados do estudo de caso de acordo com as fases seguidas da metodologia *Design Thinking* até a fase de ideação. Os resultados da fase de prototipação são apresentados na seção 4, com o detalhamento do protótipo desenvolvido.

Imersão

Para obtenção de conhecimento prévio do contexto da pesquisa e delimitação do escopo do projeto realizou-se, *a priori*, a imersão preliminar e em profundidade onde foram utilizadas as seguintes técnicas:

Pesquisa Exploratória

Nesta etapa realizou a observação do ambiente onde o problema está inserido, ou seja, saiu-se às ruas para observar os transportes públicos, sua circulação, a interação da população com o serviço público de transporte e a identificação das pessoas envolvidas neste contexto.

A partir desta fase pôde-se pensar nos temas centrais a serem abordados, e com isso, elaborar um roteiro para a pesquisa *desk*: buscar temas, problemas e soluções relacionadas ao caso estudado (transportes públicos e mobilidade urbana).

Entrevistas

Entrevistaram-se dez pessoas da população, usuários de transportes coletivos, e também cinco agentes de transporte e o próprio coordenador da secretaria de trânsito e transporte. O coordenador da secretaria de transportes pontuou as principais atividades realizadas pela secretaria, no que diz respeito ao planejamento, gestão controle e fiscalização dos transportes públicos da cidade, que são: planejamento dos itinerários dos ônibus; fiscalização dos transportes públicos (coletivos, táxi, mototáxis, vans, ônibus escolares e ônibus de transporte dos produtores agrícolas); e monitoramento dos itinerários dos coletivos. Quanto à população, foram levantados questionamentos sobre os problemas enfrentados durante sua mobilidade e os principais tipos de infrações que são evidenciadas no serviço público de transporte.

Pesquisa Desk

A partir das informações coletadas nas primeiras entrevistas e nas observações da pesquisa exploratória, realizaram-se buscas na internet utilizando-se os seguintes termos: "Planejamento dos transportes públicos nas cidades", "Mobilidade Urbana", "Fiscalização de transportes públicos", "Leis, normas e regulamentos sobre Mobilidade Urbana". Dentre as normas identificadas no contexto de mobilidade urbana, destaca-se aqui a Lei nº 12.587 (de 3 de janeiro de 2012), intrínseca a nova Política Nacional de Mobilidade Urbana (PNMU).

Procurou-se também identificar sistemas web ou aplicativos que auxiliem a população em sua mobilidade urbana, a fim de encontrar características as quais se pudesse agregar a proposta deste trabalho. Como resultado encontrou-se as seguintes aplicações, por exemplo: Rota Urbana[2], SPTrans[3] e o aplicativo Moovit[4].

Análise e síntese

Analisando as informações coletadas nas entrevistas extraíram-se as principais atividades onde podem ocorrer problemas neste contexto, tais como:

a) Planejamento das rotas dos ônibus: Dada a necessidade de criação de uma nova rota (linha), os agentes de trânsito vão a campo e neste processo, ao percorrer as ruas, eles utilizam uma prancheta e caneta para rabiscar o desenho da rota viável durante percurso;
b) Fiscalização dos transportes: Neste processo, por exemplo, são observadas em campo, pelos agentes, as situações de ocorrência de infrações dos veículos no trânsito;
c) Monitoramento dos itinerários do ônibus: Os agentes encarregados das atividades de fiscalização e monitoramento ficam alocados em pontos estratégicos no centro da cidade e, através do documento de itinerários das linhas de ônibus, é feito o monitoramento onde é observado se os motoristas dos coletivos não estão descumprindo sua rota.

Assim, os principais perfis de usuários identificados ao contexto do problema, conhecidas como *Personas*, foram:

a) Gestor: Representante do poder público. Identificou-se esse perfil como um ponto chave na comunicação entre a população e a secretaria;
b) Agente: Representante dos fiscais de transportes públicos. Este perfil representa o canal direto da comunicação entre a população e a secretaria de transportes;
c) Usuário Estudante: Usuário do transporte público – escolar;
d) Usuário Agrícola: Usuário do transporte público – ônibus agrícola;
e) Usuário Intermediário: Usuário dos demais transportes públicos – vans, táxis, mototáxis e coletivos.

Ideação

Após definidas as personas e identificados os problemas mais pertinentes realizou-se então o processo de ideação, por meio de uma sessão de *brainstorming*, com a presença de um pesquisador experiente e um pesquisador auxiliar envolvidos diretamente na elaboração do projeto. Esta sessão foi baseada nas informações obtidas nas entrevistas feitas durante a fase de Imersão. Neste processo geraram-se as ideias para implementação do protótipo, onde o foco foi analisar as personas identificadas e direcionar as ideias para soluções que atendessem às suas necessidades.

[2] Ferramenta que auxilia o cidadão a se locomover pela cidade de forma a fornecer as rotas das linhas de ônibus da região. Baseado no conceito de Crowdsourcing, os usuários podem contribuir com o registro de rotas (Rota Urbana, 20130. Sistema disponível em: <www.rotaurbana.net.br>.
[3] Sistema Web da Prefeitura de São Paulo que auxilia a população com informações sobre os itinerários e rotas das linhas de ônibus circulantes na cidade. Sistema disponível em: <www.sptrans.com.br>.
[4] Aplicativo gratuito que auxilia os usuários de transporte público a traçarem rotas de circulação durante sua mobilidade por determinado percurso. Os usuários deste app, durante suas viagens, recebem e compartilham informações sobre todos os meios de transporte público em tempo real. *Webpage* do aplicativo disponível em: <www.moovitapp.com>.

Durante aproximadamente duas horas foram expostas, em um quadro, as principais ideias organizadas dentro da regra: persona + problema enfrentado + solução. E, com base nisto, as primeiras ideias de funcionalidades para o aplicativo surgiram e foram definidas, conforme resumido no Quadro 1.

Funcionalidade	Descrição
1. Deve prover Feed de registros para os gestores de transporte baseado em filtro por tipo de denúncia e região.	Os gestores de transporte podem obter informações sobre as ocorrências de infrações cometidas pelos provedores de serviços de transporte público em suas cidades ou região.
2. Deve permitir o georeferenciamento do local onde ocorreu a infração ao se realizar uma denúncia.	Utilização de mapa ou GPS para registro da localização do ponto onde ocorreu a infração.
3. Deve gerar relatório e estatísticas sobre as denúncias registradas pela população	A população pode visualizar dados estatísticos sobre a porcentagem de infrações cometidas e denúncias realizadas.
4. Deve permitir o anexo de evidências quando o reclamante realizar uma denúncia.	Ao realizar uma denúncia, o reclamante deve anexar evidências (fotos, vídeos, imagens) que respaldem suas reclamações.
5. Deve possibilitar ao usuário (reclamante) o acompanhamento de sua denúncia.	Ao fazer uma reclamação, esta fica pendente para validação. *Status*: confirmada, pendente, falsa
6. Deve ter integração com a rede social do usuário (caso o mesmo possua).	A pessoa realiza seu *login* na aplicação através do seu perfil em rede social para compartilhar uma denúncia realizada.
7. Deve permitir a população que realize uma denúncia em tempo real ao fiscal de transporte.	População pode avisar o fiscal através da realização de uma denúncia sobre uma infração cometida.
8. Deve permitir que o gestor retorne uma resposta acerca da reclamação efetuada através de um comentário na denúncia realizada ou email.	Provê *feedback* ao reclamante sobre as medidas tomadas pelos responsáveis por fazer a fiscalização. Pode ser: • O gestor informa a população através de uma resposta à sua denúncia. • Reclamante responde ao gestor sobre o parecer tido do mesmo a cerca de uma denúncia efetuada.

Quadro 1. Funcionalidades do mínimo produto viável para o Aplicativo Monitore.
Fonte: os pesquisadores com base na sessão de *brainstorm*, 2014..

Solução proposta

A partir das ideias geradas para a aplicação na fase de ideação, assim como os requisitos a serem atendidos descritos no Quadro1, foram pensadas duas hipóteses para o protótipo. Uma estava direcionada a projetar um sistema *web*. No entanto, após analisar o perfil das personas e o que era exigido pelas mesmas verificou-se que a proposta poderia ter mais aceitação se estivesse voltada para o conceito de computação móvel. Essa decisão se justifica pelas situações em que as pessoas, que evidenciam ocorrências de infrações, possivelmente usariam algum dispositivo móvel para realizar sua denúncia, além dos representantes da secretaria de transportes, receptores das reclamações, também estarem em constante locomoção nas ruas da cidade.

Desta forma, aderiu-se à hipótese de se projetar uma solução *mobile*. Com isso, começou-se a desenhar as primeiras ideias de telas para o App (aplicativo) com papel e caneta. O aplicativo projetado como solução para o problema abordado neste trabalho foi denominado "Monitore". Após isto, a ferramenta Fluidui[5] foi utilizada para fazer o protótipo de telas, apresentado na próxima seção.

Protótipo de telas

Com base nas ideias geradas durante o brainstorming e na definição das funcionalidades que, possivelmente, assistiriam os usuários, projetaram-se então as telas do aplicativo, denominado "Monitore", conforme apresenta a Figura 2.

O Aplicativo Monitore foi projetado para atender todas as personas identificadas e tem seu foco na realização de denúncias sobre infrações cometidas nos transportes públicos da cidade, onde a população pode diretamente comunicar e evidenciar estas situações aos responsáveis por fiscalizar o serviço público de transporte, agentes e coordenador de transporte. Em sua tela principal, como mostra a Figura 2, o aplicativo disponibiliza quatro áreas principais para interação: "Registrar Denúncia", "Quero Saber", "Sou Gestor" e "Aqui, Agente!" descritas a seguir.

[5] Ferramenta Web utilizada para criação de protótipos de telas de projetos de software. Disponível em: <https://www.fluidui.com/>.

Figura 2. Tela Inicial do Aplicativo Monitore.
Fonte: autoria própria.

Área "Registrar Denúncia"

Esta seção foi direcionada para as personas representantes da população: o estudante, o agrícola e o intermediário. No entanto, as informações obtidas nesta seção favorecem também diretamente a persona gestor, representante do órgão gestor de transportes, pois nesta área pode-se efetivar uma denúncia. Para isso, é necessário que o usuário primeiramente realize o seu registro (Figura 3 – Tela 1) pois, ao se tratar deste tipo de circunstância – onde se expõe um infrator – é preferível inibir o anonimato do reclamante para se ter conhecimento de onde partiram as reclamações. Desta forma, os representantes da Secretaria podem saber a quem se dirigir, em caso de reclamações mais delicadas, como por exemplo: casos de infrações de desrespeito a deficientes físico, ou que portem qualquer outra deficiência.

Após registrar-se, como mostra a Figura 3, é preciso indicar o local onde ocorreu a infração (Tela 2), indicar o tipo de transporte envolvido no ocorrido (Tela 3), indicar o tipo de infração relacionada (Tela 4) e, por fim, descrever dados sobre o veículo (placa, por exemplo) e sobre a infração, além de anexar as evidências do fato, vídeos ou imagens (Tela 5).

Figura 3. Passos para realizar uma denúncia.
Fonte: autoria própria.

Área "Quero Saber"

Como se pode ver no esquema da Figura 4, nesta seção há três opções: Estatísticas, Feedback e Denúncias. Estas foram pensadas para que os usuários possam saber sobre estatísticas de infrações, buscando por tipo de veículo (Tela 1), ou por tipo de infração (Tela 2). Assim como aos que realizaram denúncias, são disponibilizadas duas áreas onde os mesmos podem verificar o *status* de confirmação de suas denúncias efetivadas, que podem ser consideradas como "confirmada", "falsa" ou "pendente", que não foi avaliada ainda (Tela 3), de igual forma podem falar diretamente com o representante da secretaria sobre a resposta obtida (Tela 4).

Figura 4. Área de *feedback* ao usuário.
Fonte: autoria própria.

Área "Sou Gestor"

O módulo "Sou Gestor", representado na Figura 5, foi pensado estritamente para atender a persona gestor de transportes. Para a pessoa representada por este arquétipo é necessário prover informações e relatórios sobre os registros de denúncias efetivadas na sua cidade e/ou região. Para isso o gestor deve registrar-se (Tela 1), assim o mesmo pode obter informações sobre a porcentagem de registros de reclamações dos transportes públicos, buscando por tipo de denúncia ou região (Telas 2 e 3, respectivamente) e, com isso, saber quais tipos de veículos recebem mais reclamações, bem como as infrações, além de poder solicitar recebimento de relatórios (Tela 4).

Figura 5. ÁÁrea de *feedback* ao Gestor.
Fonte: autoria própria.

O Gestor pode também avaliar as denúncias reportadas, referentes à sua cidade, conforme Figura 6. Para isto, o mesmo busca denúncias efetivadas filtrando por tipo de transporte (Tela 1), ao selecionar uma denúncia é possível verificar a descrição da mesma (Tela 2), e ao selecionar a opção "evidências" o usuário poderá ver todas as evidências anexadas (Tela 3). Após avaliar a denúncia, o gestor atribui um *Status* para a mesma e provê uma resposta ao reclamante (Tela 4).

Área "Aqui, Agente!"

Nesta seção o usuário pode, em tempo real, reportar uma reclamação sobre determinada infração evidenciada. A Figura 7 mostra as telas que o usuário visualiza quando seleciona esta área. São duas telas básicas, uma de registro (Tela 1) e outra com o chat, onde pode-se reportar uma reclamação diretamente ao Agente (Tela 2). Na mesma figura (Tela 3) é apresentada a notificação que aparece ao Agente, quando tem-se uma nova reclamação, a qual obterá resposta.

Figura 6. Análise de uma denúncia efetivada.
Fonte: autoria própria.

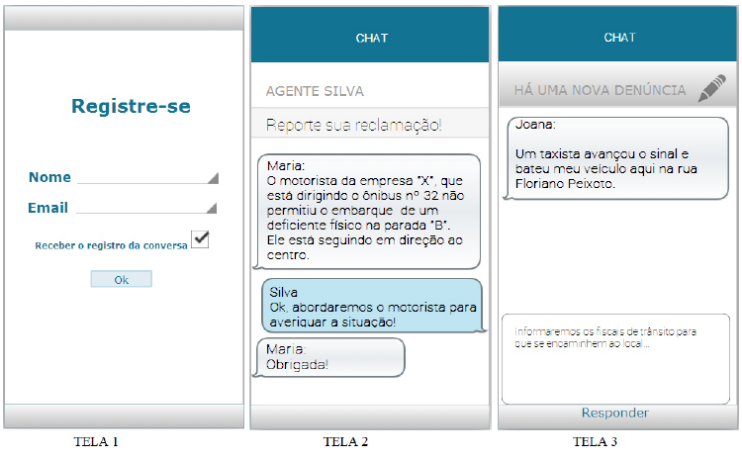


Figura 7. *Chat* com o agente de transporte.
Fonte: autoria própria.

Avaliação

Após a prototipação e para avaliar e validar a hipótese de solução proposta realizou-se um estudo campo onde se aplicaram questionários com a população para saber questões a respeito de sua mobilidade, bem como de sua avaliação sobre o protótipo, o seu esquema funcional foi apresentado na tela do celular.

Para tal, trinta pessoas – em uma faixa etária de 17 a 58 anos – foram escolhidas e abordadas aleatoriamente e solicitadas a responderem algumas questões, tais como: 1) Frequência em que utiliza os transportes públicos (coletivo, escolar, moto táxi, ônibus agrícola, táxi e van), se "nunca", "poucas vezes", "frequentemente" ou "sempre"; 2) Se já realizou algum tipo de denúncia; 3) Se utilizaria o aplicativo Monitore; 4) Se prefere realizar uma denúncia no momento em que a mesma acontece ou prefere deixar para depois (ao chegar em casa, por exemplo); 5) Indicar o grau de utilidade do aplicativo Monitore, no que atende às reclamações de problemas em transporte urbano – de 0 a 10; 6) Críticas, dicas, melhoria ou novas funcionalidade para o aplicativo.

Como resultado, observou-se que 100% dos entrevistados utiliza os transportes públicos. Destes, 90% são usuários frequentes de coletivos e10% de escolares e, 65% relatou utilizar poucas vezes os seguintes meios de transporte: moto táxi, táxi e van (dessa margem a maioria utiliza mais coletivos) e 6% utiliza sempre o ônibus de transporte agrícola. Verificou-se também que a maioria (97%) nunca realizou uma reclamação sobre infrações evidenciadas durante sua mobilidade ao utilizar o serviço público de transporte, ou "por medo de repressão", ou "por não encontrar no momento do fato um agente de transporte no local". Dos respondentes, 83% utilizariam o aplicativo para realizar uma denúncia e, destes, 3% que não o utilizariam corresponde aos usuários na faixa etária mais alta; possivelmente não familiarizados com o uso desse tipo de tecnologia. Todos os respondentes relataram que gostariam de realizar a denúncia no momento em que a mesma ocorreu, e a maioria concordou que "é melhor para não esquecer detalhes da infração". Quanto ao grau de utilidade do

app Monitore, 60% atribuíram grau "8", 23% acharam a proposta muito boa e atribuíram "9" e 16% atribuíram grau "10".

CONSIDERAÇÕES FINAIS

Neste trabalho foi apresentado o relato da experiência sobre a aplicação de uma abordagem multidisciplinar dirigida para o desenvolvimento de projetos inovadores – *Design Thinking* – na criação de um projeto de *software* relacionado ao contexto de mobilidade urbana, conforme estudo de caso abordado.

Com o estudo levantado notou-se que a aplicação de *Design Thinking* pode ser eficiente no desenvolvimento de *software*, pois se tem nessa metodologia um direcionamento na elaboração de projetos centrada na construção de produtos voltados para atender as necessidades de seus usuários, ou seja, projeta-se a solução com base no quão usual o produto será para o usuário final. Assim como suas fases e técnicas não se distanciam dos processos comumente adotados no ciclo de desenvolvimento de *software*, uma vez que suas fases podem ser facilmente mescladas, como por exemplo: A imersão, ideação e prototipação auxiliam na elicitação de requisitos e, até mesmo, no projeto do produto final.

Pensa-se, também, que a utilização das técnicas de *Design Thinking* no processo de elicitação de requisitos pode agregar maior grau de proximidade à realidade das necessidades dos usuários, isso aponta uma forma de evitar problemas na má compreensão e interpretação, quanto às requisições do cliente. Segundo Kujala (2003), o envolvimento do usuário geralmente tem efeitos positivos no sucesso do sistema e satisfação do usuário, e existe evidência de que opiniões dadas pelos usuários, como uma busca por primeiras informações, são algo efetivo da elicitação de requisitos.

Para Vertterli et al. (2013), o *Design Thinking* é consistente com as práticas iniciais de elicitação, inerentes a engenharia de requisitos, prototipagem rápida, relacionamento com o cliente e apresenta-se com um método ágil. Esta metodologia auxilia, no que tange a organização de um projeto de *software*, tanto a documentação dos requisitos, quanto a gestão de equipe e, como supracitado, seu foco é direcionado para o desenvolvimento ágil. No entanto, esta metodologia tem suas limitações, no que se refere a este contexto de aplicação, visto que no desenvolvimento de *software* sabe-se que existem documentações técnicas específicas que auxiliam a condução do projeto, desde sua fase inicial até a construção do produto final. Visto isso, existe a necessidade de agregar artefatos e documentos de Engenharia de *Software* como forma de complemento às fases de condução do *Design Thinking*.

Como forma de tornar mais eficiente algumas fases de desenvolvimento de *software*, nota-se que a mescla de Design Thinking com a metodologia Lean Startup (Ries, 2012) pode agregar efeitos positivos, em especial ao que se refere à engenharia de requisitos e validação de produtos de *software*, uma vez que esta última contribui para avaliar a aceitação do produto final com base na análise de seus usuários finais e, com isso, verificar se as hipóteses de solução são aceitáveis e se realmente satisfazem às necessidades dos usuários.

REFERÊNCIAS

Boer, G., & Bonini, L. (2010). *Design thinking: Uma nova abordagem para inovação*. [Biblioteca TerraForum]. Retirado de http://biblioteca.terraforum.com.br/BibliotecaArtigo/artigo-designthinking.pdf

Brown, T. (2010). *Design thinking: Uma metodologia poderosa para decretar o fim das velhas ideias*. Rio de Janeiro: Elsevier.

Desconsi, J. (2012). *Design thinking como um conjunto de procedimentos para a geração da inovação: Um estudo de caso do projeta do G3*. (Dissertação de Mestrado). Centro Universitário Ritter dos Reis. Porto Alegre.

Kujala, S. (2003). User involvement: A review of the benefits and challenges. *Behaviour & Information Technology*, 22(1), 1-16. Retirado de http://mcom.cit.ie/staff/Computing/prothwell/hci/papers/UserInvolvement.pdf

Loockwood, T. (2009). *Design thinking: Integrating innovation, customer experience, and brand value*. Nova York: Allworth.

Ries, E. (2012). *A startup enxuta*. São Paulo: Lua de Papel.

Rota Urbana. (2013). *Rota Urbana*. Retirado de http://www.rotaurbana.net.br/

Seyff, N., Ollmann G., & Bortenschlager, M. (2011). iRequire: gathering end-user requirements for new apps. *19th IEEE Internacional Requirements Engineering Conference*, 347-348. Retirado de http://ieeexplore.ieee.org.ez10.periodicos.capes.gov.br/stamp/stamp.jsp?tp=&arnumber=6051669

Silva, M. J. V., Silva Filho, Y. V., Adler, I. K., Lucena, B. F., & Russo, B. (2012). *Design thinking: Inovação em negócios*. Rio de Janeiro: MJV.

Como citar este artigo (ABNT):

LIMA, A. M.; ALVES, A. T.; COSTA, A. J. da.; SALES, E. O. Metodologia design thinking no projeto de software para mobilidade urbana: relato de aplicação. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 128-138, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this article* (APA):**

Lima, A. M., Alves, A. T., Costa, A. J., & Sales, E. (2014). Metodologia design thinking no projeto de software para mobilidade urbana: Relato de aplicação. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 128-138. Retrieved from http://www.atoz.ufpr.br

ISSN: 2237-826X
3(2), jul./dez. 2014

A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras

Data mining and the quality of extracted knowledge from police reports of Brazilian federal highways

Jefferson de Jesus Costa[1], Flávia Cristina Bernardini[1], José Viterbo Filho[1]

[1] *Universidade Federal Fluminense (UFF), Rio de Janeiro, RJ, Brasil*

Autor para correspondência/*Corresponding author*: Jefferson de Jesus Costa [jeffersoncosta@id.uff.br]

Agradecimentos/*Acknowlegments*: Agradecemos a todos os envolvidos no processo de confecção desse trabalho e também aos revisores por suas contribuições, que nos auxiliaram a melhorar o material, além de oferecerem interessantes considerações para trabalhos futuros.

Recebido/*Submitted*: 15 Nov. 2014
Aceito/*Approved*: 21 Dez. 2014

Resumo

Introdução: Apresenta e analisa os resultados encontrados com a aplicação do processo de Mineração de Dados nos boletins de ocorrências de rodovias federais brasileiras gerados pela Polícia Rodoviária Federal (PRF) em 2012. O objetivo desse trabalho é analisar a viabilidade da aplicação do processo de Mineração de Dados sobre os dados fornecidos pela PRF, a fim de identificar associações entre variáveis relacionadas aos acidentes de trânsito em todas as rodovias federais.
Método: Empregaram-se algoritmos de aprendizado supervisionado e simbólico e um algoritmo de regras de associação, ambos implementados na ferramenta Weka. Quanto à base de dados o estudo compreende os registros referentes ao ano de 2012. Sobre essa parcela da base de dados aplicou-se a etapa de pré-processamento dos dados, os quais foram utilizados para extração dos modelos e padrões na ferramenta Weka e, por último, avaliaram-se os modelos e os padrões extraídos.
Resultados: No aprendizado supervisionado, os resultados obtidos com os algoritmos J48 e PART foram considerados promissores, pois para todas as classes de causas de acidente, os valores obtidos de área sob a curva ROC (AUC) estiveram acima de 0,5. Além disso, utilizando-se o algoritmo *Apriori*, foram geradas 38 regras de associação com confiança maior que 0,8.
Conclusão: Conclui-se que é importante uma proposta de modelo para distribuição dos dados dessa base de dados, com o objetivo de utilizá-la para o processo de mineração de dados, bem como para outras tarefas de extração de conhecimento e tomada de decisão. Observa-se, ainda, a necessidade de melhoria da qualidade dos dados a serem disponibilizados desde a fase de coleta, ou seja, nos sistemas para cadastro dos dados.

Palavras-chave: Dados Governamentais Abertos. Mineração de Dados. Regras de Associação. Descoberta de Conhecimento em Bases de Dados.

Abstract

***Introduction**: This paper presents and analyzes the results obtained when applying Data Mining process in the bulletins of occurrences of the Brazilian federal highways generated by the Federal Highway Police (PRF) in 2012. The purpose of this work is to analyze the feasibility of implementing the Data Mining process on data provided by PRF in order to identify associations between variables related to transit accidents in all Brazilian federal highways.*
***Method**: It was used symbolic supervised learning algorithms, as well as an algorithm of generation of association rules, implemented in Weka tool. Regarding the database, it was used the records of 2012. On this portion of the database it was conducted the step of data preprocessing, which were used for extracting models and patterns in the Weka tool and, lastly, evaluated the models and extracted patterns.*
***Results**: In supervised learning, the results obtained with J48 and PART algorithms have been considered promising due to the fact that for all classes of accidents causes, the values of area under the ROC curve (AUC) were above 0.5. Furthermore, using the Apriori algorithm there have been generated 38 association rules with confidence greater than 0.8.*
***Conclusion**: It was concluded that is important to propose a model for data distribution of this database, in order to use it for data mining process, as well as other knowledge extraction tasks and decision making. It was noted still, the need to improve the quality of data to be provided from the initial stage of data gathering, that is, in the very systems used to record the data.*

***Keywords**: Open Government Data. Data Mining. Association Rules. Knowledge Discovery in Databases.*

INTRODUÇÃO

Diversos países têm demonstrado interesse em disponibilizar seus dados governamentais de forma pública, isto é, acessíveis a qualquer cidadão, visando aumentar a transparência nas ações governamentais e a participação popular. Segundo a descrição do Portal Brasileiro de Dados Abertos, esse movimento – denominado Open Data – teve início em 2009, sendo que o Brasil aderiu à iniciativa em 2011 (Brasil, 2014b). Dados deste Portal, de particular interesse para esta investigação, são os registros do Sistema BR-Brasil, desenvolvido pelo Departamento de Polícia Rodoviária Federal (DPRF), e de responsabilidade do Ministério da Justiça (Brasil, 2014a). Segundo o Portal, o Sistema

> [...] visa suprir todas as deficiências operacionais em termos de informatização e controle, substituindo a grande maioria dos serviços burocráticos associados às atividades da Polícia Rodoviária Federal e disponibilizando seus registros on-line em todo o país (Brasil, 2014b).

No Portal Brasileiro de Dados Abertos, e mais especificamente neste Sistema, podem ser encontrados os boletins de ocorrências em rodovias federais do País que aconteceram entre 2007 e 2013 (Brasil, 2014b). Uma tarefa interessante e relevante para a sociedade brasileira seria o de analisar os dados de acidentes rodoviários na tentativa de extrair algum padrão e encontrar os principais fatores que estejam causando esses acidentes. Tal tarefa pode auxiliar o processo de tomada de decisão, assim como futuros planejamentos, para que haja uma redução de acidentes nas rodovias federais brasileiras. Segundo Rezende, Pugliesi, Melanda e Paula (2003) e Witten e Frank (2009), as ferramentas e técnicas do processo de Mineração de Dados (MD) podem ser utilizadas para a descoberta de padrões e, neste particular, Reis (2013) apresenta uma proposta de dissertação que visa aplicar o processo de MD com o objetivo de encontrar padrões nas variáveis envolvidas em acidentes de trânsito na rodovia BR-381, no Estado de Minas Gerais, entre 2008 a 2012. Anteriormente, Balbo (2011) propôs um método de análise multivariada para análise dos acidentes da BR-277. Entretanto, até o momento, se desconhecem trabalhos que explores os dados de todas as rodovias federais brasileiras, bem como que descrevam a aplicação do processo de MD em toda ou em parte dessa base (Sistema BR-Brasil).

Uma das etapas do processo de MD é a de pré-processamento, a qual engloba o tratamento e a preparação dos dados. Para que sejam descobertos padrões de qualidade é importante que essa etapa seja cuidadosamente executada (Rezende et al., 2003; Witten & Frank, 2009). Ainda, segundo Facelli, Lorena, Gama e Carvalho (2011), o desempenho dos algoritmos de aprendizado de máquina geralmente é afetado pelo estado em que os dados se encontram, ou seja, pela qualidade dos dados disponíveis. Podem ser mencionadas algumas das tarefas incluídas nessa fase, a saber: limpeza dos dados, tratamento de ruídos, tratamento de dados faltantes, seleção e construção de atributos, dentre outras. Para este estudo, devido à significativa quantidade de dados disponibilizada no Portal, optou-se pela utilização das ocorrências registradas durante o ano de 2012. Ao estudar a base com maior profundidade, diversos problemas puderam ser observados, resultando em dificuldades no processo de descoberta de novos conhecimentos. Tais problemas são descritos oportunamente neste trabalho.

O objetivo desta investigação é, portanto, apresentar as dificuldades encontradas para aplicar o processo de Mineração de Dados na base de dados da Polícia Rodoviária Federal Brasileira, bem como descrever os resultados obtidos ao aplicar o referido processo. Deve ser observado que foram utilizados os dados de todas as rodovias federais do País, sendo ainda realizada uma discussão sobre alguns tratamentos de dados que foram necessários na base de dados. Para o desenvolvimento deste trabalho utilizou-se a ferramenta Weka para aplicação do processo de MD (Witten & Frank, 2009).

O artigo está organizado da seguinte maneira: na segunda seção é descrita uma breve fundamentação teórica sobre o processo de MD e de Aprendizado de Máquina, e também sobre os algoritmos utilizados. Na terceira seção apresenta-se o domínio da aplicação, ou seja, a base de dados de Boletins de Ocorrência da Polícia Rodoviária Federal; os problemas encontrados para minerar a base de dados; e as etapas de pré-processamento realizadas que permitiram a aplicação do processo de Mineração de Dados na base de dados. Na quarta seção são apresentados os resultados obtidos com os algoritmos PART, J48 e *Apriori*. Na quinta e última seção relatam-se as conclusões e apontam-se trabalhos futuros.

Mineração de dados e aprendizado de máquina

A Mineração de Dados pode ser definida como a exploração e a análise, através de meios automáticos ou semiautomáticos, de grandes quantidades de dados com o objetivo de descobrir padrões e regras significativas (Berry & Linoff, 1997). De acordo com Rezende et al. (2003) e Witten e Frank (2009), o processo de Mineração de Dados pode ser dividido, basicamente, em três diferentes etapas: (i) pré-processamento dos dados; (ii) extração de modelos e padrões; e (iii) avaliação dos modelos e padrões extraídos. A primeira fase – a de pré-processamento dos dados – envolve tarefas de limpeza dos dados, tais como aplicação de filtros, seleção e construção de atributos, preenchimento de valores faltantes, tratamento de ruídos, entre outras. O objetivo dessa fase é tornar os dados estatisticamente de melhor qualidade para extração de padrões. Na fase de extração de modelos e padrões podem ser utilizados diferentes métodos e técnicas de aprendizado de máquina (Rezende et al., 2003; Witten & Frank, 2009). Para utilizar algoritmos de aprendizado de máquina no processo de Mineração de Dados (MD) podem ser empregadas, basicamente, duas abordagens para descoberta de conhecimento: aprendizado preditivo e aprendizado descritivo (Facelli et al., 2011).

No aprendizado preditivo, o algoritmo de aprendizado é uma função que objetiva construir um estimador dado um conjunto de exemplos rotulados. O rótulo (ou etiqueta) toma valores em um domínio conhecido. Se o domínio dos rótulos, ou seja, o conjunto ao qual os rótulos dos dados pertencem, for um conjunto infinito e ordenado de valores (p. ex., o conjunto dos números reais), o problema é dito de regressão e o estimador é denominado "regressor". Porém, se o domínio dos rótulos é um conjunto finito e não ordenado de valores, o problema é dito de classificação, e o estimador é denominado "classificador". O aprendizado preditivo é também conhecido por aprendizado supervisionado, e é o tipo de tarefa de predição utilizado neste trabalho.

No aprendizado descritivo, as tarefas envolvem a identificação de informações relevantes nos dados sem um elemento externo para guiar o processo de aprendizado. As tarefas descritivas podem ser divididas em: sumarização, cujo objetivo é encontrar uma descrição mais simples e compacta dos dados; associação, cujo objetivo é buscar padrões frequentes de associações entre os atributos de um conjunto de dados; e agrupamento, cujo objetivo é identificar grupos nos dados de acordo com a similaridade entre os objetos. Neste trabalho, foi explorada a tarefa de associação para o conjunto de dados utilizado, para tentar identificar relações entre fatores de acidentes rodoviários.

Aprendizado supervisionado

No problema padrão de aprendizado de máquina supervisionado, a entrada do algoritmo consiste de um conjunto de exemplos S, com N exemplos T_i , $i = 1, \ldots, N$, escolhidos de um domínio X com uma distribuição D fixa, desconhecida e arbitrária, da forma $\{(\mathbf{x}1, y1), \ldots, (\mathbf{x}_N, y_N)\}$ para alguma função desconhecida $y = f(x)$. Os $\mathbf{x}_i$ são tipicamente vetores da forma $(x_{i1}, x_{i2}, \ldots, x_{iM})$ com valores discretos ou numéricos. x_{ij} refere-se ao valor atributo j, denominado X_j, do exemplo T_i. Neste trabalho, também se denominam os atributos X_j como atributos de descrição do domínio. Os valores de y_i referem-se ao valor do atributo Y, frequentemente denominado atributo classe. Os valores de y em problemas de classificação, como é o caso neste trabalho, são tipicamente pertencentes a um conjunto discreto de classes $C = \{C_v\}$, $v = 1, \ldots, N_{Ci}$, ***i.e*** $y \in \{C_1, \ldots, C_{NCl}\}$. O objetivo de um algoritmo de aprendizado supervisionado para problemas de classificação é construir um classificador **h**, que tem como entrada um exemplo **x**, não classificado, ou seja, um vetor de valores de atributos discretos e/ou contínuos, e a sua saída é um valor discreto, ou seja, a classe a ser predita (Domingos, 2012). O classificador **h** é também denominado hipótese da desconhecida e verdadeira função **f**, tal que, para todo $\mathbf{x} \in X$, $\mathbf{f}(\mathbf{x}) = y$ (Mitchell, 1997).

Para avaliar os conhecimentos gerados a partir da base de dados de interesse neste trabalho, utilizando aprendizado supervisionado para problemas de classificação, foram definidos os algoritmos de aprendizado de máquina PART (Witten & Frank, 2009) e J48, que é uma implementação do algoritmo C4.5 (Quinlan, 1993; Witten & Frank, 2009). Ambos os algoritmos oferecem como saída conjuntos de regras facilmente interpretáveis por seres humanos. O objetivo do algoritmo PART é induzir um classificador composto por regras de decisão. Já o J48 tem como finalidade gerar uma árvore de decisão baseada no conjunto de dados de treinamento.

As árvores de decisão possuem um custo computacional baixo, por isso têm sido largamente utilizadas em problemas de classificação. Além disso, são fáceis de entender, fato que aumenta a confiabilidade neste tipo de estrutura. A ideia por trás dos algoritmos de indução de árvore de decisão é decompor a classificação em um conjunto de escolhas sobre cada variável em etapas, iniciando na raiz da árvore e percorrendo as folhas, onde ocorre a classificação. Os diversos algoritmos de árvore de decisão existentes utilizam basicamente o mesmo princípio: a árvore é construída de maneira gulosa, começando pela raiz, escolhendo o atributo com mais informação a cada iteração (Quinlan, 1988). O algoritmo C4.5 (Quinlan, 1993), escrito na linguagem C e usado para gerar árvores de decisão, deu origem ao algoritmo J48, que é uma implementação *open source* em Java do C4.5 para o *software* Weka (Witten & Frank, 2009). O objetivo do J48 é gerar uma árvore de decisão baseada em um conjunto de dados rotulados. O J48 envolve variáveis qualitativas contínuas e discretas presentes na base de dados, por isso a sua expressiva utilização no processo de descoberta de conhecimento e de geração de árvores de decisão. Além disso, é considerado o algoritmo que apresenta o melhor resultado na indução de árvores de decisão, a partir de um conjunto de dados de treinamento. Para induzir uma árvore de decisão, o J48 utiliza a abordagem de dividir-para-conquistar, ou seja, divide um problema complexo em subproblemas mais simples, aplicando recursivamente a mesma estratégia a cada subproblema, dividindo o espaço definido pelos atributos em subespaços, associando-se a eles uma classe (Witten & Frank, 2009). Uma característica

interessante da árvore de decisão está relacionada a cada caminho da árvore gerar uma regra de decisão, e entre as regras não existe intersecção de cobertura de exemplos. Em outras palavras, não existe sobreposição dessas regras no espaço de descrição dos exemplos (Baranauskas & Monard, 2000).

O PART foi desenvolvido tendo como base o algoritmo J48. Ele gera uma lista de decisão e, assim como o J48, também usa a técnica de dividir-para-conquistar. O algoritmo constrói uma árvore de decisão C4.5 parcial a cada iteração e coloca a melhor folha dentro de uma regra (Witten & Frank, 2009). O processo de geração das regras de associação acontece da seguinte maneira: as regras são induzidas a partir de uma árvore e posteriormente são refinadas. Para cada regra criada, é estimada a sua cobertura das instâncias da base de dados. Isso acontece repetidas vezes até que todas as instâncias estejam cobertas. As regras com coberturas mais altas são mantidas e apresentadas para o usuário e as demais são descartadas (Frank & Witten, 1998). A diferença de um algoritmo de indução de regras de decisão em relação a um algoritmo de árvore de decisão reside no fato de que as regras de decisão são induzidas para cobrir um conjunto de exemplos e dessa maneira pode haver sobreposição das regras construídas no espaço de descrição dos exemplos (Baranauskas & Monard, 2000). Dessa maneira, os conceitos aprendidos com esses diferentes algoritmos podem ser bastante distintos, tendo sido utilizados – para fins deste estudo – dois algoritmos de indução de classificadores simbólicos.

Aprendizado de regras de associação

Neste trabalho foi utilizado o algoritmo de construção de regras de associação *Apriori* (Borgelt & Kruse, 2002), cujas regras produzidas associam atributos do domínio de descrição dos exemplos. O algoritmo *Apriori* foi proposto por Agrawal, Imielinski e Swami (1993), e consiste na busca por padrões que indicam o relacionamento entre conjuntos de itens. O Apriori é um dos algoritmos mais utilizados para a descoberta de regras de associação, pois executa diversas leituras na base de dados de transações, sendo capaz de trabalhar com um número grande de atributos. Como resultado, o algoritmo obtém várias alternativas combinatórias entre eles. Ainda assim, devido ao processo de otimização para a geração das regras, o algoritmo consegue ter um bom desempenho em termos de processamento. O *Apriori* também utiliza a técnica de dividir-para-conquistar, com o objetivo de encontrar regras de associação para todas as expressões possíveis.

Seja $A=\{a_1,\ldots,a_q\}$ o universo de q itens. Os itens podem ser produtos, ou valores específicos de atributos, de um conjunto de dados (p. ex., "leite" e "pão" são itens em um domínio de supermercado, ou "idade = jovem", se idade for um dos atributos de descrição do domínio). Um conjunto de itens I é um subconjunto de A, ou seja, $I \subseteq A$. As regras de associação são definidas como: "Se I_i então I_j" ou "I_i => I_j", onde I_i e I_j são conjuntos de itens, $I_i \cap I_j = \emptyset$, I_i é o antecedente da regra, e I_j é o consequente da regra. No *Apriori*, dado um conjunto de transações, ou conjunto de dados, S_{trans}, a busca pelas regras de associação é realizada em duas fases: geração e poda. Na primeira fase, o algoritmo percorre todo o conjunto de dados para gerar todas as combinações de valores possíveis. Em seguida, são mantidas apenas as combinações com uma frequência maior que um valor mínimo pré-determinado, denominado suporte. O suporte de um conjunto de itens I, denominado sup(I), é definido pela Equação 1, onde o símbolo # significa "número de" (Carvalho, Sampaio, & Mongiovi, 1999).

$$\sup(I) = \frac{\#\ transações\ que\ contém\ os\ os\ elementos\ do\ conjunto\ de\ itens\ I}{\#\ total\ de\ transações} \quad (1)$$

Na segunda fase, as regras são construídas a partir dos conjuntos de itens, e é utilizada outra medida para seleção das regras consideradas relevantes – o fator de confiança de uma regra R: $I_i \rightarrow I_j$. A confiança de uma regra *conf*(R) é definida pela Equação 2:

$$\mathrm{conf}(R) = \frac{\#\ transações\ que\ possuem\ I_i\ e\ I_j}{\#\ transações\ que\ possuem\ somente\ I_i} \quad (2)$$

Além de utilizar a confiança como parâmetro de seleção das regras, essa medida também é usada para avaliar a qualidade das regras construídas.

Deve ser observado que, como o *Apriori* espera que cada atributo de descrição do domínio possua itens para serem relacionados, é necessário que o domínio de cada atributo seja discreto, ou seja, possua um número limitado de valores possíveis.

Avaliação dos classificadores

Para avaliar um classificador **h**, inicialmente é necessário coletar informações das decisões tomadas pelo classificador em um conjunto de teste S_{te}, não utilizado na fase de treinamento de **h**. Para isso, é construída uma matriz bidimensional, cujas dimensões são denominadas classe verdadeira e classe predita. A essa matriz dá-se o nome de matriz de confusão, mostrada na Tabela 1. Cada elemento $M(C_i,C_j)$ da matriz, definido pela Equação 3, indica o número de exemplos que pertencem à classe C_i e foram preditos como pertencentes à classe C_j. Nessa equação, $||\mathbf{h}(\mathbf{x}) = C_j||$ é igual a 1 se a igualdade $\mathbf{h}(\mathbf{x}) = C_j$ for verdadeira, ou é igual a 0 se a igualdade for falsa. O número de predições corretas para cada classe são os números apresentados na diagonal principal da matriz de confusão, ou seja, os valores associados a $M(C_i,C_j)$. Todos os outros elementos da matriz $M(C_i,C_j)$, para $i \neq j$, são referentes ao número de erros cometidos em cada classe. Para cada classe Cv, $v = 1$,..., NCl, pode-se calcular as taxas de verdadeiros positivos (*TP*, do inglês *True Positives*), verdadeiros negativos (*TN*, do inglês *True Negatives*), falsos positivos (*FP*, do inglês *False Positives*), e falsos negativos (*FN*, do inglês **False Negatives**). A ideia é considerar cada classe C_v como sendo a classe positiva e todas as outras como compondo a classe negativa em relação à C_v. Assim, $TP_{Cv} = M(C_v,C_v)$; $TN_{Cv} = \sum_{\forall Ci \forall C | i \neq v} M(C_i,C_i)$; $FP_{Cv} = \sum_{\forall Ci \forall C | i \neq v} M(C_v,C_i)$; e $FN_{Cv} = \sum_{\forall Ci \forall C | i \neq v} M(C_i,C_v)$. Neste trabalho, utiliza-se a taxa de erro de **h** – *err*(**h**) –, que é a soma de todos os valores $M(C_i,C_j)$ tais que $i \neq j$, bem como outras medidas para avaliar o comportamento do classificador nas classes. Tais medidas são precisão – *Prec*(**h**), definida pela Equação 6; sensibilidade, ou *recall* – *Rec*(**h**), definida pela Equação 7; F – *F*(**h**), definida pela Equação 8; e área sob a curva ROC – *AUC*(**h**). A curva ROC é um gráfico que trata a relação entre as taxas de TP e FP, pois o ideal é que TP seja 1, e FP seja 0, Assim, quanto mais próximo um classificador possua o par (*TP*, *FP*) ao ponto (1,0), melhor o classificador. Deve ser observado que classificadores cujos pares (*TP*, *FP*) tais que *TP* = *FP* são modelos considerados aleatórios. Daí, a área sob a curva ROC (*AUC*, do inglês *Area Under Curve*) pode ser calculada – quanto mais próximo de 1 o valor da *AUC*, melhor o modelo construído. A média das *AUC*s, calculada sobre a *AUC* para cada classe C_v, pode ser então calculada.

Classe Verdadeira	Predita C_1	Predita C_2	...	Predita C_{NCl}
C_1	$M(C_1, C_1)$	$M(C_1, C_2)$	...	$M(C_1, C_{NCl})$
C_2	$M(C_2, C_1)$	$M(C_2, C_2)$	...	$M(C_2, C_{NCl})$
...	...	...	...	...
C_{NCl}	$M(C_{NCl}, C_1)$	$M(C_{NCl}, C_2)$	...	$M(C_{NCl}, C_{NCl})$

Tabela 1. Matriz de Confusão.
Fonte: autoria própria.

$$M(C_i, C_j) = \sum_{\forall (\mathbf{x},y) \in S_{te} | y = C_i} ||\mathbf{h}(\mathbf{x}) = C_j|| \quad (3)$$

$$Prec(\mathbf{h}) = \frac{\sum_{v=1}^{NCl} TP_{C_v}}{\sum_{v=1}^{NCl} TP_{C_v} + FP_{C_v}} \quad (4)$$

$$Rec(\mathbf{h}) = \frac{\sum_{v=1}^{NCl} TP_{C_v}}{\sum_{v=1}^{NCl} TP_{C_v} + FN_{C_v}} \quad (5)$$

$$F(\mathbf{h}) = \frac{2 \times Prec(\mathbf{h}) \times Rec(\mathbf{h})}{Prec(\mathbf{h}) + Rec(\mathbf{h})} \quad (6)$$

A fim de estimar as medidas previamente descritas, existem diversas técnicas para construir o conjunto de treinamento e teste. É comum utilizar a técnica de validação cruzada para estimar a taxa de erro de um classificador, bem como as outras medidas. Na explicação a seguir, detalha-se a estimativa da medida *err*(**h**); de maneira análoga, podem ser estimadas as outras medidas. Na técnica de validação cruzada com *K* partições, o conjunto de dados *S* é dividido aleatoriamente em *K* partições $S_1,...,S_K$, disjuntas, sendo todas as partições de conjuntos de dados de aproximadamente o mesmo tamanho. Após, são executadas *K* iterações de indução e teste de um classificador. Na primeira iteração, é induzido o classificador $\mathbf{h}_1$ utilizando os conjuntos de dados $S_2,...,S_K$. Daí, $\mathbf{h}_1$ é testado com o conjunto S_1, obtendo assim a taxa de erro $err(\mathbf{h}_1)$. Na segunda iteração, é induzido o classificador $\mathbf{h}_2$ utilizando os conjuntos de dados S_1 e $S_3,...,S_K$. Daí, $\mathbf{h}_2$ é testado com o conjunto S_2, obtendo assim a taxa de erro $err(\mathbf{h}_2)$, e assim sucessivamente. Então, a média $m_{err}(\mathbf{h})$ e o erro padrão $se_{err}(\mathbf{h})$ dessas taxas de erro são calculados, definidos respectivamente pelas Equações 7 e 8, do modelo final. Esse

modelo final é construído utilizando todos os exemplos disponíveis. Estimados a média e o erro padrão da taxa de erro do classificador **h**, pode ser utilizado o teste t de *Student* para comparar o poder de predição dos dois algoritmos de aprendizado de máquina para um mesmo conjunto de dados (Baranauskas & Monard, 2000).

$$m_{Err}(\mathbf{h}) = \frac{1}{K}\sum_{k=1}^{K} err(\mathbf{h}_k) \tag{7}$$

$$se_{Err}(\mathbf{h}) = \sqrt{\frac{1}{K-1} \times \frac{1}{K}\sum_{k=1}^{K} (err(\mathbf{h}_k) - m_{Err}(\mathbf{h}))^2} \tag{8}$$

No Weka se implementa a validação cruzada estratificada com *K* partições. A diferença está na maneira em que é feita a divisão do conjunto de dados original em *K* partições. Neste tipo de técnica com *K* partições, as partições são feitas de modo que seja respeitada a distribuição dos exemplos nas classes. Ou seja, se no conjunto de dados original existem 20% dos exemplos na classe C_1 e 80% dos exemplos na classe C_2, então em cada partição construída com a técnica de validação cruzada estratificada com *K* partições, existem aproximadamente 20% dos exemplos pertencentes à classe C_1 e 80% dos exemplos pertencentes à classe C_2.

Pré-processamento de dados

Um conjunto de dados pode conter diversos tipos de ruídos e/ou imperfeições, como valores incorretos, inconsistentes, duplicados ou ausentes. Frequentemente são utilizadas técnicas de pré-processamento de dados para melhorar a qualidade dos mesmos, essas técnicas podem ser de eliminação ou minimização dos problemas citados. Dados processados – onde estão presentes apenas atributos relevantes para o domínio – levam à indução de conceitos mais precisos e mais enxutos, o que também implica em uma maior facilidade na interpretação dos padrões extraídos. Técnicas de pré-processamento são úteis também para tornar os dados mais adequados para um determinado algoritmo, como por exemplo, a substituição do domínio de um atributo contínuo por um domínio discreto, tarefa necessária quando se utiliza o algoritmo *Apriori* (Facelli et al., 2011).

Nem todos os atributos do conjunto de dados original são necessários para determinada tarefa de aprendizado de máquina como, por exemplo, um atributo que possua o mesmo valor para todas as instâncias. Quando um atributo não contribui para a estimativa do valor do atributo classe, ele é considerado irrelevante. (Facelli et al., 2011)

Na seção a seguir é descrita a base de dados de interesse para o desenvolvimento deste trabalho.

Base de dados: boletins de ocorrências em rodovias federais brasileiras da Polícia Rodoviária Federal

Os boletins de ocorrências em rodovias federais brasileiras, disponíveis no Portal Brasileiro De Dados Abertos (Brasil, 2014b), são caracterizados como Dados Abertos Governamentais (DAG) ou dados públicos, pois são disponibilizados na internet para livre utilização pela sociedade (Agune, Gregorio Filho, & Bolliger, 2010). A comunidade envolvida com os DAG afirma que, para que os dados sejam definidos como tal, eles devem seguir oito princípios listados a seguir (The Annotated 8 principles of Open Government Data, 2014):

a) Completos. Todos os dados públicos são disponibilizados. Dados são informações eletronicamente gravadas, incluindo – mas não se limitando a – documentos, bancos de dados, transcrições e gravações audiovisuais. Dados públicos são dados que não estão sujeitos a limitações válidas de privacidade, segurança ou controle de acesso, regulados por estatutos;
b) Primários. Os dados são publicados na forma coletada na fonte, com a mais fina granularidade possível, e não de forma agregada ou transformada;
c) Atuais. Os dados são disponibilizados o quão rapidamente seja necessário para preservar seu valor;
d) Acessíveis. Os dados são disponibilizados para o público mais amplo possível e para os propósitos mais variados possíveis;
e) Processáveis por máquina. Os dados são razoavelmente estruturados para possibilitar o seu processamento automatizado;
f) Acesso não discriminatório. Os dados estão disponíveis a todos, sem que seja necessária identificação ou registro;

g) Formatos não proprietários. Os dados estão disponíveis em um formato sobre o qual nenhum ente tenha controle exclusivo;
h) Livres de licenças. Os dados não estão sujeitos a regulações de direitos autorais, marcas, patentes ou segredo industrial. Restrições razoáveis de privacidade, segurança e controle de acesso podem ser permitidas na forma regulada por estatutos.

Deve ser observado que o segundo princípio determina que os dados abertos devam ser publicados como coletados na fonte, princípio este que pode ser percebido nos boletins de ocorrências da PRF. Entretanto, observa-se que esse princípio dificultou a etapa de pré-processamento para o processo de Mineração de Dados, especialmente devido: a) ao volume significativo de dados; b) a uma grande quantidade de dados faltantes; e c) aos diversos problemas de dados errôneos encontrados (corrigidos manualmente quando possível). Vários desses problemas são descritos adiante.

Os dados analisados neste trabalho foram obtidos no referido portal de dados abertos e fazem parte da base de dados da Polícia Rodoviária Federal. Primeiramente, foi analisada qual parcela de dados seria considerada útil para analisar as ocorrências que foram registradas durante todo o ano de 2012, pois este era o ano mais recente e com os dados dos dois semestres publicados. Infelizmente, os dados do segundo semestre de 2013 ainda não tinham sido publicados, o que reforçou a opção por se utilizarem os dados de 2012. Pode-se observar que este fato fere o terceiro princípio dos DAG, que diz respeito à atualidade dos dados. Nessa porção da base de dados foi identificado que algumas tabelas e diversos campos possuem informações consideradas irrelevantes e/ou desnecessárias para o processo de mineração de dados, incluindo a aquisição de novos conhecimentos. Um exemplo são os atributos: (i) *ocotipo*, da tabela '*OCORRENCIA*', que registra o tipo da ocorrência. Neste, todos os registros contêm o valor "1" (que significa acidentes rodoviários). Ao se considerar que esse atributo apresenta um dado redundante, procedeu-se a retirada; (ii) *ocostatus*, também presente na tabela de ocorrências, que registra o *status* da ocorrência. Todos os registros contêm o valor "S", que significa que a ocorrência já foi encerrada, e por isso o atributo também foi removido; (iii) o campo *oacdanoterc*, presente na tabela '*OCORRENCIAACIDENTE*', registra se aconteceu dano a terceiro. A maioria dos registros (384211 de 390974 registros totais) contém o valor 'N', e, por esse motivo, esse campo também foi removido; e (iv) na tabela '*PESSOA*', os campos *pesteccodigo*, que identifica o estado civil do envolvido em uma ocorrência, e o campo *pestgicodigo*, que identifica o grau de instrução da pessoa, foram desconsiderados, pois ambos eram campos do tipo chave estrangeira para outra tabela, que não estão disponíveis no portal de dados abertos. Questionou-se, ainda, a utilidade de alguns dados, devido à significativa quantidade de dados faltantes em alguns atributos. O Diagrama de Entidade e Relacionamento (DER), que pode ser visualizado na Figura 1, ilustra as entidades utilizadas neste trabalho para a extração dos dados[1]. Uma descrição detalhada deste DER pode ser encontrada em Anexo.

Dentre os problemas identificados, durante o pré-processamento dos dados, é importante destacar:

a) alguns atributos estão presentes na base de dados, porém nem todos são úteis para a descoberta de conhecimento e para o próprio sistema BR-Brasil, pois foram descontinuados nas alterações de versões do sistema, ou seja, alguns campos utilizados em determinada versão do sistema foram retirados em outras;
b) significativa quantidade de dados faltantes – nos 390.973 registros totais, foram encontradas 181.428 ocorrências de dados faltantes, que tiveram que ser substituídos por '?';
c) cidades e códigos de rodovias federais inexistentes e/ou repetidos;
d) dicionário de dados incompleto e de difícil compreensão, pois alguns atributos não são descritos e, por outro lado, outros, como o atributo *oacgirofundo* (tabela de ocorrências dos acidentes), não possuem uma identificação de sua finalidade;
e) algumas tabelas, como a que identifica o modelo de pista, não estão no conjunto de dados publicados;
f) DER incompleto e desatualizado;
g) alguns campos possuem diferentes opções de escolha, mas em todos os registros somente um valor é escolhido, como acontece no atributo que registra o tipo de envolvido no acidente: dentre 16 opções, apenas duas – passageiro e condutor – são utilizadas;

[1] O DER completo de toda a base da PRF pode ser visualizado em http://migre.me/iehd4.

h) o campo que demonstra o estado físico do acidentado não agrega valor ao processo de descoberta de conhecimento, pois não é preenchido sempre, e quando o é, apresenta erros de digitação; além disto, cada pessoa envolvida em um acidente pode ter um tipo de estado físico diferente;
i) na tabela '*TIPOVEICULO*' encontram-se valores repetidos e/ou similares, como por exemplo, "reboque", "semi-reboque e reboque" ou "semi-reboque";
j) pelos valores observados no atributo município, acredita-se que o preenchimento do nome do município onde ocorreu o acidente (bem como da BR) são campos de texto, ou seja, o usuário pode digitar sem uma padronização (valores predeterminados), o que implica em vários erros de digitação;
k) campos que poderiam agregar valor ao processo de MD, tais como o registro do estado da pista na hora do acidente, ou a ocorrência de danos ao ambiente, não são preenchidos na maioria das vezes;
l) a separação das ocorrências por semestre é feita com base na data da finalização da mesma, o que implica que ocorrências de 2001, por exemplo, podem ser encontradas nos dados de 2012, pois sua finalização ocorre nesse ano; e
m) falta de padronização nas entradas dos dados, tais como valores de textos em campos numéricos e vice-versa, nomes de cidades grafados erroneamente e/ou digitados de maneiras distintas, códigos de rodovias federais inexistentes, dentre outros.

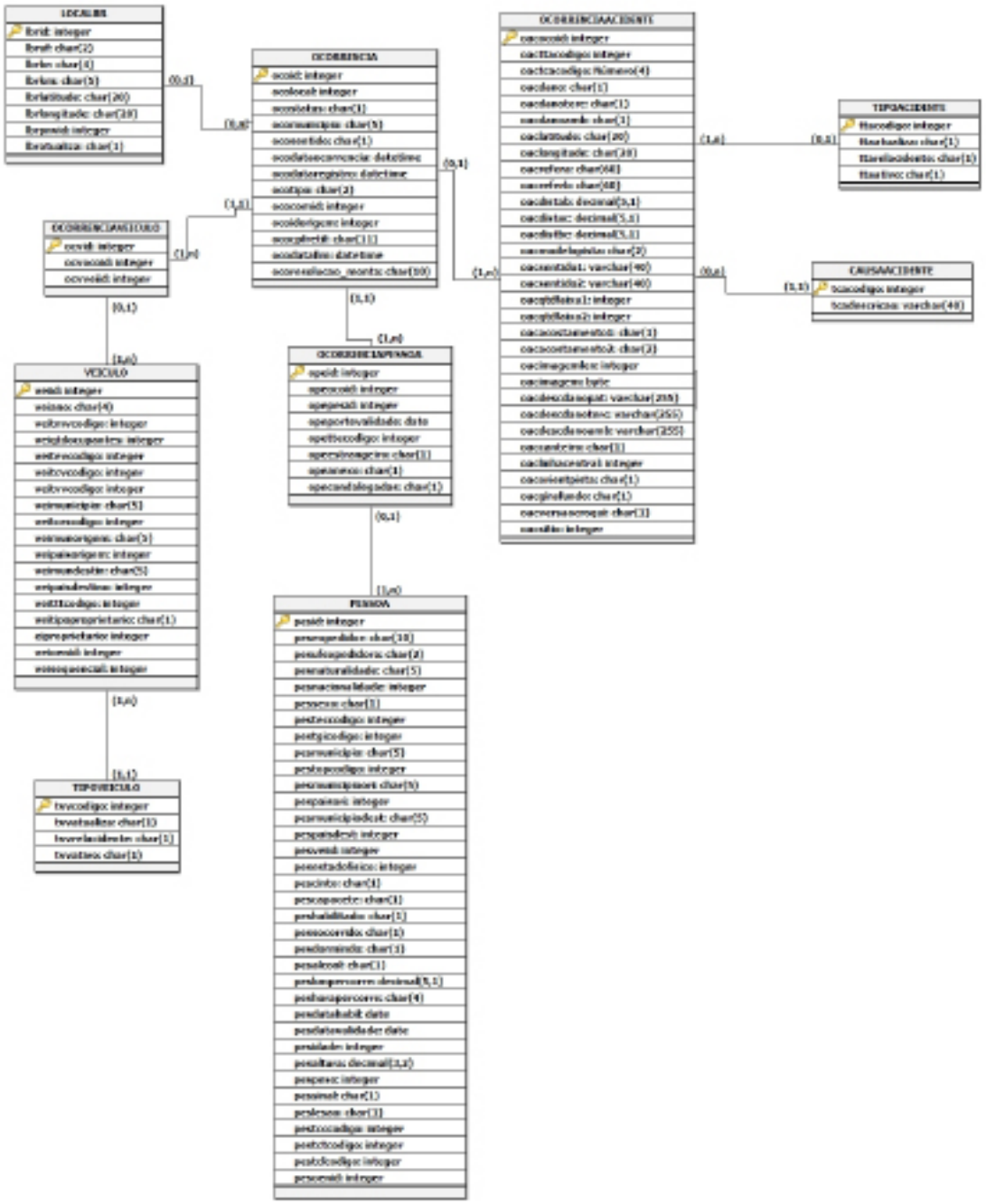


Figura 1. Diagrama de entidade e relacionamento da base de dados utilizada.
Fonte: autoria própria.

Descrição do estudo: aplicação de ferramentas de mineração de dados nos dados de boletins de ocorrências da PRF em 2012

Para esse estudo foram utilizados os dados do ano de 2012, presentes na base de dados de ocorrências em rodovias federais, da PRF. Em seguida, realizou-se a limpeza desses dados, principal tarefa da etapa de pré-processamento. Utilizou-se o *software* Weka (Witten & Frank, 2009) que requer que os dados estejam em formato atributo-valor. Efetivaram-se, ainda, outras pequenas transformações, devido a restrições da ferramenta e do uso do algoritmo *Apriori* (esse, por sua vez, exige que todos os atributos de descrição do domínio possuam um domínio discreto). Realizaram-se as seguintes substituições:

a) os atributos do tipo data foram alterados para dia da semana;
b) o horário do acidente foi transformado para o período do dia (manhã, tarde, noite ou madrugada) referente ao acidente;
c) a data de fabricação do veículo serviu para categorizarmos o veículo em novo, seminovo, usado ou com mais de 10 anos de uso;
d) a data de nascimento foi substituída pela faixa etária da pessoa (criança, adolescente, jovem, adulto ou idoso);
e) a data de vencimento da Carteira Nacional de Habilitação (CNH) foi substituída por um campo que informa se o motorista estava com a habilitação vencida ou não;
f) os dados faltantes foram substituídos por '?'.

Na Tabela 2 são exibidas as características da base de dados processada, na qual # At. Discretos e # At. Contínuos são, respectivamente, o número de atributos de descrição do domínio que são discretos e contínuos; # Exs. é o número total de exemplos presentes na base de dados utilizada; # Classes (*NCl*) é o número de classes presentes na base de dados; e Distribuição dos Exs. nas Classes apresenta as classes presentes na base de dados, o número de exemplos presente em cada classe (# Exs.) e o percentual de exemplos em cada classe (% Exs.). Deve ser observado que na coluna # Exs. existem dois valores. O primeiro valor é referente a todos os exemplos presentes no conjunto de dados. No entanto, dos 390.973 exemplos, somente 294.480 são rotulados com a classe "Causa do Acidente". Portanto, somente esses exemplos foram utilizados nos algoritmos de predição (J48 e PART), e considerados para computar a distribuição dos exemplos nas classes, exibida na respectiva coluna da Quadro 1.

# At. Discretos	# At. Contínuos	# Exs.	# Classes (*NCl*)	Distribuição dos Exs. nas Classes		
				Classe	# Exs.	% Exs.
				Animais na Pista	7211	2,4%
				Defeito Mecânico no Veículo	13088	4,4%
				Defeito na Via	4346	1,5%
				Desobediência à Sinalização	19266	6,5%
8	0	390.973 / 294.480	10	Dormindo	8638	2,9%
				Falta de Atenção	136342	46,3%
				Ingestão de Álcool	16007	5,4%
				Não Guardar Distância de Segurança	47558	16,1%
				Ultrapassagem Indevida	11904	4,0%
				Velocidade Incompatível	30120	10,2%

Quadro 1. Características da base de dados processada – ano de 2012.
Fonte: autoria própria.

No Quadro 2, é apresentada uma descrição do conteúdo de cada atributo e, no Quadro 3, exibidos os atributos e os valores possíveis nos atributos.

Atributo	Conteúdo
Tipo de Veículo	Tipo de veículo envolvido na ocorrência, como por exemplo, automóvel, motocicleta, etc.
Ano do Veículo	Categorização do veículo de acordo com o seu ano de fabricação. Na base original esse valor é numérico e representa o ano de fabricação do veículo, mas para que seja possível extrair conhecimento, esses números foram transformados em variáveis que agregam valor. A categorização foi feita da seguinte maneira: veículos fabricados há menos de três anos, foram considerados seminovos; entre três e 10 anos, veículos usados e o restante foi classificado como veículos com mais de 10 anos de fabricação.
Estado Físico da Pessoa	Estado físico em que a pessoa se encontrava quando os agentes da PRF chegavam ao local do acidente.
Faixa Etária da Pessoa	Idade das pessoas envolvidas na ocorrência. Na base original, esse valor pode ser encontrado através da data de nascimento do envolvido, a partir dessa data calculou-se a idade. Em seguida, categorizou-se a idade em faixas etárias: entre 0 e 12 anos, atribuiu-se o valor 'criança'; entre 13 e 17 anos, 'adolescente'; entre 18 e 25, 'jovem; entre 26 e 59 anos, 'adulto'; pessoas com mais de 60 anos, 'idoso';
Tipo de Acidente	Tipo de acidente da ocorrência como, por exemplo, "atropelamento de pessoa".
Modelo da Pista	Modelo da pista do local do acidente. Por exemplo, se o acidente foi em uma reta ou em uma curva acentuada.
Período do Dia	Na base de dados original é disponibilizada a hora em que o acidente ocorreu, assim como outros campos descritos acima. Esse valor numérico não agrega valor aos algoritmos que foram utilizados no processo de descoberta de conhecimento. Assim, categorizou-se a hora do acidente da maneira como segue: ocorrências que aconteceram entre 6h e 11h receberam o valor 'manhã'; entre 12h e 18h, 'tarde'; entre 19h e 23h, 'noite' e as que ocorreram entre 0h e 5h, foram substituídas pelo valor 'madrugada'.
Dia da Semana	Nos boletins de ocorrências, o dia da semana do fato não é registrado, mas sim a data (dia, mês e ano) da ocorrência. Com essa informação, foi possível descobrir o dia da semana em que ocorreu o acidente.

Quadro 2. Atributos de descrição e uma descrição sobre seu conteúdo.
Fonte: autoria própria.

Atributo	Domínio do Atributo
Tipo de Veículo	Bicicleta, Ciclomotor, Motoneta, Motocicleta, Triciclo, Quadriciclo, Automóvel, Micro-ônibus, Ônibus, Bonde / Trem, Reboque, Semireboque, Charrete, Caminhão, Carroça, Carro de Mão, Caminhonete, Utilitário, Caminhão Trator, Trator de Rodas, Trator de Esteiras, Trator Misto, Camioneta, Caminhão Tanque, Não Identificado.
Ano do Veículo	Seminovo, Usado, Mais de 10 Anos.
Estado Físico da Pessoa	Ileso, Lesões Leves, Lesões Graves, Morto, Ignorado.
Faixa Etária da Pessoa	Criança, Adolescente, Jovem, Adulto, Idoso.
Tipo de Acidente	Atropelamento de Animal, Atropelamento de Pessoa, Capotamento, Colisão com Bicicleta, Colisão com Objeto Fixo, Colisão Frontal, Colisão Lateral, Colisão Traseira, Incêndio, Colisão Transversal, Tombamento, Saída de Pista, Derramamento de Carga, Colisão com Objeto Móvel, Queda de Motocicleta / Bicicleta / Veículo, Danos Eventuais.
Modelo da Pista	Reta, Curva Aberta, Ponte, Bifurcação, Bifurcação com Rotatória, Curva Acentuada, Curva Acentuada a Direita, Curva Diamante, Curva 180º, Sinuosa, Cruzamento, Cruzamento com Rotatória, Cruzamento com Viaduto, Cruzamento com Canteiro, Retorno 1, Retorno 2, Início / Fim de Pista Dupla, Vicinal, Vicinal Dupla, Saída, Cruzamento com viaduto 2.
Período do Dia	Manhã, Tarde, Noite, Madrugada.
Dia da Semana	Domingo, Segunda-feira, Terça-feira, Quarta-feira, Quinta-feira, Sexta-feira, Sábado.

Quadro 3. Atributos de descrição e seus respectivos domínios da base de dados processada – ano de 2012.
Fonte: autoria própria.

Com os dados pré-processados e transformados, realizou-se a etapa de extração de padrões. Foram utilizados os algoritmos PART (indução de regras de conhecimento), J48 (árvores de decisão) e *Apriori* (construção de regras de associação), descritos anteriormente[2]. Optou-se pela utilização de algoritmos de aprendizado simbólico, já que os classificadores induzidos por tais algoritmos podem ser transformados em conjuntos de regras proposicionais R = B → H, que são mais facilmente interpretadas por seres humanos (Bernardini, 2006).

Resultados e análise – algoritmos PART e J48

Na Tabela 2 são exibidos os resultados obtidos com os algoritmos J48 e PART para cada uma das medidas $mea(\mathbf{h}) \in \{err(\mathbf{h}), prec(\mathbf{h}), rec(\mathbf{h}), F(\mathbf{h}), AUC(\mathbf{h})\}$. Deve ser observado que todas as medidas foram estimadas utilizando a técnica de validação cruzada estratificada com 10 partições, descrita anteriormente. Como a classe majoritária conjunto de dados é "Falta de Atenção", com 46,3% dos exemplos, o erro majoritário da base de

[2] Deve ser observado que o Erro Majoritário de uma base de dados é o limite superior que a taxa de erro de um classificador deve atingir. Se a taxa de erro de um classificador for maior que o erro majoritário, o classificador é menos eficiente que predizer, para os exemplos futuros, a classe majoritária.

dados utilizada é de 53,7%, e a média das taxas de erro obtidas para os algoritmos J48 e PART são menores ou iguais a esse valor – $m_{err}(\mathbf{h}_{J48}) = 45{,}88\%$ e $m_{err}(\mathbf{h}_{J48}) = 46{,}36\%$ –, tais fatos indicam que os classificadores obtidos são melhores preditores do que classificadores dos exemplos na classe majoritária.

Na Tabela 3, foram marcados com o símbolo ▲ os casos em que o algoritmo apresenta melhor resultado em relação ao outro, com 95% de confiança segundo o teste t de *Student*, e com o símbolo ○ os casos em que ambos os algoritmos não apresentaram diferença estatística, também segundo o teste t de *Student*.

Algoritmo		$err(\boldsymbol{h})$		$prec(\boldsymbol{h})$		$rec(\boldsymbol{h})$		$F(\boldsymbol{h})$		$AUC(\boldsymbol{h})$	
J48	$m_{mea}(\mathbf{h}_{J48})$	45,88%	▲	46,2%	▲	8,9%		15%		83,4%	○
	$se_{mea}(\mathbf{h}_{J48})$	0,04%		1,0%		0,2%		0,4%		0,2%	
PART	$m_{mea}(\boldsymbol{h}_{PART})$	46,36%		42,2%		14,7%	▲	21,8%	▲	83,3%	○
	$se_{mea}(\boldsymbol{h}_{PART})$	0,07%		0,7%		0,3%		0,4%		0,2%	

Tabela 2. Resultados dos algoritmos J48 e PART.
Fonte: autoria própria.

Pode-se observar, nessa tabela, que o algoritmo J48 apresenta melhor comportamento para a base de dados utilizada para as medidas *err*(**h**) e *prec*(**h**). Já o algoritmo PART apresenta melhor comportamento em relação ao J48 para as medidas *rec*(**h**) e *F*(**h**). Ainda assim, para ambos os casos, observa-se que os valores das medidas *rec*(**h**) e *F*(**h**) são baixos, o que indica uma melhor análise em relação às predições nas classes. Sendo assim, observam-se também as medidas em cada uma das classes. Nas Tabelas 3 e 4 são mostradas a taxa de *TP* e de *FP* e as medidas $prec(C_v)$, $rec(C_v)$, $F(C_v)$ e $AUC(C_v)$ para os algoritmos J48 e PART, respectivamente, e para cada uma das classes $C_v \in C$, gerou-se um gráfico com esses dados, que podem ser visualizados nas Figuras 2 e 3.

Classe	Taxa de *TP*	Taxa de *FP*	$prec(C_v)$	$rec(C_v)$	$F(C_v)$	$AUC(C_v)$	% Ex.
Animais na Pista	0,758	0,001	0,943	0,758	0,84	0,924	2,4%
Defeito Mecânico no Veículo	0,34	0,014	0,536	0,34	0,416	0,777	4,4%
Defeito na Via	0,074	0,002	0,327	0,074	0,121	0,725	1,5%
Desobediência à Sinalização	0,089	0,007	0,462	0,089	0,15	0,834	6,5%
Falta de Atenção	0,918	0,67	0,542	0,918	0,681	0,675	2,9%
Ingestão de Álcool	0,107	0,012	0,346	0,107	0,163	0,675	46,3%
Motorista Dormindo	0,22	0,008	0,462	0,22	0,298	0,772	5,4%
Não Guardar Distância de Segurança	0,019	0,005	0,425	0,019	0,036	0,824	16,1%
Ultrapassagem Indevida	0,242	0,012	0,455	0,242	0,316	0,805	4,0%
Velocidade Incompatível	0,496	0,046	0,55	0,496	0,522	0,778	10,2%

Tabela 3. Resultados do algoritmo J48 em cada classe.
Fonte: autoria própria.

Classe	Taxa de *TP*	Taxa de *FP*	$prec(C_v)$	$rec(C_v)$	$F(C_v)$	$AUC(C_v)$	% Ex.
Animais na Pista	0,761	0,002	0,925	0,761	0,835	0,925	2,4%
Defeito Mecânico no Veículo	0,35	0,016	0,503	0,35	0,413	0,8	4,,4%
Defeito na Via	0,092	0,003	0,299	0,092	0,141	0,742	1,5%
Desobediência à Sinalização	0,147	0,014	0,421	0,147	0,218	0,833	6,5%
Falta de Atenção	0,839	0,576	0,557	0,839	0,669	0,688	2,9%
Ingestão de Álcool	0,166	0,019	0,337	0,166	0,223	0,772	46,3%
Motorista Dormindo	0,255	0,011	0,417	0,255	0,317	0,811	5,4%
Não Guardar Distância de Segurança	0,165	0,046	0,41	0,165	0,235	0,839	16,1%
Ultrapassagem Indevida	0,258	0,014	0,438	0,258	0,325	0,809	4,0%
Velocidade Incompatível	0,482	0,046	0,544	0,482	0,511	0,796	10,2%

Tabela 4. Resultados do algoritmo PART em cada classe.
Fonte: autoria própria.

Pode-se observar, nas Tabelas 3 e 4, e nos gráficos das Figuras 2 e 3, que, para as medidas Taxa de *FP*, $prec(C_v)$, $rec(C_v)$ e $F(C_v)$, somente a classe "Animais na Pista" apresentou valores altos, diferente das outras classes, que apresentaram valores baixos. Observa-se também que todas as classes apresentaram valores acima de 0,5 na medida de $AUC(C_v)$, indicando que houve aprendizado em cada uma das classes, apesar dos valores estarem abaixo de 0,8 em alguns casos. Ainda, apesar de "Defeito Mecânico no Veículo" ter somente 4,4% dos exemplos da base de dados, as taxas apresentadas para essa classe são promissoras, já que as medidas de *prec* e *AUC* apresentam valores maiores que 0,5. Por outro lado, a classe "Ingestão de Álcool" possui 46,3% dos exemplos da base de dados, mas ainda assim possui baixa taxa de *TP*, *prec*, *rec* e *F*, o que indica que, para essa base, a distribuição dos dados não tem impacto direto nos resultados obtidos. Adiante, ainda nesta seção, é apresentado um estudo mais aprofundado em relação ao aprendizado de classificadores simbólicos para cada uma das classes.

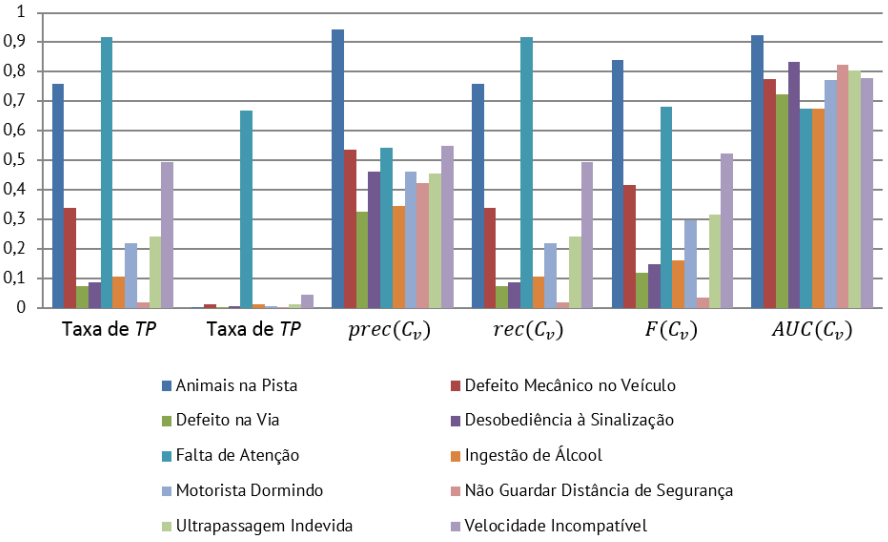


Figura 2. Valor das medidas por classe - algoritmo J48.
Fonte: autoria própria.

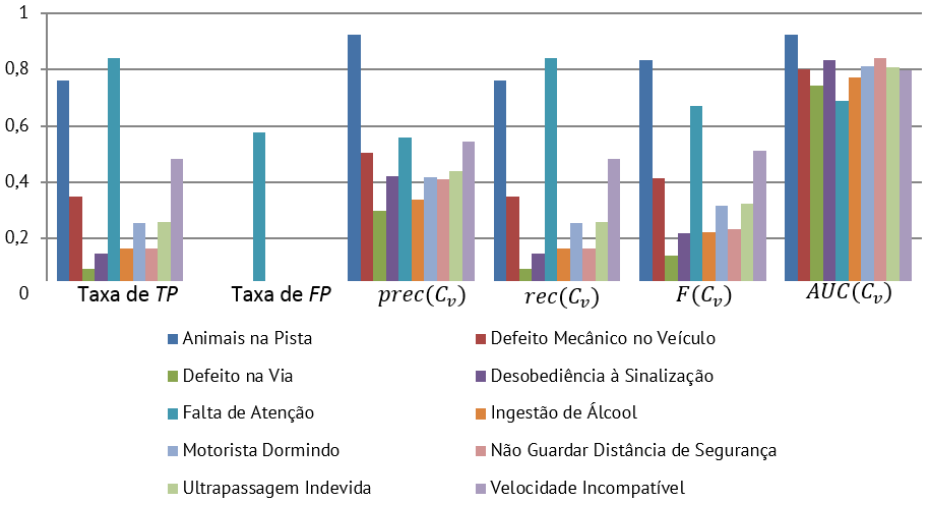


Figura 3. Valor das medidas por classe – algoritmo PART.
Fonte: autoria própria.

Pode-se observar também que, em relação ao número de regras criadas por ambos os algoritmos[3], o PART construiu um classificador com todos os exemplos de treinamento, que possui um total de 12.600 regras de decisão. Já o J48 construiu uma árvore de decisão com 14.311 nós folhas, ou seja, 14.300 regras de decisão.

Dentre as regras geradas pelo algoritmo J48, são mostradas, nos Quadros 4 e 5, as que foram julgadas mais interessantes. O critério de escolha foi considerar regras que traziam novos conhecimentos a respeito dos acidentes rodoviários e que, do ponto de vista dos autores, poderiam ser utilizadas pelas autoridades, e pela própria população, para a diminuição de acidente em rodovias federais. Entretanto, deve ser observado que

[3] Cada caminho do nó raiz até um nó de decisão, ou nó folha, de uma árvore de decisão pode ser reescrito como uma regra de decisão.

não houve uma validação junto aos órgãos competentes quanto a essa validade. Nos resultados do PART, foram selecionadas algumas regras que possuem alto valor de precisão e cobertura da regra, que são listadas na Figura 6. Nessas figuras, #Cob é o número de casos cobertos corretamente pela regra e #Incorr é o número de exemplos incorretamente cobertos pela regra.

- SE Tipo de Acidente = Capotamento E Modelo da Pista = Reta e Tipo de Veículo = Automóvel E Período do Dia = Manhã E Dia da Semana = Domingo E Estado Físico da Pessoa = Lesões Graves E Ano do Veículo = Usado ENTÃO Causa do Acidente = Ingestão de Álcool. (#Cob = 3,52; #Incorr = 2,52)
- SE Tipo de Acidente = Colisão com Objeto Fixo E Modelo da Pista = Reta E Tipo de Veículo = Motocicleta E Estado Físico da Pessoa = Lesões Graves E Período do Dia = Noite E Dia da Semana = Terça-feira ENTÃO Causa do Acidente = Falta de Atenção. (#Cob = 16; #Incorr = 1)
- SE Tipo de Acidente = Colisão com Objeto Fixo E Modelo da Pista = Reta E Tipo de Veículo = Automóvel E Estado Físico da Pessoa = Lesões Graves E Período do dia = Madrugada E Ano do Veículo = Mais de 10 Anos ENTÃO Causa do Acidente = Ingestão de Álcool. (#Cob = 8,94; #Incorr = 2,94)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Tipo de Veículo = Motocicleta ENTÃO Causa do Acidente = Falta de Atenção. (#Cob = 16; #Incorr = 7)

Quadro 4. Regras selecionadas do classificador induzido pelo algoritmo J48 (PARTE 1).
Fonte: autoria própria.

- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Período do Dia = Manhã E Dia da Semana = Segunda-feira E Tipo de Veículo = Automóvel E Estado Físico da Pessoa = Morto ENTÃO Causa do Acidente = Ultrapassagem Indevida. (#Cob = 19.25; #Incorr = 11)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Período do Dia = Noite E Dia da Semana = Sábado E Tipo de Veículo = Automóvel E Estado Físico da Pessoa = Morto E Ano do Veículo = Usado ENTÃO Causa do Acidente = Ultrapassagem Indevida. (#Cob = 16,13; #Incorr = 9,13)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Reta E Período do Dia = Noite E Dia da Semana = Quinta-feira E Tipo de Veículo = Micro-ônibus ENTÃO Causa do Acidente = Ultrapassagem Indevida. (#Cob = 27; #Incorr = 4)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Curva 180º ENTÃO Causa do Acidente = Velocidade Incompatível. (#Cob = 287,87; #Incorr = 75,04)
- SE Tipo de Acidente = Colisão Frontal E Modelo da Pista = Cruzamento E Tipo de Veículo = Motocicleta ENTÃO Causa do Acidente = Defeito Mecânico no Veículo. (#Cob = 13,11; #Incorr = 4,04)

Quadro 5. Regras selecionadas do classificador induzido pelo algoritmo J48 (PARTE 2).
Fonte: autoria própria.

Algumas regras obtidas trazem informações interessantes como, por exemplo a terceira regra do Quadro 6, que revela que acidentes com incêndio, ocorridos na quarta-feira de manhã, aconteceram por que o motorista dormiu ao volante. A quarta regra apresentada também chama a atenção para um fato muito comum em acidentes rodoviários: a falta de atenção. De acordo com os resultados da última regra, a maioria dos atropelamentos de pessoas em retas é causada por falta de atenção do motorista.

- SE Tipo de Acidente = Atropelamento de Animal E Modelo da Pista = Reta E Tipo de Veículo = Automóvel E Período do Dia = Noite ENTÃO Causa do Acidente = Animais na Pista. (#Cob = 1036; #Incorr = 17)
- SE Tipo de Acidente = Incêndio E Estado Físico da Pessoa = Ileso E Modelo da Pista = Reta ENTÃO Causa do Acidente = Defeito Mecânico no Veículo. (#Cob = 628; #Incorr = 7)
- SE Tipo de Acidente = Incêndio E Dia da Semana = Quarta-feira E Período do Dia = Manhã ENTÃO Causa do Acidente = Motorista Dormindo. (#Cob = 16)
- SE Tipo de Acidente = Atropelamento de Pessoa E Modelo da Pista = Reta E Período do Dia = Tarde e Tipo de Veículo = Automóvel ENTÃO Causa do Acidente = Falta de Atenção. (#Cob = 343; #Incorr = 85)

Quadro 6. Regras selecionadas da árvore de decisão induzida pelo algoritmo PART.
Fonte: autoria própria.

Devido ao significativo número de classes, foi avaliado também o comportamento dos algoritmos PART e J48 para cada uma das classes $C_v \in C$, em cada uma das medidas previamente mencionadas. Analogamente aos experimentos anteriores, utilizamos a técnica de validação cruzada estratificada com 10 partições. Foi gerado um conjunto de dados para cada classe $C_v \in C$, no qual os exemplos $\mathbf{x} \in C_v$ são rotulados como positivos, e os exemplos $\mathbf{x} \notin C_v$ são rotulados como negativos. Essa abordagem está relacionada à técnica de divisão de problemas de classificação denominada "um-contra-todos" (Facelli et al., 2011). Nas Tabelas 5 e 6 são apresentados os resultados obtidos para os algoritmos J48 e PART para cada uma das classes. Em cada uma dessas tabelas, é apresentada a média $m_{mea}(\mathbf{h})$ e o erro padrão $se_{mea}(\mathbf{h})$ para cada um dos algoritmos, onde $mea(\mathbf{h}) \in \{err(\mathbf{h}), prec(\mathbf{h}), rec(\mathbf{h}), F(\mathbf{h}), AUC(\mathbf{h})\}$. Na última coluna das tabelas, é apresentado também o número médio de regras dos classificadores induzidos em cada iteração da validação cruzada estratificada com 10 partições. Nessas tabelas, também foram marcados com o símbolo ▲ os casos em que o algoritmo apresenta melhor

resultado em relação ao outro, com 95% de confiança segundo o teste t de *Student*, e com o símbolo ○ os casos em que ambos os algoritmos não apresentaram diferença estatística, também segundo o teste t de *Student*. Ainda, foram marcados com * os resultados que estão iguais ou piores que o erro majoritário, indicando que o algoritmo não conseguiu aprender os conceitos nas classes minoritárias.

Nas Tabelas 5 e 6, pode-se observar que as taxas geradas por ambos os algoritmos são praticamente iguais. Com base nas taxas geradas pelo J48, pode-se observar que, para a classe "Animais na Pista", foram obtidos bons resultados em todas as medidas, já que todas as medidas apresentaram valores acima de 50%. Tal fato reforça o resultado para essa classe exibido para o algoritmo J48 considerando todas as classes em conjunto, cujo resultado é exibido na Tabela 6. Em relação à classe "Falta de Atenção", observa-se que, apesar dos exemplos dessa classe corresponderem a 46,3% dos exemplos do conjunto de dados, a taxa de erro é a mais alta dentre todas as taxas de erro. No entanto, todas as demais taxas estão acima de 50% para essa classe. É importante observar, na Tabela 6 (J48), as classes cujo número médio de regras foi aproximadamente 1. Isso indica que o classificador gerado é aquele que classifica todos os exemplos na classe majoritária, ou seja, não foi induzido nenhum conhecimento. Daí, para as classes "Defeito na Via", "Desobediência à Sinalização", "Ingestão de Álcool" e "Não Guardar Distância de Segurança", considera-se que o algoritmo J48 não é capaz de aprender conhecimento.

Classe C_v		$err(h)$		$prec(h)$		$rec(h)$	$F(h)$		$AUC(h)$	# Reg.
Animais na Pista	$m_{mea}(\mathbf{h}_{J48})$	0,68%	○	97,80%	▲	75,86%	84,15%	○	97,98%	16
	$se_{mea}(\mathbf{h}_{J48})$	0,01%		0,73%		0,45%	0,34%		0,21%	
Defeito Mecânico no Veículo	$m_{mea}(\mathbf{h}_{J48})$	3,87%	▲	76,60%	▲	18,74%	30,11%		77,77%	550
	$se_{mea}(\mathbf{h}_{J48})$	0,01%		0,67%		0,21%	0,29%		0,21%	
Defeito na Via	$m_{mea}(\mathbf{h}_{J48})$	1,18%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Desobediência à Sinalização	$m_{mea}(\mathbf{h}_{J48})$	6,54%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Falta de Atenção	$m_{mea}(\mathbf{h}_{J48})$	35,18%		63,92%	▲	55,15%	59,21%		69,16%	4541
	$se_{mea}(\mathbf{h}_{J48})$	0,07%		0,09%		0,26%	0,15%		0,08%	
Ingestão de Álcool	$m_{mea}(\mathbf{h}_{J48})$	5,11%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Motorista Dormindo	$m_{mea}(\mathbf{h}_{J48})$	2,78%	▲	76,74%	▲	7,81%	14,16%		72,55%	352
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		1,45%		0,29%	0,46%		0,42%	
Não Guardar Distância de Segurança	$m_{mea}(\mathbf{h}_{J48})$	16,15%	*	0,00%		0,00%	0,00%		50,00%	1
	$se_{mea}(\mathbf{h}_{J48})$	0,00%		0,00%		0,00%	0,00%		0,00%	
Ultrapassagem Indevida	$m_{mea}(\mathbf{h}_{J48})$	3,87%		67,17%	▲	8,07%	14,10%		80,53%	677
	$se_{mea}(\mathbf{h}_{J48})$	0,01%		0,74%		0,31%	0,49%		0,26%	
Velocidade Incompatível	$m_{mea}(\mathbf{h}_{J48})$	8,40%	▲	67,36%	▲	54,65%	45,77%		76,38%	2479
	$se_{mea}(\mathbf{h}_{J48})$	0,06%		0,52%		0,30%	0,38%		0,20%	

Tabela 5. Resultados do algoritmo J48 para cada classe C_v.
Fonte: autoria própria.

Já na Tabela 6, cujos resultados são referentes ao algoritmo PART, pode-se observar que as taxas de erro estão levemente mais altas que as do J48, isso significa que o algoritmo PART conseguiu gerar regras, porém regras "ruins".

Classe C_v		$err(h)$		$prec(h)$		$rec(h)$		$F(h)$		$AUC(h)$		# Reg.
Animais na Pista	$m_{mon}(h_{PART})$	0,67%	◦	97,77%		74,94%	▲	84,63%	◦	93,90%	▲	211
	$se_{mon}(h_{PART})$	0,01%		0,77%		0,12%		0,30%		0,24%		
Defeito Mecânico no Veículo	$m_{mon}(h_{PART})$	3,90%		67,18%		23,88%	▲	35,22%	▲	80,98%	▲	1255
	$se_{mon}(h_{PART})$	0,01%		0,58%		0,20%		0,20%		0,17%		
Defeito na Via	$m_{mon}(h_{PART})$	1,48%	•	47,07%	▲	2,46%	▲	4,67%	▲	76,37%	▲	474
	$se_{mon}(h_{PART})$	0,01%		3,55%		0,26%		0,47%		0,52%		
Desobediência à Sinalização	$m_{mon}(h_{PART})$	6,60%	•	47,59%	▲	8,54%	▲	14,47%	▲	84,14%	▲	1581
	$se_{mon}(h_{PART})$	0,02%		0,83%		0,24%		0,37%		0,11%		
Falta de Atenção	$m_{mon}(h_{PART})$	2,77%	▲	62,74%		56,95%	▲	59,70%	▲	68,93%	▲	6415
	$se_{mon}(h_{PART})$	0,01%		0,10%		0,13%		0,09%		0,07%		
Ingestão de Álcool	$m_{mon}(h_{PART})$	35,59%	•	13,52%	▲	5,57%	▲	9,86%	▲	79,19%	▲	1474
	$se_{mon}(h_{PART})$	0,08%		1,02%		0,21%		0,35%		0,19%		
Motorista Dormindo	$m_{mon}(h_{PART})$	5,55%		61,89%		14,67%	▲	25,70%	▲	84,86%	▲	906
	$se_{mon}(h_{PART})$	0,02%		0,86%		0,34%		0,48%		0,26%		
Não Guardar Distância de Segurança	$m_{mon}(h_{PART})$	16,35%	•	42,96%	▲	3,82%	▲	7,02%	▲	84,19%	▲	1521
	$se_{mon}(h_{PART})$	0,01%		0,52%		0,13%		0,22%		0,08%		
Ultrapassagem Indevida	$m_{mon}(h_{PART})$	3,81%	▲	60,71%		13,80%	▲	22,18%	▲	82,62%	▲	1066
	$se_{mon}(h_{PART})$	0,02%		0,78%		0,55%		0,55%		0,50%		
Velocidade Incompatível	$m_{mon}(h_{PART})$	8,52%		64,59%		56,95%	▲	47,01%	▲	81,21%	▲	2613
	$se_{mon}(h_{PART})$	0,05%		0,45%		0,22%		0,26%		0,14%		

Tabela 6. Resultados do algoritmo PART para cada classe C_v.
Fonte: autoria própria.

Resultados e análise – algoritmo APRIORI

Foram geradas 38 regras de associação com confiança maior que 0,8. No Quadro 7 são listadas aquelas com maior valor de confiança. Valores de confiança menores que esse valor gerava um número de regras bastante grande para serem analisadas. Por outro lado, aumentando o valor de confiança para 90%, apenas duas regras, listadas na Quadro 8, foram geradas. Em ambos os casos, o suporte mínimo utilizado foi de 0,1.

- SE Tipo de Veículo = Automóvel E Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Traseira E Modelo da Pista = Reta ENTÃO Estado Físico da Pessoa = Ileso. (Conf = 93%)
- SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Tipo de Acidente = Colisão Traseira. (Conf = 88%)
- SE Ano do Veículo = Seminovo E Tipo de Acidente = Colisão Traseira ENTÃO Estado Físico da Pessoa = Ileso. (Conf = 86%)
- SE Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Lateral ENTÃO Estado Físico da Pessoa = Ileso (Conf 86%)
- SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Estado Físico da Pessoa = Ileso (Conf 86%)
- SE Tipo de Acidente = Colisão Traseira E Período do Dia = Manhã ENTÃO Estado Físico da Pessoa = Ileso (Conf = 85%)
- SE Tipo de Acidente = Colisão Traseira E Modelo da Pista = Reta E Período do Dia = Tarde ENTÃO Estado Físico da Pessoa = Ileso (Conf = 85%)
- SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Modelo da Pista = Reta (Conf = 82%)

Quadro 7. Resultados do algoritmo *Apriori* com confiança maior que 0,8.
Fonte: autoria própria.

- SE Tipo de Veículo = Automóvel E Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Traseira E Modelo da Pista = Reta ENTÃO Estado Físico da Pessoa = Ileso (Conf = 93%)
- SE Tipo de Veículo = Automóvel E Faixa Etária da Pessoa = Adulto E Tipo de Acidente = Colisão Traseira ENTÃO Estado Físico da Pessoa = Ileso (Conf = 93%)

Quadro 8. Resultados do algoritmo *Apriori* com confiança maior que 0,9.
Fonte: autoria própria.

Nas regras obtidas pelo *Apriori*, pode-se observar algumas interessantes, como, p. ex., a regra "SE Causa do Acidente = Não Guardar Distância de Segurança ENTÃO Tipo de Acidente = Colisão Traseira", que relaciona a falta de distância de segurança com uma pista reta em acidentes. Deve ser observado que tal relação está presente em 82% dos casos de acidentes nas rodovias federais do País.

CONCLUSÕES

Tendo em vista que o papel principal dos Dados Abertos Governamentais (DAG) é auxiliar no processo de criação de um governo mais transparente e participativo - tornando as informações mais compreensíveis e próximas dos cidadãos – é necessário que tais dados sejam disponibilizados seguindo um padrão aceito internacionalmente e que possibilite sua ampla reutilização, tanto por máquinas quanto por humanos. Entretanto, observam-se diversos problemas nos dados utilizados neste trabalho, considerados de extrema importância não somente para a população em geral, como para novas medidas e decisões governamentais relação aos acidentes rodoviários que acontecem diariamente nas rodovias federais brasileiras. Revisar o processo de inserção e de coleta desses dados, assim como o modelo dos mesmos, pode melhorar o resultado do processo de MD, uma vez que os diversos erros encontrados fizeram com que a confiabilidade e a qualidade dos resultados gerados pelos algoritmos não tenham sido de todo satisfatórias (ainda que regras interessantes tenham sido observadas). Porém, dada a natureza dos dados e os resultados gerados, conclui-se que a experimentação de novos algoritmos, como os que consideram a incerteza – por exemplo, aqueles que utilizem redes Bayesianas – pode ser válida para a obtenção de melhores resultados.

É importante ressaltar que os dados disponíveis no Portal Brasileiro de Dados Abertos e, consequentemente, os dados utilizados nesse trabalho, seguem o segundo princípio dos DAG, o que determina que tais dados sejam disponibilizados em sua forma primária, ou seja, tais como são coletados na fonte. Porém, essa forma de disponibilizar os dados, os torna mais difíceis de serem entendidos por homem e por máquina. Dados sem qualidade e sem padrões dificultam o processo de MD e, até mesmo, a sua reutilização em outros segmentos. Baseado nos resultados encontrados sugere-se que a forma como os dados estão sendo disponibilizados seja amplamente revista e discutida, bem como seja criado ou utilizado um padrão na forma de publicá-los. Uma solução simples seria a disponibilização destes dados de duas formas diferentes: a primeira respeitando os princípios dos DAG; e, como segunda solução, disponibilizando os dados públicos em um formato padronizado e amplamente utilizado (tais como triplas RDF, por exemplo).

Foi possível observar que alguns resultados obtidos com os algoritmos J48 e PART são promissores em relação à classificação das causas de acidentes. Os valores obtidos de área sob a curva ROC (AUC) estiveram acima de 0,5 e, ao se utilizar o algoritmo *Apriori*, foram geradas 38 regras de associação com confiança maior que 0,8. Porém, a baixa taxa de precisão dos classificadores gerados indica que há a necessidade de maior exploração nos dados para tentar extrair melhores resultados no processo. Ainda, ferramentas de visualização das estatísticas dos dados também podem ser interessantes e enriquecer os resultados encontrados, facilitando a sua reutilização e o entendimento de todos os cidadãos, o que auxiliaria a alcançar uma audiência mais ampla e, consequentemente, os objetivos dos DAG.

Observa-se também que o acesso e a interpretação desses dados não é uma tarefa trivial para o cidadão. Uma maneira de contornar esse problema é disponibilizar os dados de tal maneira que sistemas computacionais possam interpretá-los, bem como a construir interfaces mais simples para a visualização destes dados. Esse é um dos princípios da *Web* Semântica, uma extensão da *web* atual que provê um framework, composta por diversas tecnologias, dentre elas o *Resource Description Framework* (RDF), o OWL, e o SPARQL, para permitir que dados sejam compartilhados e reutilizados por aplicações, empresas e comunidades (Breitman, 2005). A *Web* Semântica fornece um ambiente onde uma aplicação pode consultar esses dados, realizar inferências usando vocabulários específicos de domínio, extrair padrões, etc. Como trabalho futuro, pretende-se propor um método para coletar os dados da base da PRF, em seu formato original, e disponibilizá-los em triplas RDF utilizando uma ontologia para modelagem.

REFERÊNCIAS

Agrawal, R., Imielinski, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *ACM Sigmod Conference*. Retirado de http://www.it.uu.se/edu/course/homepage/infoutv/ht08/agrawal93mining.pdf

Agune, R. M., Gregorio Filho, A. S., & Bolliger, S. P. (2010). Governo aberto SP: disponibilização de bases de dados e informações em formato aberto. *Congresso Consad de Gestão Pública*. Retirado de http://www.prefeitura.sp.gov.br/cidade/secretarias/upload/controladoria_geral/arquivos/C3_TP_GOVERNO%20ABERTO%20SP%20DISPONIBILIZACAO%20DE%20BASES%20DE%20DADOS.pdf

Balbo, F. A. N. (2011). *Análise multivariada aplicada aos acidentes da BR-277 entre janeiro de 2007 e novembro de 2009*. (Dissertação de Mestrado em Métodos Numéricos em Engenharia). Universidade Federal do Paraná. Retirado de http://www.ppgmne.ufpr.br/arquivos/diss/239.pdf

Baranauskas, J. A., & Monard, M. C. (2000). *Reviewing some machine learning concepts and methods*. Relatórios Técnicos do ICMC/USP, 102.

Bernardini, F. C. (2006). *Combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéricos*. (Tese de Doutorado em Ciências – Ciências de Computação e Matemática Computacional). Universidade de São Paulo/São Carlos. Retirado de http://www.teses.usp.br/teses/disponiveis/55/55134/tde-29092006-110806/

Berry, M. J. A., & Linoff, G. (©1997). *Data mining techniques: For marketing, sales, and customer support*. New York: John Wiley & Sons.

Borgelt, C., & Kruse, R. (2002). Induction of association rules: Apriori implementation. *15th Conference on Computational Statistics*. Retirado de http://www.borgelt.net/papers/cstat_02.pdf

Brasil. Ministério da Justiça. (2014a). *Sistema BR-Brasil: boletins de ocorrências em rodovias federais*. Retirado de http://dados.gov.br/dataset/acidentes-rodovias-federais

Brasil. Portal Brasileiro de Dados Abertos. (2014b). *O que são Dados Abertos? 2014*. Retirado de http://www.governoeletronico.gov.br/acoes-e-projetos/Dados-Abertos

Breitman, K. (2005). *Web semântica: a Internet do futuro*. Rio de Janeiro: LTC.

Carvalho, J. V., Sampaio, M. C., & Mongiovi, G. (1999). Utilização de técnicas de "Data Mining" para o reconhecimento de caracteres manuscritos. *14º Simpósio Brasileiro de Bancos de Dados*, 235-249. Retirado de http://www.dsc.ufcg.edu.br/~sampaio/Artigos/reconhecimentocaracteresmanuscritos.pdf

Domingos, P. A. (2012). Few useful things to know about machine learning. *Communications of the ACM*, 55(10), 78-87. Retirado de http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf

Facelli, K., Lorena, A. C., Gama, J., & Carvalho, A. C. P. L. F. (2011). *Inteligência Artificial: Uma abordagem de aprendizado de máquina*. Rio de Janeiro: LTC.

Frank, E., & Witten, I. H. (1998). *Generating accurate rule sets without global optimization*. Hamilton, New Zealand: University of Waikato.

Mitchell, T. (1997). *Machine Learning*. New York: McGraw Hill.

Quinlan, J. R. (1988). Decision trees and multi-valued attributes. In: Hayes, J. E., Michei, D., & Richards, J. (Orgs.). *Machine Intelligence*, 11. New York: Oxford University. Retirado de http://aitopics.org/sites/default/files/classic/Machine_Intelligence_11/MI11-Ch13-Quinlan.pdf

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Francisco: Morgan Kaufmann.

Reis, C. V. R. (2013). *O uso da descoberta de conhecimento em Banco de Dados nos acidentes da BR-381*. (Projeto de pesquisa – Mestrado Profissional em Sistemas de Informação e Gestão do Conhecimento). Universidade FUMEC. Retirado de http://www.fumec.br/revistas/sigc/article/view/1508

Rezende, S. O., Pugliesi, J. B., Melanda, E. A., & Paula, M. D. (2003). Mineração de dados. In: REZENDE, S.O. (Org.). *Sistemas inteligentes: Fundamentos e aplicações*. São Paulo: Manole.

The Annotated 8 principles of Open Government Data. (2014). Retirado de http://opengovdata.org/

Witten, I. H., & Frank, E. (2009). *Data Mining: Practical machine learning tools and techniques with java implementations*. Burlington, Massachusetts: Morgan Kaufmann.

Como citar este artigo (ABNT):

COSTA, J. de J.; BERNARDINI, F. C.; VITERBO FILHO, J. A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 139-157, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this article* (APA):**

Costa, J. J., Bernardini, F. C., & Viterbo Filho, J. (2014). A mineração de dados e a qualidade de conhecimentos extraídos dos boletins de ocorrência das rodovias federais brasileiras. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 139-157. Retrieved from http://www.atoz.ufpr.br

ANEXO – Descrição das tabelas do DER da base de dados utilizada

LOCALBR: armazena o local onde ocorreu a ocorrência, ou seja, qual a BR onde foi registrado o acidente. Além disso, identifica o estado onde ocorreu o acidente. Algumas informações dessa tabela (tais como em qual quilômetro da BR ocorreu o acidente, altitude e longitude da ocorrência), seriam úteis para o processo de descoberta de conhecimento. Porém estes são campos que nem sempre são preenchidos. Esta é uma tabela de domínio e está associada à tabela OCORRENCIAACIDENTE e UNIDADEOPERACIONAL. Colunas:

- lbrid - identifica a chave primaria da tabela;
- lbruf - identifica a UF da ocorrência;
- lbrbr - identifica a BR da ocorrência;
- lbrkm - identifica o KM da ocorrência;
- lbrlatitude - identifica a latitude da ocorrência;
- lbrlongitude - identifica a longitude da ocorrência;
- lbrpnvid - identifica a da ocorrência;
- lbratualiza - registra a atualização do local da ocorrência;

OCORRENCIA: essa tabela contém o registro de ocorrências confirmadas a partir das comunicações recebidas. A maioria das tabelas do sistema tem algum tipo de relacionamento com essa tabela. Colunas:

- ocoid - identificação única da ocorrência;
- ocolocal - identificação do local da BR onde aconteceu a ocorrência;
- ocostatus – registra o status da Ocorrência (Aberta, Encerrada, Anulada, Estatística, Retificada e Em Processo)
- ocomunicipio - identificação do município da ocorrência;
- ocosentido - identificação do sentido da via (Crescente ou Decrescente);
- ocodataocorrencia - data da ocorrência;
- ocodataregistro - data do registro da ocorrência;
- ocotipo - tipo da ocorrência (Acidentes Rodoviários, Retenção, Apreensão e Recuperação de Veículos; Pessoa Detenção/Auxílio; Apreensão de CNH; Apreensão de Documento; Apreensão de Carga; Interdição de Rodovias; Ocorrências);
- ococomid - chave estrangeira que identifica a comunicação;
- ocoidorigem – FK (foreign key) da ocorrência retificada;
- ococpfretif - CPF que identifica o executar da retificação;
- ocodatafim - data de fim da ocorrência.

OCORRENCIAVEICULO: cadastro do veículo da pessoa envolvida na ocorrência. Colunas:

- ocvid - identificação única da ocorrência com veículo;
- ocvocoid - chave estrangeira que identifica a ocorrência;
- ocvveiid - chave estrangeira que identifica o veículo;

VEICULO: tabela que contém o cadastro dos dados do veículo. Colunas:

- veiid - identificação única do veículo;
- veiano - ano do veículo;
- veiqtdocupantes - quantidade de ocupantes do veículo;
- veimunicipio - município do veículo;
- veimunorigem - município de origem do veículo;
- veipaisorigem - país de origem do veículo;
- veimundestino - município de destino do veículo;
- veipaisdestino - país de destino do veículo;
- veiproprietario - identificação do tipo de proprietário;
- tvvcodigo - identifica a chave primária da tabela tipoveiculo (chave desligada);

TIPOVEICULO: tabela que identifica os tipos de veículos envolvidos nos acidentes. Esta é uma tabela de domínio e está associada à tabela VEICULO. Colunas:

- tvvcodigo - identificador único do tipo do veículo;
- tvvatualiza - identifica se o registro pode ser atualizado (manutenção de histórico);
- tvvativo – indica se o veículo está ativo.

OCORRENCIAPESSOA: registro das pessoas envolvidas no acidente. Colunas:

- opeid - identificador único da ocorrência de pessoas;
- opeocoid - chave estrangeira que identifica a ocorrência;
- opepesid - chave estrangeira que identifica a pessoa;
- opeportevalidade - validade do porte de arma;
- opettecodigo - chave estrangeira que identifica o tipo de envolvido;
- opeanexo - declaração em anexo('S','N');
- opecondalegadas - condições alegadas para a ocorrência.

PESSOA: tabela que registra os dados da pessoa envolvida na ocorrência. Colunas:

- pesid - identificador único de pessoa;
- pesexpedidor - órgão expedidor;
- pesufexpedidora - UF expedidora;
- pesdatanascimento – data de nascimento;
- pesnaturalidade – naturalidade;
- pesnacionalidade – nacionalidade;
- pessexo – sexo;
- pesteccodigo - chave estrangeira que identifica o estado civil;
- pestgicodigo - chave estrangeira que identifica o grau de instrução das pessoas;
- pesmunicipio – munícipio de residência;
- pescep – CEP;
- pestopcodigo - chave estrangeira que identifica a ocupação principal da pessoa;
- pesmunicipioori - município de origem da pessoa;
- pespaisori – país de origem da pessoa;

- pesmunicipiodest - município de destino da pessoa;
- pespaisdest – país de destino da pessoa;
- pesveiid - chave estrangeira que identifica o veículo da pessoa;
- pescinto - identifica se a pessoa estava utilizando cinto;
- pescapacete - identifica se a pessoa estava usando capacete;
- peshabilitado - identifica se a pessoa é habilitada;
- pessocorrido - identifica se a pessoa foi socorrida;
- pesdormindo - identifica se a pessoa estava dormindo;
- pesalcool - identifica se a pessoa estaca alcoolizada;
- peskmpercorre - indica quantos quilômetros ela percorreu;
- peshorapercorre - identifica o tempo que ela percorreu a quilometragem;
- pescategcnh - identifica a categoria da CNH;
- pesregistrocnh - numero do registro da CNH;
- pesufcnh - UF em que tirou a habilitação;
- pespaiscnh – país onde tirou a CNH;
- pesdatahabil – data da habilitação;
- pesdatavalidade – validade da habilitação;
- pesapelido - apelido atribuído a pessoa;
- pesidade – idade da pessoa;
- pesaltura – altura;
- pespeso – peso da pessoa;
- pescicatriz - identifica se possui cicatriz;
- pestatuagem - identifica se possui tatuagem;
- pessinal - identifica se a pessoa possui sinal;
- peslesao - identifica se a pessoa possui lesão;
- pestcccodigo - chave estrangeira que identifica a cor cabelo;
- pestctcodigo - chave estrangeira que identifica a cor da cútis;
- pestclcodigo - chave estrangeira que identifica a cor do olho;
- pespertences - descreve os pertences das pessoas no local da ocorrência;
- pesdadoscomplement - dados complementares das pessoas;
- vestigios_drogas - indicador de vestígios de droga;
- descricao_cicatriz – descrição da cicatriz da pessoa;
- descricao_tatuagem - descrição da tatuagem da pessoa
- descricao_sinal - descrição dos sinais da pessoa
- descricao_deficiencia – descrição da deficiência física da pessoa.

OCORRENCIAACIDENTE: cadastro da ocorrência envolvendo veículos. Colunas:
- oacocoid - identificador único do acidente;
- oacttacodigo - chave estrangeira que identifica o tipo de acidente;
- oactcacodigo - chave estrangeira que identifica a causa do acidente;
- oacdano - dano causado no acidente;
- oacdanoterc - dano causado a terceiro;
- oacdanoamb - dano causado ao acidente;
- oacrefera - referência ponto A (descreve o ponto de referencia);
- oacreferb - referência ponto B (descreve o ponto de referencia);
- oacdistab - distância entre os pontos A e B;
- oacdistac - distância entre os pontos A e C;
- oacdistbc - distância entre os pontos B e C; oacmodelopista - identificação do modelo de pista;
- oacsentido1 - descrição do sentido 1;
- oacsentido2 - descrição do sentido 2;
- oacqtdfaixa1 - quantidade de faixas no sentido 1;
- oacqtdfaixa2 - quantidade de faixas no sentido 2;
- oacacostamento1 - indicador de acostamento no sentido 1;
- oacacostamento2 - indicador de acostamento no sentido 2;
- oacimagem - indicador da existência de imagem;
- oacdescdanopat - descrição do dano causado ao patrimônio;
- oacdescdanoterc - descrição dos danos causados a terceiros;
- oacdescdanoamb - descrição dos danos causados ao ambiente;
- oaccanteiro - descreve se o local da ocorrência possui ou não canteiro;
- oaclinhacentral - descreve se a pista possui ou não linha central;
- oacorientpista - descreve se o acidente aconteceu no sentido crescente ou decrescente da pista de rolamento;
- oacversaocroqui - informa se foi realizado ou não um croqui para ocorrência.

CAUSAACIDENTE: qualifica as causas do acidente. Colunas:
- tcacodigo – identificador da causa do acidente;
- tcadescricao – descrição da causa do acidente (Velocidade incompatível, Ultrapassagem indevida, Ingestão de álcool, Desobediência à sinalização, Defeito mecânico, Defeito na via, Falta de atenção, Dormindo, Animais na pista, Não guardar distância de segurança, Outras, Não informado)

TIPOACIDENTE: qualifica os tipos de acidente. Colunas:
- ttacodigo - identificador único do tipo de acidente;
- ttaatualiza - indica se o registro permite atualização;
- ttaativo - indica se o registro está ativo.

AtoZ: novas práticas em informação e conhecimento
Av. Prefeito Lothário Meissner, 632 - Campus III
Jardim Botânico
80210-170 - Curitiba, PR - Brasil
www.atoz.ufpr.br | revistaatoz@ufpr.br