PICTOREA: UM MÉTODO PARA AUXÍLIO NA CONDUÇÃO DE PROCESSOS DE DATA SCIENCE

André Montevecchi
Pontificia Universidade Católica de Minas Gerais, MG, Brasil
Luis Enrique Zárate
Pontificia Universidade católica de Minas Gerais, MG, Brasil

RESUMO

Atualmente, dentro do contexto de Ciência dos Dados, o processo de descoberta de conhecimento em banco de dados é bastante utilizado na academia e relativamente pouco utilizado no mercado e indústria. As organizações que utilizam esse processo geralmente o fazem adquirindo softwares com metodologias definidas por esses ambientes. Porém, grande parte das aplicações desse processo é feita utilizando metodologias próprias do responsável pelo processo. Essas metodologias geralmente não seguem um padrão. Através do método científico interpretativista, dos conceitos de Domain-Driven Data Mining - D3M, e com auxílio da Metodologia para Modelagem de Processos (Business Process Management - BPM) neste trabalho é proposto um método canônico com caráter pedagógico, denominado PICTOREA, para desenvolvimento, acompanhamento e documentação das etapas e atividades de projetos de descoberta de conhecimento em banco de dados. Resultados da aplicação do método são apresentados para reforçar a relevância do método proposto.

Palavras-chave: Ciência dos dados; Descoberta de Conhecimento em Bases de Dados; Mineração de Dados, Metodologias para KDD; Mineração de dados por manipulação de domínio

ABSTRACT

Nowadays, in the Data Science context, the process of knowledge discovery in databases (Knowledge Discovery in Databases - KDD) is widely used in academia and yet little used in the market and industry. Organizations using KDD process usually do with acquiring software and methodologies defined by its environments. However, most KDD applications are made using own methodologies. These methodologies generally do not follow a pattern. Through interpretive scientific method, concepts of Domain-Driven Data Mining - D3M e with the aid of Methodology for Process Modeling (Business Process Management - BPM), this paper proposed a pedagogic canonical method, called PICTOREA for developing, monitoring and documentation of the steps and activities of a project KDD. Results of application of method are presented to enhance the relevance of the proposed method.

Keywords: Data Science; Knowledge Discovery in database; Data mining, KDD methodology; Domain-Driven Data Mining

[1] http://www.kdnuggets.com/polls/

1. INTRODUÇÃO

Atualmente existe uma grande quantidade de dados sendo gerados a todo instante por meio de diferentes fontes tais como a internet e seus serviços, sensores (fixos e móveis), dispositivos de captação de mídias, telecomunicações, consumo, financeiros, meteorológicos, dentre muitos outros. Qualquer setor, seja industrial, financeiro, do agro-negócio, de saúde, de serviços, transportes, logística, planejamento urbano, consumo de energia, climatologia, meteorologia, etc., estão interessados em como explorar esse volume gigantesco de dados. Este novo cenário tem propiciado o surgimento do conceito de Ciência dos dados (DS - *Data Science*) (Dhar, 2013), que envolve os principais fundamentos que sustentam um processo de descoberta de conhecimento e geração de informação útil para a tomada de decisão a partir de fontes diversas de dados. Esse novo conceito tem uma forte relação com a mineração de dados (DM - *Data Mining*), desde que este último envolve técnicas e algoritmos para a identificação de padrões e extração de conhecimento. O detalhamento desses métodos, procedimentos e técnicas caracterizam e fundamentam o conceito de *Data Science*.

Nos setores de Tecnologia da Informação (TI) das empresas, a padronização dos processos traz significantes benefícios como a interoperabilidade, a garantia de qualidade, a compatibilidade e a agilidade de comunicação. No entanto, dentro do cenário da ciência dos dados é uma realidade a falta de padronização na condução de processos de Descoberta de Conhecimento em Banco de dados (da sigla em inglês KDD – *Knowledge Discovery in DataBase*) (Fayyad et al., 1996; Fayyad & Stolorz, 1997; Kaber & Han, 2001; Kantardzic, 2003). Esta situação leva à ausência de documentação das ações aplicadas nas diversas etapas, e das decisões tomadas durante um processo de descoberta de conhecimento.

Segundo documento que relata os 10 principais desafios em KDD (Yang & Wu, 2006), um deles aponta para a falta de um método unificado que permita aos responsáveis por projetos de DS aplicarem com sucesso um processo KDD. Apesar da divulgação desse desafio já ter uma década, poucas contribuições tem sido feitas nessa direção. É possível observar que normalmente as metodologias aplicadas aos processos de KDD são próprias e específicas ao problema sendo tratado, o que tem dificultado o desenvolvimento e adoção de uma teoria unificada acerca do método. É possível observar que erroneamente muitos responsáveis por projetos de BI - *Business Intelligence* associam um processo de descoberta de conhecimento ao uso exclusivo de ferramentas. Em Yang e Wu (2006) é exposto que os softwares disponíveis para mineração de dados possuim uma boa interface para o usuário mas não possuim intrinsecamente um método que sustente as ações tomadas pelos mesmos. Encontrar um método canônico para processos de KDD é um desafio latente para os pesquisadores da área de ciência dos dados.

[1] http://www.kdnuggets.com/polls/

De acordo com pesquisas de opinião realizadas[1], a metodologia mais utilizada é a CRISP-DM, vinculada à ferramenta de mineração de dados mais vendida do mercado, a SPSS-Clementine, hoje de propriedade da IBM (sobre o nome IBM SPSS Modeler). A metodologia SEMMA, desenvolvida pela SAS, foi a terceira metodologia mais utilizada. O uso de metodologias próprias, propostas pelo responsáveis pela condução de um processo KDD, foi a segunda metodologia mais utilizada, cenário que ainda se mantêm na atualidade. Essas pesquisas mostram que nem sempre as ferramentas de mercado atendem e suportam as necessidades para condução de um processo de KDD. Há especificidades referentes ao domínio do problema e ao banco de dados que requerem uma abordagem própria que não são representadas e suportadas pelas ferramentas e metodologias a estas vinculadas. Existe então um campo fértil para propor um método geral que permita conduzir um processo de descoberta de conhecimento e explorar de forma mais eficiente os recursos computacionais oferecidos pelas ferramentas existentes. Esse método deve ser suficientemente detalhado para permitir a condução de projetos de ciência dos dados por parte de profissionais menos experientes, sendo este aspecto o principal objetivo da presente proposta.

Segundo Cao (2005,2007,2010), atualmente existe uma grande preocupação por processos de KDD corretamente executados e que sejam aderentes à realidade e necessidades das empresas. Com o intuito de atender a essa preocupação tem sido proposto um novo conceito: Mineração de Dados Orientada ao Domínio (D3M - *Domain-Driven Data Mining*) que tem por objetivo tornar os aspectos e demandas das organizações presentes durante toda a condução de um processo KDD. O paradigma D3M induze ao surgimento de uma nova geração de metodologias para aplicação de processos de KDD orientadas ao negócio. De acordo ao conceito D3M cada etapa do processo KDD deve ser acompanhada e validada por um especialista de domínio. Segundo Cão (2010), com a colaboração entre especialistas de TI, especialistas em KDD e especialistas no domínio do problema, um projeto de ciência dos dados pode-se tornar mais aderente e útil para apóio à tomada de decisão.

Frente ao cenário do uso de metodologias próprias, da necessidade do correto desenvolvimento de um projeto de ciência dos dados, e considerando as necessidades das empresas em projetos aderentes aos interesses das mesmas, abre-se um espaço potencialmente fértil para a proposta da padronização dos processos de KDD. O uso de um processo padronizado que leve em conta os princípios de D3M pode orientar melhor especialistas de domínio e de mineração de dados na condução do processo. Isso permitiria a condução de projetos de DS de forma organizada, viabilizando o controle de interações, adaptações e alterações às quais todo projeto de TI estão sujeito.

Neste trabalho é proposto um método canônico para descoberta de conhecimento em banco de dados denominado PICTOREA, que preferimos

[1] http://www.kdnuggets.com/polls/

defini-lo como um método pedagógico, uma vez que o próprio fluxo mantém a concisão das etapas do projeto e pode ser acompanhada por profissionais menos experientes, num cenário industrial ou acadêmico.

Este artigo está dividido em cinco seções. Na segunda seção, uma revisão da literatura e discussão acerca do cenário atual do processo KDD é apresentada. Na terceira seção a metodologia para desenvolvimento e o método PICTOREA são apresentados. Na quarta seção, resultados da aplicação do método como ferramenta pedagógica em grupo de estudantes de um curso de TI são apresentados e discutidos. Finalmente as conclusões e trabalhos futuros são apresentados.

2. REVISÃO DA LITERATURA

2.1 UMA DISCUSSÃO SOBRE PROJETOS DE KDD

Na sua concepção inicial, segundo Fayyad et. al. (1996), o processo KDD possui 5 etapas principais: seleção de dados, pré-processamento, transformação, mineração de dados e interpretação. Todas em conjunto buscam um objetivo, a descoberta de conhecimento útil e não óbvio. As diversas metodologias como CRISP-DM e SEMMA, vinculadas às suas respectivas ferramentas, obedecem ou adaptam essas 5 etapas. A experiência prática mostra que existem outras sub-etapas que poderiam ser aplicadas para a correta descoberta de conhecimento, embora muitas vezes essas etapas são negligenciadas. Como exemplo, podemos mencionar que é comum ignorar a análise do mecanismo de ausência de dados durante o processo de imputação de dados ausentes. O processo de imputação pode trazer resultados desastrosos se o mecanismo for não ignorável (Rubin 2006). Este problema ainda pode piorar se o banco de dados possui massivos dados aussentes (Zárate et al., 2007). Outro problema não menos importante é a definição do tamanho da amostra de dados. Um tamanho inadequado da amostra pode impor restrições à representatividade do conhecimento extraído.

Na prática é possível observar que os responsáveis por projetos de DS se apresam para encontrar padrões e construir modelos por meio da aplicação das ferramentas disponíveis no mercado. Porém, um processo KDD executado cuidadosa e criteriosamente requer sempre um tempo maior. Experiências em projetos de KDD mostram que aproximadamente 75% dos custos do projeto (tempo e recurso financeiro) normalmente são gastos na etapa do pré-processamento (integração e seleção de dados, limpeza de dados, análise de outliers, recuperação de dados ausentes, etc) etapa essencial para a correta descoberta de conhecimento. Na prática tempo suficiente não é gasto na etapa do pré-processamento e por consequencia o conhecimento extraído pode não ser correto. É importante ressaltar que a condução de um projeto de KDD, levando em conta todas as ações de enriquecimento, melhoramento, detecção de

[1] http://www.kdnuggets.com/polls/

inconsistências, tratamento de dados ausentes, balanceamento, etc. pode demandar de um a dois anos de trabalho cuidadoso. Como mencionado, em Yang e Wu (2006), um dos principais desafios envolvendo a descoberta de conhecimento em banco de dados é encontrar uma teoria canônica que possa gerar uma metodologia unificada capaz de orientar processos de KDD de forma a minimizar a ocorrência de erros de projeto.

Os erros em projetos de DS podem ser evitados ou minimizados se o especialista de domínio pudesse participar durante o desenvolvimento do processo KDD. Além disso, em Freitas et al. (2005), os autores evidenciam a necessidade de uma documentação detalhada das etapas e tarefas aplicadas durante o processo KDD, uma vez que citam dificuldades enfrentadas como a falta de documentação, o retrabalho, a falta de medições dos esforços e de custos não esperados. Esses problemas poderiam ser minimizados caso houvesse uma documentação das etapas de forma que o processo pudesse ser mais bem gerenciado. Esses requisitos são suportados pelo método PICTOREA.

Como pode ser observado através da literatura, inumeros trabalhos e diversos métodos de aplicação do processo de KDD têm sido propostos. Muitos deles com alto nível de abstração que geralmente são focados sobre um domínio de problema específico. Segundo Boente et al. (2006) a complexidade inerente ao processo KDD decorre diretamente da enorme diversidade de alternativas de ações que podem ser executadas e das novas ações que surgem ao longo do processo.

Em projetos de mineração de dados do mundo real se observa uma grande variedade de cenários, fatores organizacionais, necessidades e preferências dos usuários. Entretanto, os atuais algoritmos de mineração de dados e ferramentas geralmente pecam ao entregar padrões que satisfazem apenas as expectativas técnicas. Geralmente, pessoas de negócios não têm interesse sobre como e o que se faz tecnicamente para obter os resultados. Há um sério conflito de interesses entre academia e indústria. Para reduzir este conflito, tornar os fatores do mundo real relevantes à mineração de dados, e torná-os mais úteis ao suporte de decisão no mundo real, é também um desafio que deve ser considerado.

Como mencionado, o método D3M propõe que cada etapa do processo KDD seja acompanhada e validada por um especialista de negócio. Com a colaboração entre especialistas em BI, especialistas em ciência dos dados e especialistas de domínio, o processo KDD se torna mais aderente e útil para tomada de decisão. Na proposta do método PICTOREA, durante o processo KDD, há pontos de avaliações constantes de forma que o especialista de domínio possa participar na avaliação das ações e proponha ajustes necessários à continuidade do processo.

[1] http://www.kdnuggets.com/polls/

É possível verificar que um processo KDD possui características semelhantes a um processo de negócios, principalmente no que concerne aos papéis dos envolvidos, necessidade de documentação, possibilidade de retornos e possibilidade de evolução. Entendemos que para a modelagem do PICTOREA é necessário modelar os processos do seu fluxo principal. Com esse intuito foram investigados na literatura aspectos fundamentais para se modelar esse fluxo de etapas. A seguir uma revisão dos principais aspectos considerados relevantes para construção do método PICTOREA são apresentados.

2.2 REVISÃO ACERCA DA MODELAGEM DE PROCESSOS

Segundo Graham (1999), um fluxo de tarefas além de conduzir os envolvidos num processo torna mais fácil o entendimento acerca do mesmo. Dessa forma, os benefícios de se ter um processo modelado vão desde a melhora da produtividade da equipe até à diminuição do esforço, tempo e custos.

Em Genvigir e Filho (2003), o autor salienta que os processos precisam ser modelados e padronizados de forma a entendermos o seu funcionamento. Isso possibilita o melhor treinamento e a proposta de melhorias. Em Humphrey (1989), o autor define os principais objetivos a serem alcançados ao modelar um processo: a) possibilitar a comunicação e o entendimento efetivo do processo; b) facilitar a reutilização; c) apoiar a evolução do processo; e d) facilitar o gerenciamento do mesmo. Ainda segundo Humphrey (1989), as principais razões que levam à padronização dos processos são: a) permitir o treinamento, gerenciamento e revisão das ferramentas de suporte; b) contribuir para a melhoria dos processos; e c) fornecer uma base estrutural para medição de tempos e custos. Devido ao crescimento da popularidade na modelagem de processos, cresceu também o número de técnicas e ferramentas para este fim. Hommes (2004) analisou as principais metodologias e ferramentas existentes e desenvolveu um método para avaliar a qualidade das metodologias para modelagem de processos.

Para construir o método PICTOREA foi necessário definir uma metodologia para construir o método. A seguir é brevemente apresentado e justificado o método científico utilizado para construir o método que sustenta PICTOREA.

De acordo com Hommes, não existe uma única metodologia para todas as pesquisas científicas. São as características individuais de cada projeto que determinam a metodologia mais adequada. Partindo deste princípio, neste trabalho é proposto o desenvolvimento de um método canônico, como uma estrutura principal detalhada, que sirva como suporte para projetos de DS, a partir do qual, seja possível definir as etapas específicas para um projeto KDD e definir, por tanto, um método específico para um problema específico.

[1] http://www.kdnuggets.com/polls/

Hommes também ressalta duas tradições na pesquisa moderna: o positivismo e o interpretativismo. Na tradição positivista, a realidade é objetivamente estudada e descrita de forma independente do pesquisador. Nesse caso, o papel da pesquisa científica consiste em sistematicamente adquirir conhecimento objetivo sobre o fenômeno a ser conhecido. Em sua tese, Hommes aponta problemas relacionados a essa abordagem. O autor argumenta que há sempre uma observação individual da realidade e assim não há garantia que a imagem percebida seja realmente a realidade, uma vez que até os filósofos esclarecem que temos limitações na percepção dessa realidade.

A tradição interpretativista tenta superar os problemas relativos à confiabilidade das percepções por ser mais moderada em suas reivindicações. A percepção e interpretação da possível realidade não podem ser separadas do pesquisador. O conhecimento depende das interpretações do fenômeno na realidade.

O método PICTOREA foi concebido levando em consideração a tradição interpretativista. Neste trabalho, o método proposto é baseado em conhecimentos tanto explicitos como tácitos, este último adquirido através de entrevistas com especialista para as definições dos principais requisitos, necessários para desenvolvimento de um projeto de ciência dos dados com qualidade. Espera-se que este método seja uma contribuição a caminho de um método unificado para descoberta de conhecimento em banco de dados.

3. METODOLOGIA PARA DESENVOLVIMENTO DO PICTOREA

Como já mencionado anteriormente, um método que estabeleça os procedimentos gerais e específicos a serem observados para o desenvolvimento de um processo KDD pode contribuir com os cientistas de dados que utilizam método próprio. O método PICTOREA permite um bom gerenciamento de todas as etapas de um processo KDD permitindo a um cientista de dados mais experiente, coordenar um ou mais projetos executados por profissionais menos experientes. A seguir, de forma resumida, apresentamos a metodologia proposta para construção do método PICTOREA.

Entendimento do objeto-problema - Como mostra a Fig. 1, para o entendimento acerca do objeto-problema (a proposta de um método canônico) foi utilizado o método científico interpretativista. Nesse sentido, temos duas fontes para o trabalho: o conhecimento tácito vindo de um especialista em projetos KDD e o conhecimento explícito, adquirido na literatura.

Para captura do conhecimento explícito foram definidas três grandes fontes: 1) Os conceitos em torno da Mineração de Dados Orientada ao Domínio – D3M;

[1] http://www.kdnuggets.com/polls/

2) aplicações de projetos KDD; e 3) o estudo de métodos para conduzir projetos KDD. Através do estudo destes três aspectos buscou-se identificar quais aspectos e características o método PICTOREA deveria incorporar.

Para captura do conhecimento tácito de um especialista em processos KDD houve uma série de interações com o especialista KDD com o objetivo de identificar os aspectos e etapas principais para a condução de um processo KDD. Em cada interação, feita por meio de entrevistas, as informações encontradas na literatura foram confrontadas com a experiência do especialista para a definição mais exata das etapas do novo método proposto. Como resultado, foi criado um método unindo o empírico ao teórico.

Figura 1- Entendimento do Objeto-Problema

Representação do Problema - O fluxo principal do método PICTOREA foi modelado utilizando a ferramenta para modelagem de processos em notação BPMN (BizAgi Process Modeler) (Bizagi, 2014). Como mencionado, a modelagem de um processo deve ser conduzida de modo a possibilitar o entendimento e a padronização do mesmo.

O PICTOREA foi modelado considerando 13 etapas no seu fluxo principal (Fig. 2) ao invés de 5 como proposta por Fayyad et al. (1996). Nas Figs. 3a e 3b são mostrados os diagramas de atividades e responsabilidades dos especialistas de domínio e de KDD respectivamente. A seguir cada etapa do método PICTOREA serão descritos.

[1] http://www.kdnuggets.com/polls/

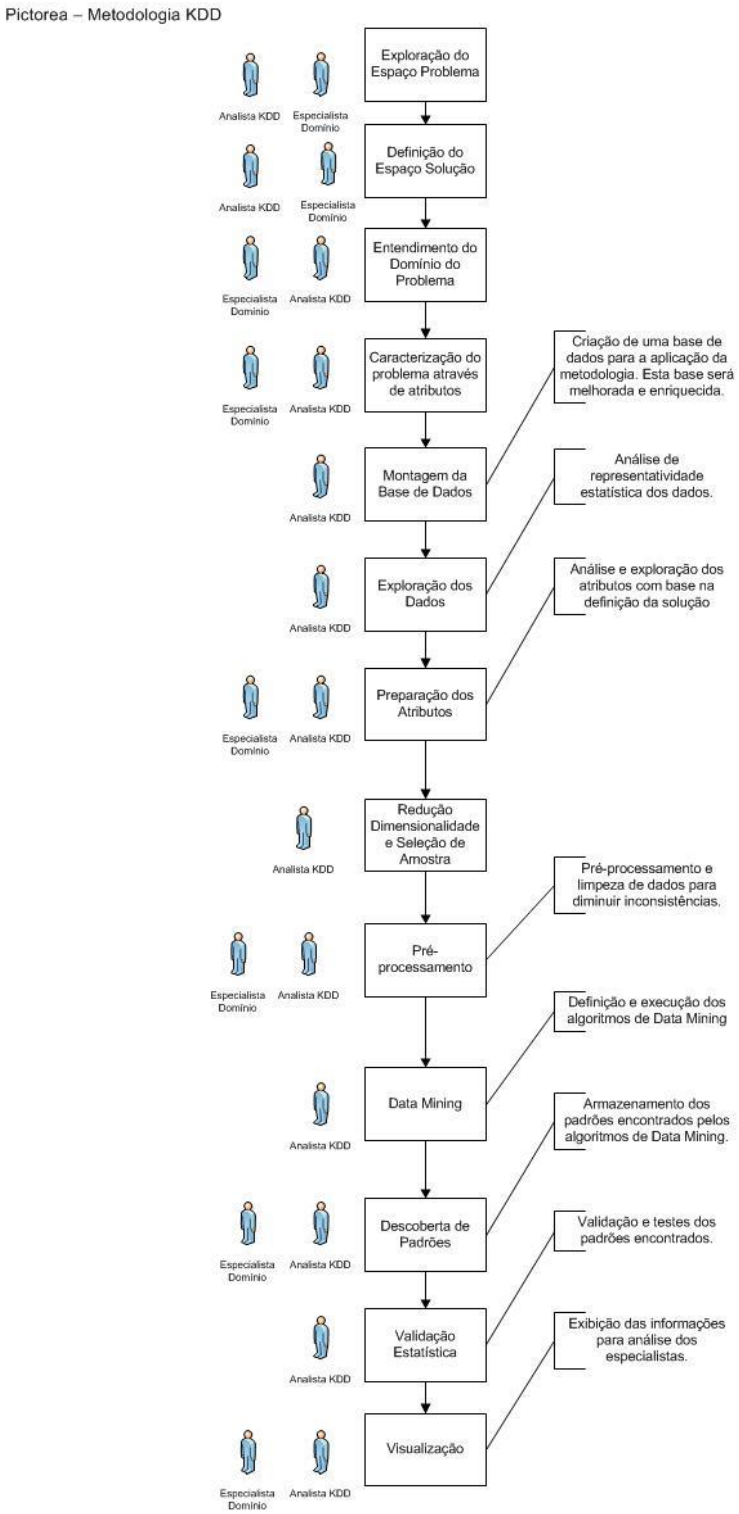

Figura 2- Fluxo principal do método PICTOREA

[1] http://www.kdnuggets.com/polls/

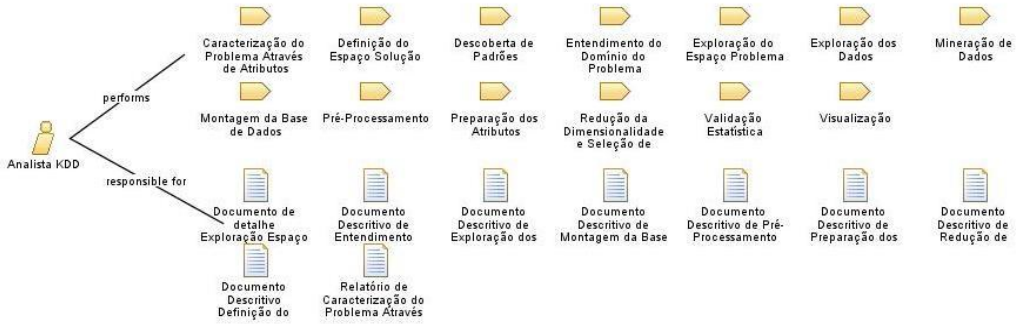

Figura 3a - Diagrama de atividades – especialista KDD

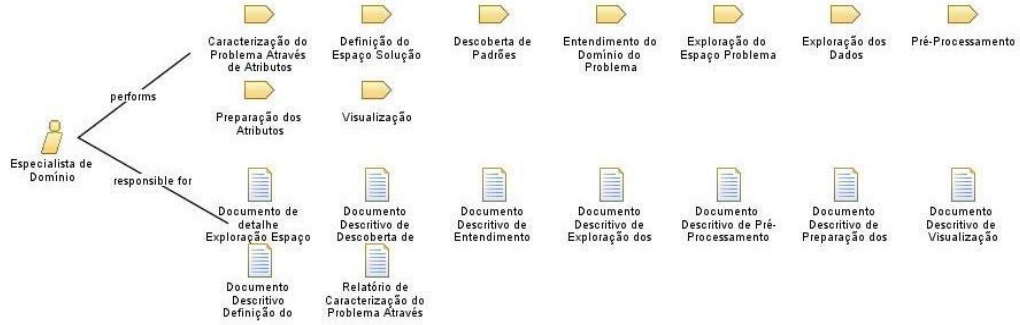

Figura 3b - Diagrama de atividades – especialista de domínio

1) Exploração do Espaço Problema – Como ponto de início, é necessário um conhecimento acerca dos problemas que determinam o espaço problemas do negócio ou empresa. O conhecimento acerca de cada domínio de problema, bem como as regras de negócio envolvidas é transmitido ao especialista KDD pelos especialistas de cada domínio.

A participação dos especialistas de domínio é imprescindível para a definição e listagem dos possíveis problemas, potenciais alvos para descoberta de conhecimento. Cada especialista de domínio contribui colocando peso aos possíveis problemas, visto que a resposta do processo a determinados problemas pode gerar mais valor para os negócios do que outros.

Segundo Pyle (1999), uma matriz de problemas *pairwise* pode ser utilizada para identificar e priorizar o domínio de problema a ser investigado. A Fig. 4, mostra o diagrama da etapa de Exploração do Espaço Problema e as responsabilidades dos responsáveis envolvidos.

[1] http://www.kdnuggets.com/polls/

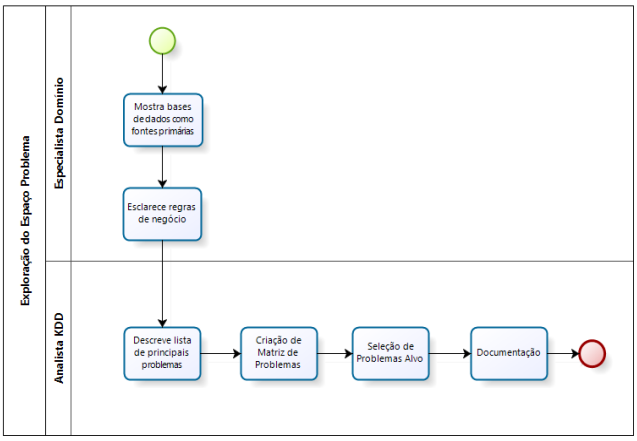

Figura 4 - Diagrama da etapa de exploração do espaço problema

2) Definição do Espaço Solução - Após a priorização e identificação do problema a ser investigado, o especialista KDD com o auxílio do especialista de domínio definirá as expectativas sobre o resultado e as saídas esperadas. Para isto, são definidas as técnicas de mineração de dados (Kaber & Han, 2001; Kantardzic, 2003; Wu et al., 2008) e de visualização (Keim, 2002; Fayyad, et al., 2002) a serem utilizados de forma a atender as expectativas do especialista de domínio. A Fig. 5 mostra o diagrama da etapa de Definição do Espaço Solução e as responsabilidades de ambos especialistas.

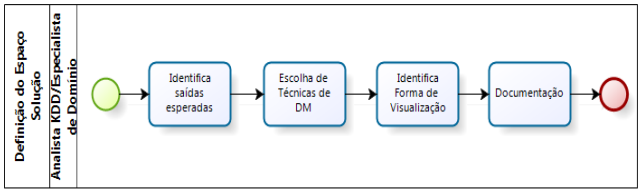

Figura 5 - Diagrama da etapa de Definição do Espaço Solução

3) Entendimento do Domínio de Problema - Nesta etapa o especialista KDD, sob a orientação do especialista de domínio, deverá comprender o domínio do problema em profundidade e caracterizá-lo utilizando modelos de ontologia ou mapas conceituais (Hong & Han, 2002). O objetivo desta etapa é identificar as características (atributos) que possam enriquecer o banco de dados tornando o conhecimento útil e não óbvio. A experiência mostra que o conhecimento não óbvio é resultado muitas vezes das características consideradas *julgamentos*. A descoberta de conhecimento em banco de dados baseado unicamente em informações de *fatos* pode não levar a conhecimento relevante. Por exemplo, consideremos a busca de padrões para os acidentes de trâncitos. Dar atenção somente aos fatos: <imprudência, efeito do álcool, velocidade excessiva e falha

[1] http://www.kdnuggets.com/polls/

mecânica> não traz conhecimento útil. Variáveis de *julgamento* como <perfil dos condutores, perfil dos acompanhantes, etc> pode enriquecer a base de dados trazendo conhecimento não óbvio. A Fig. 6 mostra o diagrama desta etapa.

Observe que a participação do especialista de domínio é relevante. Pois este auxilia na definição do problema, na exploração do espaço solução e no entendimento do domínio de problema escolhido.

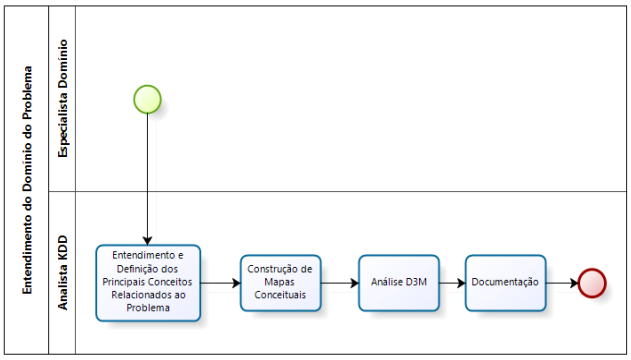

Figura 6 - Diagrama da etapa de entendimento do domínio do problema

4) Caracterização do Problema Através de Atributos - Após entendimento do domínio de problema, da definição das expectativas e dos resultados esperados, é necessário identificar os atributos relevantes para compor o banco de dados. Este procedimento é chamado de seleção conceitual de atributos. Cada atributo deve ser avaliado conceitualmente pelos especialistas KDD e de domínio selecionando o atributo de acordo com a sua relevância em relação ao problema tratado. Será documentado cada atributo, seu tipo e faixa de valores. Ver diagrama da Fig. 7 para observar as responsabilidades de ambos especialistas.

5) Montagem e verificação da consistência da Base de Dados - Com os atributos selecionados na etapa anterior será montado um arquivo de dados preliminar para o projeto. Alguns atributos poderam ser combinados com outros atributos com o objetivo de reduzir a dimensionalidade (Pyle, 1999; Zhan, 2003; Cooley et al., 1999).

É comum o especialista KDD ter a necessidade de buscar em fontes de dados externas dados para os atributos identificados como essenciais durante as etapas 2 e 3. Esses dados podem ser importados a partir de bancos de dados externos para enriquecimento do banco de dados principal. Além disso, pode ser também necessário diminuir a granularidade de algum atributo, o que é chamado de melhoramento de atributos. Deverão ser também verificadas a consistência e coerência dos valores dos atributos e das instâncias, a presença de poluição nos dados, a integridade e a duplicidade de instâncias. Ao final desta etapa, é feita

[1] http://www.kdnuggets.com/polls/

uma avaliação da representatividade do banco de dados criada a partir da análise dos domínios dos atributos (entende-se por representatividade conter dados suficientes para descrever o domínio de problema). Caso o banco de dados resultante não seja representativo o suficiente para a descoberta de conhecimento, o especialista KDD pode decidir por prosseguir, pode-se voltar a alguma etapa anterior do método ou impor restrições ao conhecimento a ser extraído. Caso o especialista KDD opte por não prosseguir, os motivos são documentados e o processo de descoberta de conhecimento é cancelado. Ver diagrama da Fig. 8 para observar os pontos de bifurcação e tomada de decisão.

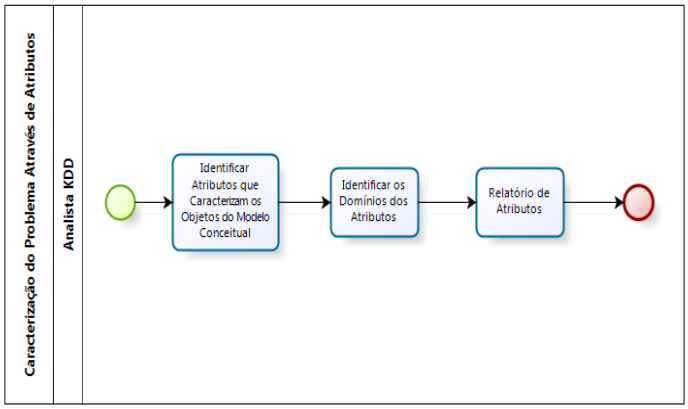

Figura 7 - Diagrama da etapa de caracterização do problema através de atributos

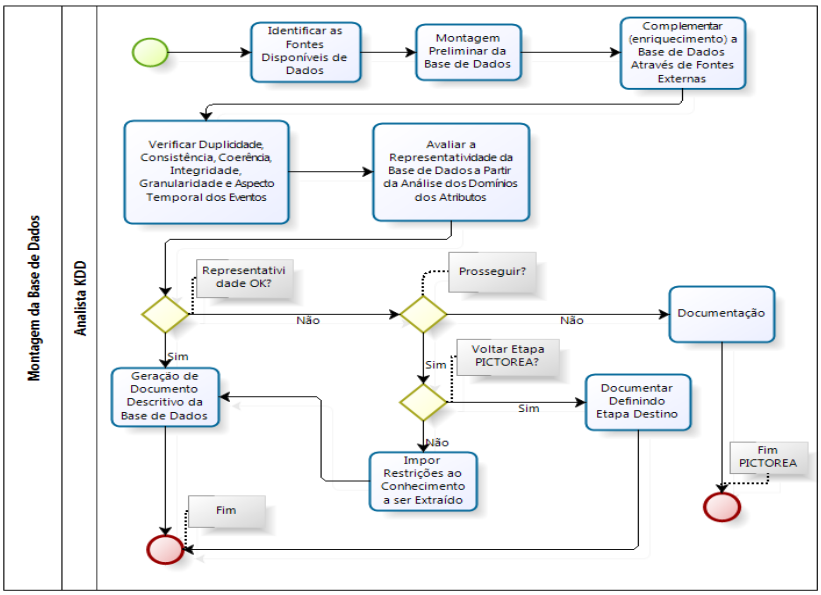

Figura 8 - Diagrama da etapa de Montagem e avaliação de consistência da base de dados

6) Exploração dos Dados - Nesta etapa será feita a caracterização dos atributos identificando o tipo de dado, intervalo de valores, frequência de dados ausentes, valores mínimos, valores máximos, histogramas, função de densidade de probabilidade dos atributos, etc. O especialista KDD precisa conhecer e explorar os dados que serão utilizados no processo de descoberta de conhecimento (Pyle, 1999).

Será feita uma análise estatística descritiva dos atributos com o intuito de identificar médias, medianas, desvio padrão, intervalos de confiança, entalpia, etc. Estes procedimentos são úteis para análise de outliers, tratamento de dados aussentes e discretização de atributos. Ao final da etapa, tem-se um documento descritivo da exploração dos dados. Ver Diagrama da Fig. 9.

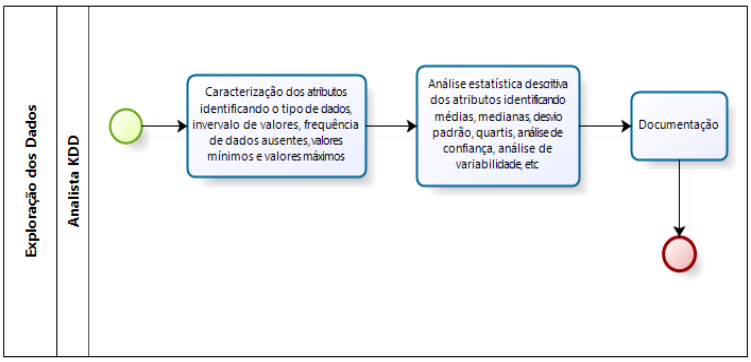

Figura 9 - Diagrama da etapa de exploração dos dados

7) Análise dos dados - Nesta etapa serão realizadas as análises de *Outliers* e de dados aussentes. *Outliers* são dados ou registros com comportamento muito diferente dos demais que fogem ao padrão dos dados (Hodge & Austin, 2004). Estes dados precisam ser identificados e analisados, pois possivelmente trata-se de erros no banco de dados. Deve ser feita também uma análise de dados ausentes com o intuito de verificar o impacto que esta ocorrência terá na descoberta de conhecimento. Decisões devem ser tomadas em relação a eliminar registros contendo dados aussentes, imputar valor ou selecionar técnicas de mineração de dados que lidam com instâncias contendo dados aussentes (Silva & Zárate, 2014). Cabe ao especialista KDD definir uma estratégia para tratar *outliers* e os dados ausentes. Ainda nesta etapa, será feita uma análise de representatividade dos dados a partir da análise dos domínios dos atributos observados na etapa 3. Se a base de dados não for representativa, o especialista KDD poderá decidir entre voltar em qualquer etapa do processo ou impor restrições ao conhecimento a ser extraído ou inclusive encerrar o projeto. Ver diagrama da Fig. 10 para observar pontos de tomada de decisão.

[1] http://www.kdnuggets.com/polls/

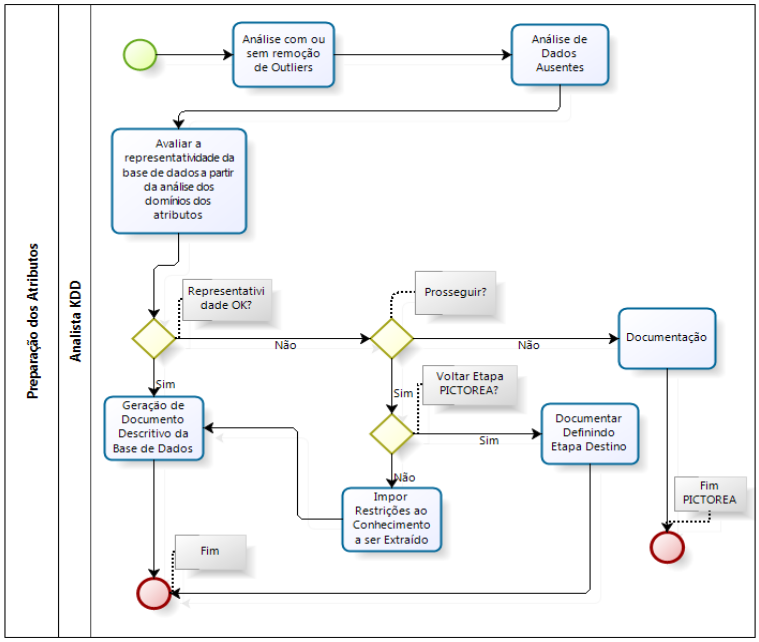

Figura 10 - Diagrama da etapa de análise dos dados

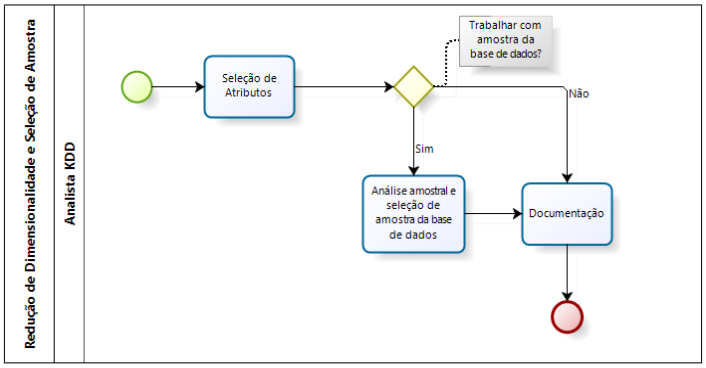

Figura 11 - Diagrama da etapa de redução de dimensionalidade e seleção de amostra

8) Redução da Dimensionalidade e Seleção de Amostra - Nesta etapa o analista KDD poderá aplicar técnicas de seleção de atributos (filtros, wrappers), redução de dimensionalidade como análise de componentes principais, análise de correlação entre atributos, etc. (Guyon & Elisseeff, 2003; Hall & Holmes, 2003). Nesta etapa, o especialista KDD deverá decidir se trabalhará com uma amostra ou com a totalidade de instâncias do banco de dados. Caso faça opção de utiliar uma amostra deve ser feita uma análise para seleção de amostra (Liu et al., 2012). Em relação à seleção de amostras deve-se ter sempre presente que o tamanho ideal de

[1] http://www.kdnuggets.com/polls/

um banco de dados para um processo KDD corresponde a: $Num\_Instâncias = n.m.p....$ sendo *n*, *m*, *p*,... o número de valores distintos de cada atributo. Ao final da etapa as análises e decisões tomadas serão documentadas. Ver diagrama da Fig. 11.

9) Transformação de dados - Os dados precisam ser transformados de forma que possam servir de entrada para os algoritmos de mineração de dados. Normalmente se faz necessária a transformação, discretização, mudanças de escala, ou normalização dos dados garantindo que se preservem as características dos valores originais García, et al., 2013). Segundo Pyle (1999), a melhor forma de transformar os dados é verificar quais requisitos a solução precisa atender e quais são os requisitos que a técnica de mineração de dados impõe. Ao final da etapa, tem-se um documento descritivo dos procedimentos adotados. Ver diagrama da Fig. 12.

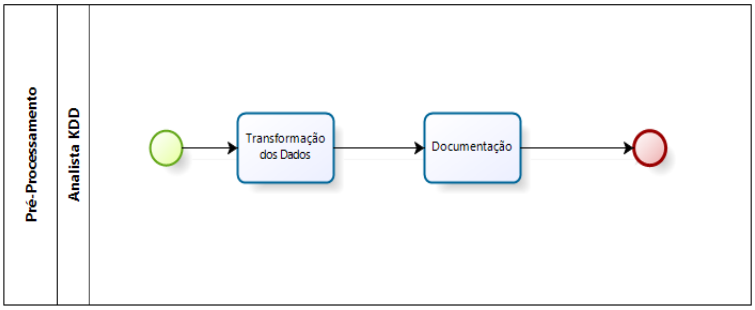

Figura 12 - Diagrama da etapa de transformação de dados

10) Mineração de Dados - O analista KDD deverá definir o conjunto de dados para treinamento e validação. Deverá identificar a técnica apropriada para a extração de conhecimentos que atendam as expectativas documentadas. Dentre as técnicas de aprendizado de máquina mais utilizadas podemos citar agrupamento, classificação e associação dentre as mais convencionais. A escolha da técnica de mineração de dados depende da análise das necessidades do especialista de domínio juntamente com as características da técnica (Britos, et al., 2006; Goebel & Gruenwald, 1999; Wu et al., 2008).

Como mencionado anteriormente, as ferramentas para mineração de dados existentes no mercado apresentam variações dos algoritmos clássicos que executam as técnicas de mineração de dados. Por esse motivo, não será discutido nenhum software específico para execução de uma técnica de mineração de dados. Uma vez definida a técnica, esta poderá ser aplicada sobre os dados de uma amostra selecionada para posteriormente ser aplicada ao banco de dados completo. Ver diagrama da Fig. 13.

[1] http://www.kdnuggets.com/polls/

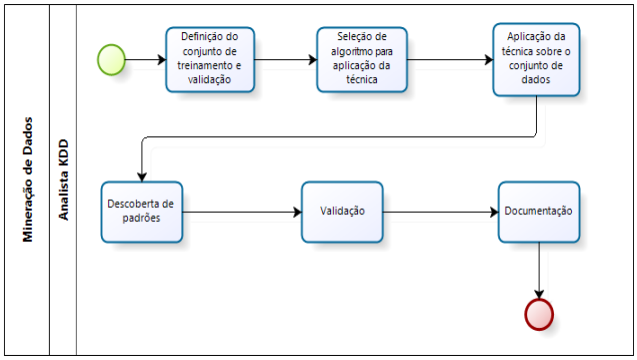

Figura 13 - Diagrama da etapa de mineração de dados

11) Descoberta de Padrões - Após a aplicação dos algoritmos de mineração de dados, os padrões encontrados deverão ser identificados, analisados e armazenados para posterior análise da qualidade do conhecimento extraído. Ver Fig. 14.

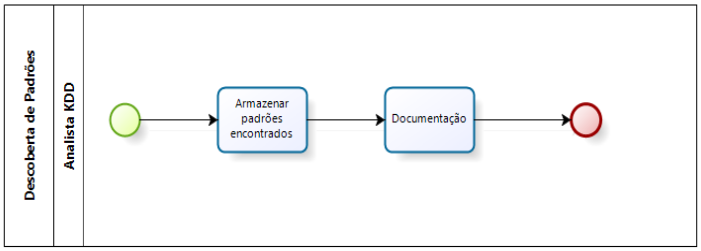

Figura 14 - Diagrama da etapa de descoberta de padrões

12) Validação Estatística - Os padrões descobertos deverão ser verificados estatisticamente. Para isso, pode ser utilizada a técnica *Cross-Validation*, a qual é um método estatístico geralmente utilizado para validação e comparação do resultado dos algoritmos de aprendizado de máquina. Basicamente os dados são divididos em dois grupos: um deles utilizado para o treinamento e outro para validação e suas combinações. Este procedimento possui fundamentação estatística e permite a construção de intervalos de confiança.

Ainda em relação a *Cross-Validation* encontra-se na literatura várias formas de aplicação tais como: *Resubstitution Validation, Hold-Out Validation, K-Fold Cross-Validation, Leave-One-Out Cross-Validation e Repeated K-Fold Cross-Validation* (Payam, et al. 2009). Ver diagrama da Fig. 15 para as responsabilidades dos especialistas.

É importante notar que as etapas 11, 12 e 13 podem ser executadas em ciclos evolutivos até chegar a conhecimento representativo, válido e útil.

[1] http://www.kdnuggets.com/polls/

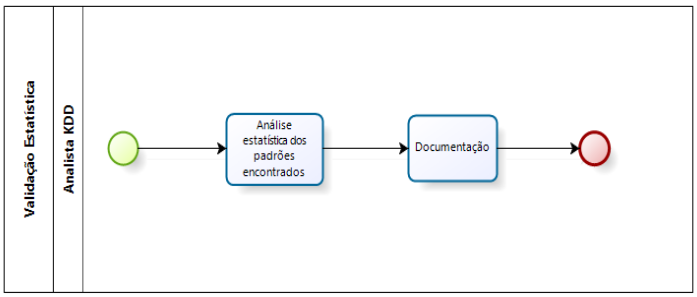

Figura 15 - Diagrama da etapa de validação estatística

13) Visualização - De acordo com a técnica de mineração de dados utilizada, o especialista KDD deverá selecionar a técnica de visualização mais apropriada para a exibição do padrão encontrado (Keim, 2002; Fayyad et al., 2002). O especialista de domínio fará a validação dos dados decidindo se o padrão encontrado corresponde às expectativas ou responde ao problema definido. Caso nesta etapa o especialista de domínio não encontre relevância nos padrões encontrados, o especialista KDD deverá verificar os resultados de cada etapa e poderá refazer o processo a partir da etapa que decidir.

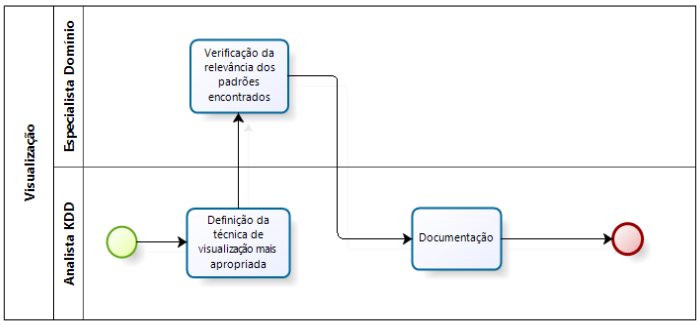

Figura 16 - Diagrama da etapa de visualização

4. Aplicação do método PICTOREA

O caráter pedagógico do método PICTOREA foi avaliado na disciplina de Tópicos Especiais de um curso de graduação em Tecnologia da Informação de uma IES durante o 2º semestre de 2015. A turma considerada cursava o 3º ano e estava formada por 30 alunos, dos quais 22 responderam às seguintes questões:

Q1: Conhecimento prévio acerva de Mineração de Dados e KDD.
Q2: Importância dada à etapa do pré-processamento.
Q3: Entendimento acerca do domínio de problema.
Q4: Tempo necessário gasto durante a etapa do pré-processamento.
Q5: Importância de um método padrão para guiar um processo KDD.

[1] http://www.kdnuggets.com/polls/

O perfil dos entrevistados é mostrado na Figura 17f. A totalidade desses alunos atua na área de tecnologia da informação embora menos de 13,64% atua como Analista de Dados e Inteligência de Negôcios. A turma foi dividida em 6 grupos para desenvolver projetos de KDD seguindo a metodologia PICTOREA. Ao final do semestre a turma foi convidada a responder aos quesitos anteriores.

Em relação ao quesito Q1, Fig. 17a, pode ser observado claramente que após a execução dos projetos, o conhecimento acerca da mineração de dados e do processo KDD aumentou consideravelmente de 0% para 68,18% nos itens "Bom" e "Muito bom". Isto mostra que o método PICTOREA possui um caráter pedagógico, permitindo a condução de projetos KDD por parte de profissionais menos experientes. Em relação ao quesito Q3, Fig. 17c, pode ser observado que a importância dada ao entendimento do domínio de problema subiu de 30,91% para 90,91% para os itens "Altamente relevante" e "Essencial". Com este resultado é possivel observar que se não for dada atenção debida ao conhecimento prévio acerca do domínio de problema um projeto KDD pode estar destinado ao insucesso.

Em relação ao quesito Q2, Fig. 17b, relativo à importância dada à etapa do pré-processamento subiu de 18,19% para 95,55% para os itens "Altamente relevante" e "Essencial". É observado na prática que muitos responsáveis de BI se apressam para aplicar ferramentas muito antes de gastar tempo na preparação adequada da base de dados. Em relação ao quesito Q4, Fig. 17d, relativo ao tempo necessário a ser gasto na etapa do pré-processamento 77,27% dos entrevistados consideram um tempo de 50 a 75% do tempo total gasto no projeto. Em relação à necessidade de um método padrão que sirva para conduzir projetos de KDD subiu de 31,82% para 63,63% para a faixa de 75 a 100% que consideram necessário um método padrão.

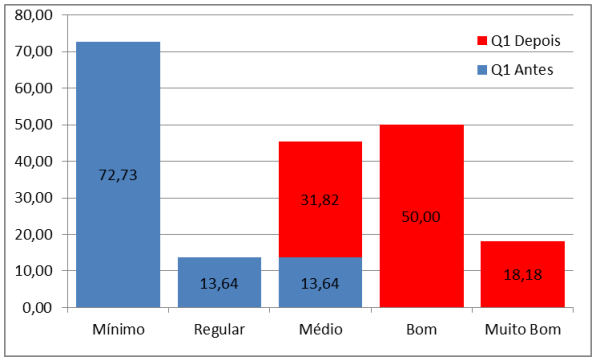

Figura 17a - Resultados de pesquisa referente a Q1: Conhecimento prévio acerca de Mineração de Dados e KDD

[1] http://www.kdnuggets.com/polls/

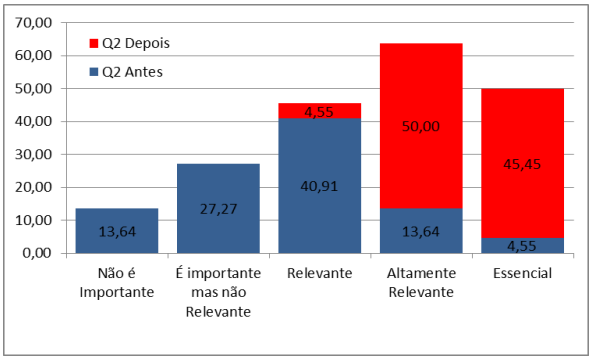

Figura 17b - Resultados de pesquisa referente a Q2: Importância dada à etapa do pré-processamento

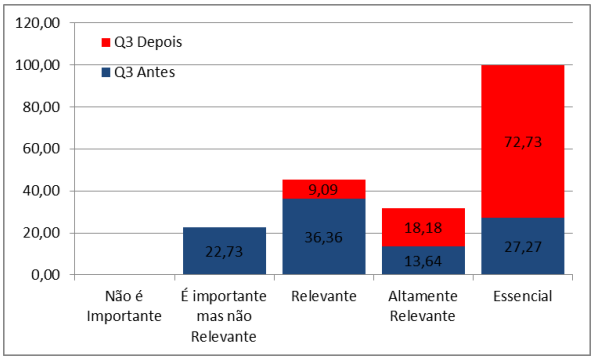

Figura 17c - Resultados de pesquisa referente a Q3: Entendimento acerca do domínio do problema.

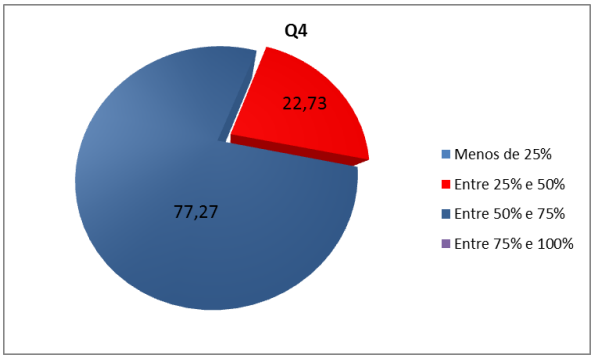

Figura 17d - Resultados de pesquisa referente a Q4: Tempo necessário a ser gasto durante a etapa do pré-processamento

[1] http://www.kdnuggets.com/polls/

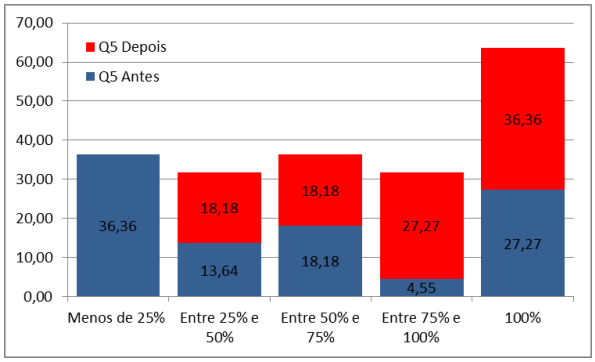

Figura 17e - Resultados de pesquisa referente a Q5: Importância de um Método padrão para guiar um processo de KDD

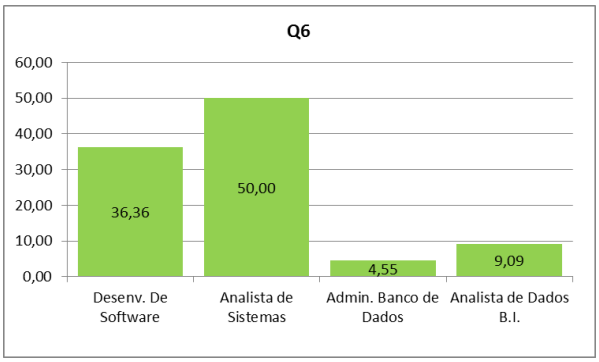

Figura 17f - Resultados de pesquisa referente a Q6: Perfil de atuação do respondente

5. CONCLUSÕES

Neste trabalho foi proposto um método denominado PICTOREA para conduzir a execução de processos de descoberta de conhecimento em bancos de dados em projetos de Ciência dos Dados. Foi reforçado o desafio e a necessidade em encontrar um método canônico que possa garantir a padronização, a qualidade, o retorno, e a diminuição dos custos.

Diante da falta de um padrão e da necessidade de métodos canônicos, utilizamos do conhecimento especialista (conhecimento tácito) por meio da metodologia interpretativista aliado à revisão da literatura (conhecimento explícito) para fundamentar as etapas do método PICTOREA. Essas etapas são necessárias para correta condução e aplicação de um projeto KDD. Entendemos que não há uma única metodologia para todos os projetos KDD, pois cada projeto possue características individuais que determinam a metodologia mais adequada. Porém, a nossa proposta de um método canônico, a partir do qual é possível

[1] http://www.kdnuggets.com/polls/

definir um método específico para um projeto KDD, tem por objetivo guiar os responsáveis por projetos de KDD a opterem projetos mais aderentes aos objetivos do negócio.

Um aspecto relevante do método PICTOREA, é que este pode ser conduzido por profissionais menos experientes sob a supervisão de um especialista, além de ser um facilitador na aprendizagem dos aspectos relevantes de projetos de ciência dos dados, vale destacar o caráter pedagógico para apoio à aprendizagem. O PICTOREA favorece a padronização com intuito de garantir maior qualidade, controle e o desenvolvimento de futuras métricas de avaliação de projetos KDD. Além disso, observa-se o potencial do método para trabalhos em grupo aumentando a produtividade dos setores de *Business Intelligence* e *Data Science* das empresas.

AGRADECIMENTOS

Os autores agradecem ao apóio financeiro recebido da Fundação de Apóio à Pesquisa do Estado de Minas Gerais, FAPEMIG, do Conselho Nacional para o Desenvolvimento Científico e Tecnológico, CNPq, Brasil.

REFERÊNCIAS

Boente, A.N.P., Goldschmidt, R.R., & Estrela, V.V. (2006). Uma metodologia para apoio e realização do processo de descoberta de conhecimento em bases de dados. In *Anai III Simpósio de Excelência em Gestão e Tecnologia*, 1-13.
http://www.boente.eti.br/publica/artigo2wcc.pdf

Britos, P., Merlino, H., Fernaández, E., Ochoa, M.A., Diez, E., & García-Martínez, R. (2006). Tool selection methodology in data mining. In: *V Jornadas Iberoamericanas de Ingeniería de Software e Ingeniería del Conocimiento, Memoria Técnica, JIISIC*. Puebla, Pue. México, 85-90.
http://iidia.com.ar/rgm/comunicaciones/JIISIC-2006-Tool-Selection-Methodology-in-Data-Mining.pdf

Cao, L. (2007). Domain-Driven Actionable Knowledge Discovery, *IEEE Intelligent Systems*, vol. 22, no. 4, 78-89.
DOI: 10.1109/MIS.2007.67

Cao, L., & Zhang, C. (2005). Domain-Driven Data Mining: A Practical Methodology, *Int',l J. Data Warehousing and Mining*, vol. 2, no. 4, 49-65.
http://www.igi-global.com/article/domain-driven-data-mining/1774

[1] http://www.kdnuggets.com/polls/

Cao, L. (2010). Domain Driven Data Mining: Challenges and Prospects. In: *IEEE Transactions on Knowledge and Data Engineering*, *22* (6), 755-769.
http://www.computer.org/csdl/trans/tk/2010/06/ttk2010060755-abs.html

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communication ACM*, New York, NY, USA, 39(11), 27-34.
DOI: 10.1145/240455.240464

Freitas, P.B.A., Brazdil, P., Pereira, A.C. (2005). Mining hospital databases for management support. In: *IADIS Virtual Multi Conference on Computer Science and Information Systems.* 207-212.
https://www.researchgate.net/profile/Pavel_Brazdil/publication/266875100_MINING_HOSPITAL_DATABASES_FOR_MANAGEMENT_SUPPORT/links/546378cf0cf2c0c6aec4c052.pdf

Goebel, M., & Gruenwald, L. (1999). A survey of data mining and knowledge discovery software tools. In: *SIGKDD Explorations ACM* SIGKDD. 1, Issue 1, June, 20-33.
DOI: 10.1145/846170.846172

Graham, BB. (2004). Business Process Improvement Methodology. Ohio, USA: *The Ben Graham Corporation*.

Hommes, L.J. (2004). The Evaluation of Business Process Modeling Techniques. Tese (Doutorado) - Delft University of Technology, 2004.
http://repository.tudelft.nl/view/ir/uuid%3A1d209c45-4b2a-41f2-9e94-a54b8ee76d78/

Pyle, D. (1999). *Data preparation for data mining*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Yang, Q., & WU, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, 5(4), 597-604, 2006.
http://www.cs.uvm.edu/~icdm/10Problems/10Problems-06.pdf

Genvigir, E.C., & Filho, L.F. (2003). Modelagem de processos de software através do SPEM - software process engineering meta model - conceitos e aplicação. In: *WORKSHOP DOS CURSOS DE COMPUTAÇÃO APLICADA DO INPE*, 3.

[1] http://www.kdnuggets.com/polls/

http://mtc-m16c.sid.inpe.br/col/lac.inpe.br/worcap/2003/10.31.14.46/doc/artigo_worcap_elias_2003.PDF

Humphrey, W.S. (1989). *Managing the software process*. Boston: Addison-Wesley.

Bizagi. (2011). Bizagi Process Modeler. Disponivel em: < http://www.bizagi.com/products/bizagi-bpm-suite/download/ >. Acesso em: 08 fev. 2014.

Rubin, D.B. (2006). Conceptual, computational and inferential benefits of the missing data perspective in applied and theoretical statistical problem. *Statistisches Archiv*, 90(4), 501-513.
http://link.springer.com/article/10.1007/s10182-006-0004-z

Zárate, L.E., Nogueira, B.M., & Santos, T.R.A. (2007). Comparison of Classifiers Efficiency on Missing Values Recovering: Application in a Marketing Database with Massive Missing Data. In *Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, IEEE CIDM*, 66-72.
DOI: 10.1109/CIDM.2007.368854

Dhar, V. (2013). Data science and prediction. *Communications of the ACM* 56 (12): 64.
DOI:10.1145/2500499.

García, S., Luengo J., Sáez, J.A., López, V., & Herrera, F. (2013). Survey of Discretization Techniques: Taxonomy and Empirical Analysis in Supervised Learning. *IEEE Tranc. on Knowledge and data Engineering*, 25, 4, 734-750.
DOI: 10.1109/TKDE.2012.35

Fayyad, U., & Stolorz, P. (1997). Data Mining and KDD: promise and challenges. *Future Generation Computer Systems*, 13, 99-115, Elsevier Science.
DOI: Data Mining and KDD: promise and challenges

Kamber, H., & Han, J. (2001). *Data Mining: Concepts and Techniques*, Morgan Kaufmam Pub.

Kantardzic, M. (2003). *Data Mining: Concept, Models and Algorithms*, IEEE Press.

[1] http://www.kdnuggets.com/polls/

Silva, L.O., & Zárate, L.E. (2014). A brief review of the main approaches for treatment of missing data. *Intelligent Data Analysis*. 18, 6, 1177-1198.
DOI: 10.3233/IDA-140690

Keim, D.A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*. 1-8, Jan/Mar.
DOI: 10.1109/2945.981847

Fayyad, U.M., Wierse, A, & Grinstein, G.G. (2002). *Information visualization in data mining and knowledge discovery*. Academic Press, USA.

Hong, T., & Han, I. (2002). Knowledge-based data mining of news information on internet using cognitive maps and neural networks. *Expert Systems with Applications*, 23, 1-8.
DOI: 10.1016/S0957-4174(02)00022-2

Zhang, S., Zhang, C., & Yan, Q. (2003). Data preparation for data mining. *Applied Artificial Intelligence*, 17, Iss. 5-6, 375-381.
DOI: 10.1080/713827180

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. *Knowledge and Information Systems*. Feb. 1, Issue 1, 5-32.
DOI: 10.1007/BF03325089

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, Ph.S., Zhou, Zhi-Hua, Steinbach, M., Hand, D.J., & Steinberg, D. (2008). Top 10 algorithms in data mining. *Knowl Inf Syst* 14,1–37.
DOI: 10.1007/s10115-007-0114-2

Hodge, V.J., Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22 (2), 85-126.
http://eprints.whiterose.ac.uk/767/1/hodgevj4.pdf

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3, 1157-1182.
http://www.jmlr.org/papers/volume3/guyon03a/guyon03a.pdf

Hall, M.A., & Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on knowledge and data Engineering*, 15, 6, 1437-1447.

[1] http://www.kdnuggets.com/polls/

DOI: 10.1109/TKDE.2003.1245283

Liu, Ch.F., Yeh, Ch.Y., & Lee, Sh.J. (2012). A Novel Prototype reduction approach for supervised learning. *International Journal of Innovative Computing, Information and Control,* 8, 6, 3963-3980. http://itlab.ee.nsysu.edu.tw/paper/A%20NOVEL%20PROTOTYPE%20REDUCTION%20APPROACH%20FOR%20SUPERVISED.pdf

Payam, R., Tang, L., & Liu, H. (2009). Arizona State University. Cross-Validation, In *Encyclopedia of Database Systems (EDBS),* Editors: Ling Liu and M. Tamer Özsu. Springer.

[1] http://www.kdnuggets.com/polls/