PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS
Programa de Pós-Graduação em Informática

André Luiz Dias Montevecchi

PICTOREA: UM MÉTODO PARA DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS CONVENCIONAIS

Belo Horizonte
2012

André Luiz Dias Montevecchi

PICTOREA: UM MÉTODO PARA DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS CONVENCIONAIS

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Orientador: Luis Enrique Zárate Gálvez

Belo Horizonte

2012

FICHA CATALOGRÁFICA
Elaborada pela Biblioteca da Pontifícia Universidade Católica de Minas Gerais

Montevecchi, André Luiz Dias
M781p Pictorea: um método para descoberta de conhecimento em bancos de dados convencionais / André Luiz Dias Montevecchi. Belo Horizonte, 2012.
98f.: il.

Orientador: Luis Enrique Zárate Gálvez
Dissertação (Mestrado) – Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática.

1. Banco de dados. 2. Mineração de dados (Computação). I. Gálvez, Luis Enrique Zárate. II. Pontifícia Universidade Católica de Minas Gerais. Programa de Pós-Graduação em Informática. III. Título.

CDU: 681.3.011

André Luiz Dias Montevecchi

PICTOREA: UM MÉTODO PARA DESCOBERTA DE CONHECIMENTO EM BANCOS DE DADOS CONVENCIONAIS

Dissertação apresentada ao Programa de Pós-Graduação em Informática da Pontifícia Universidade Católica de Minas Gerais, como requisito parcial para obtenção do título de Mestre em Informática.

Prof. Luis Enrique Zárate (Orientador) – PUC Minas

Prof. Humberto Torres Marques Neto – PUC Minas

Prof. Clodoveu Augusto Davis Junior - UFMG

Belo Horizonte, 05 de março de 2012

AGRADECIMENTOS

Agradeço ao meu orientador, Professor Zárate, pela ajuda, paciência e dedicação.

Agradeço aos meus pais pelo apoio e paciência durante todo este trabalho.

Agradeço a minha querida esposa, Eyre Montevecchi, pelo carinho e paciência.

Agradeço a CAPES pelo financiamento da pesquisa e pela oportunidade.

Agradeço aos meus grandes amigos e colegas de mestrado, Henrique Batista e Adriana Monteiro, pelo companheirismo nos estudos.

RESUMO

Atualmente, a descoberta de conhecimento em banco de dados (*Knowledge Discovery in Databases* - KDD) é bastante utilizada na academia e pouco utilizada no mercado. As organizações que utilizam o processo KDD, geralmente o fazem adquirindo softwares com metodologias definidas como o CRISP-DM. Porém, grande parte das aplicações de KDD é feita utilizando a metodologia própria do responsável pelo processo, as quais geralmente não seguem um padrão. Através do método científico interpretativista, dos conceitos de *Domain-Driven Data Mining* - D3M, com auxílio da Metodologia para Modelagem de Processos (*Business Process Management* - BPM) e do SPEM - *Softwares and Systems Process Engineering Meta-Model*, este trabalho propõe um novo método, com caráter pedagógico, denominado PICTOREA, para desenvolvimento, acompanhamento e documentação das etapas e atividades de um projeto KDD.

Palavras-chave: Processo KDD, Mineração de Dados, Descoberta de Conhecimento em Bancos de Dados.

ABSTRACT

Currently, knowledge discovery in databases – KDD is widely used in academia and little used in the market. Organizations using the KDD process, usually do so by purchasing software with established methodologies such as CRISP-DM. However, most KDD applications are made using the own methodology of the responsible for the process. These methods are not standardized. Through the interpretative scientific method, the concepts of Domain-Driven Data Mining - D3M, with Business Process Management - BPM and SPEM - Software and Systems Process Engineering Meta-Model, this study proposes a new pedagogical method called PICTOREA, for developing, monitoring and documenting the steps and activities of a KDD project.

Keywords: Data mining, KDD, Domain-Driven Data Mining.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

1.1 Considerações iniciais e caracterização do problema

Segundo Brusse e Wenning (2005), padronização (*standardization*) (ou norma) é o processo de desenvolver e combinar normas técnicas. Uma norma (padrão) é um documento que permite estabelecer engenharia uniforme ou especificações técnicas, critérios, métodos, processos, ou práticas aplicáveis em fluxos de processos de informação.

Em diferentes áreas, há sempre uma grande preocupação e busca por estabelecimento de padrões. Especificamente na área de TI, dentre os principais benefícios da padronização se encontram compatibilidade e interoperabilidade, garantia de qualidade e agilidade de comunicação. A padronização pode contribuir para remover barreiras técnicas, abrir novos mercados e desenvolver novos modelos de negócios.

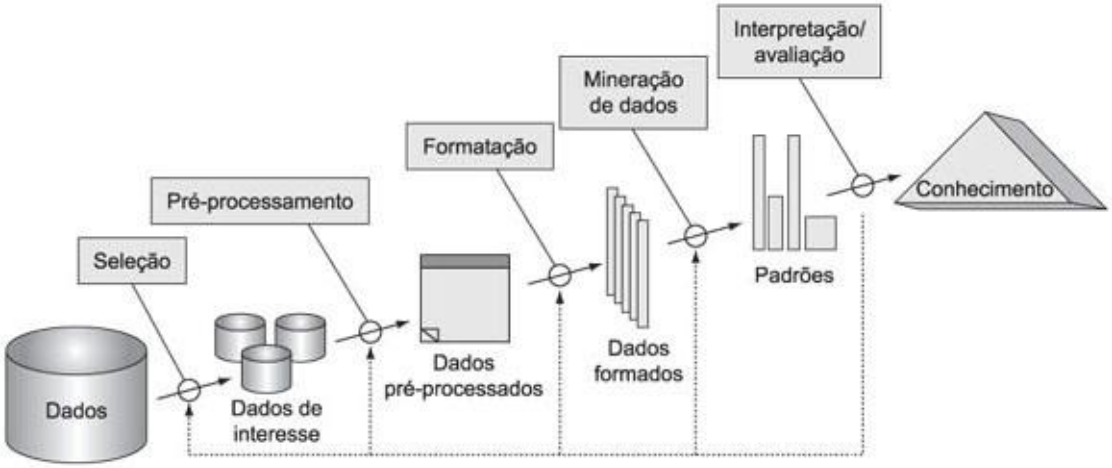
É sabido que diariamente as organizações acumulam grande volume de dados provenientes de suas atividades operacionais. Esses dados possuem um tipo de informação vital para a organização, pois contêm conhecimento sobre processos internos, perfis de funcionários, clientes e operações em geral. O volume de informações é geralmente tão grande que inviabiliza a análise e assimilação por parte do ser humano. Nesse cenário surgiu o Processo de Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases* - KDD), o qual procura por padrões desconhecidos e informações ocultas nos bancos de dados, gerando assim, novo conhecimento para tomada de decisão. O processo KDD, tal qual é hoje, possui cinco etapas básicas: seleção, pré-processamento, transformação, mineração de dados e interpretação/avaliação, conforme demonstrado na Figura 1. (FAYYAD; PIATETSKY-SHAPIRO; PADHRAIC, 1996).

É conhecido que a falta de padronização de um processo KDD leva também à falta de uma documentação das etapas e decisões tomadas durante esse processo. Segundo documento que relata os 10 principais desafios em KDD (YANG; WU, 2006), um deles aponta que a maioria das técnicas de processos KDD são desenvolvidas para problemas individuais. Com isso, não há ainda uma teoria unificada acerca do processo KDD. Segundo os autores, há várias pesquisas

também no sentido de incorporar metodologias de extração de conhecimento em ferramentas de mineração de dados.

Em Yang e Wu (2006), é exposto que o software *Clementine*, talvez a ferramenta mais popular utilizada em processos de KDD, possui uma boa interface com o usuário, mas não possui uma teoria que sustente suas operações. Segundo os autores, encontrar uma teoria canônica sobre KDD é um desafio futuro para os pesquisadores da área.

Figura 1 - Etapas do Processo KDD



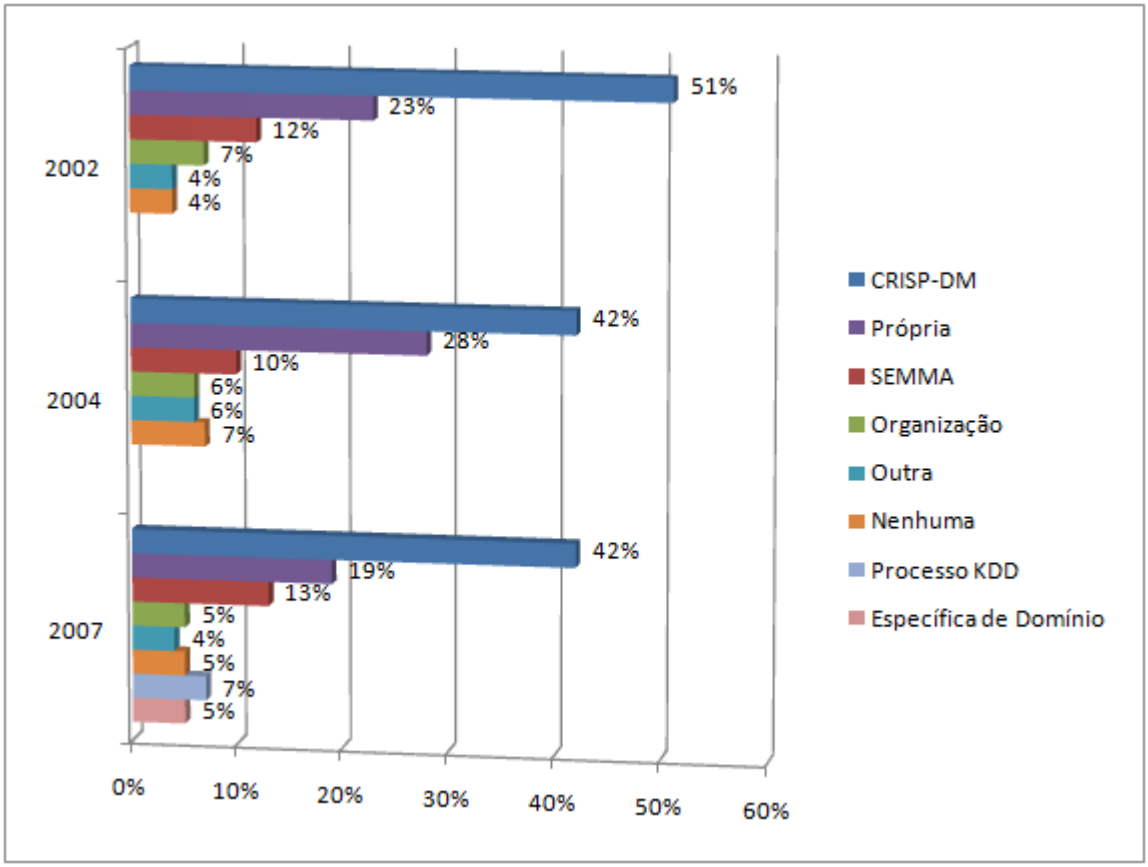
Fonte: FAYYAD; PIATETSKY-SHAPIRO; PADHRAIC, 1996

Atualmente, há duas principais metodologias para aplicação de processos de KDD voltadas para o mercado. Ambas aplicam as etapas de KDD propostas por Fayyad, Piatetsky-Shapiro e Smyth (1996). São elas a CRISP-DM (*Cross Industry Standard Process for Data Mining*), desenvolvida originalmente pela SPSS e recentemente adquirida pela IBM e a SEMMA desenvolvida pela SAS (*Statistical Analysis System*). A CRISP-DM ou Processo Padrão Inter-Indústrias para Mineração de Dados, é a metodologia que dá suporte à ferramenta *Clementine* e é voltada para dar suporte às aplicações de mercado.

A Figura 2 demonstra o uso de metodologias de Mineração de Dados nos anos de 2002, 2004, 2007. Observamos que a metodologia mais utilizada nesses anos é a CRISP-DM, pois é vinculada à ferramenta de Mineração de Dados mais vendida no mercado, a SPSS-Clementine. A metodologia SEMMA, desenvolvida pela SAS, foi a terceira metodologia mais utilizada nos anos citados.

Observamos também que o uso de metodologias próprias ocupa a segunda posição do ranking, em todos os anos da pesquisa. Em 2002, o uso de metodologias próprias era da ordem de 23%. Em 2004, seu uso teve um aumento para 28%. No ano de 2007, seu uso diminuiu, caindo para 19%. Isso mostra que nem sempre as ferramentas de mercado atendem às necessidades de aplicação de um processo KDD eque há especificidades referentes ao domínio do problema que necessitam de uma abordagem própria que pode não ser representada pelas ferramentas e metodologias vinculadas já citadas.

Figura 2 - Ranking de Metodologias KDD - 2002, 2004, 2007



Fonte: (KDNUGGETS, 2002), (KDNUGGETS, 2004), (KDNUGGETS, 2007)

A utilização de metodologias próprias tem boa representatividade, porém não há padrões para esse tipo de aplicação. Nesses casos, a aplicação do processo é feita utilizando experiência e metodologias próprias do responsável pelo processo KDD.

Analisando especificamente a metodologia CRISP-DM, está constituída de 6 etapas: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e visualização. Por meio da Figura 3 é possível observar que essas etapas são semelhantes às etapas proposta por (FAYYAD; PIATETSKY-SHAPIRO; PADHRAIC, 1996).

Figura 3 - Metodologia CRISP-DM

Fonte: CHAPMAN, CLINTON, *et al.*, 2000

Uma das características desta metodologia é a definição das tarefas que compõem cada etapa da metodologia CRISP-DM. Na etapa de entendimento do negócio, por exemplo, encontram-se as seguintes tarefas: determinar os objetivos do negócio, avaliação da situação, determinar metas de Mineração e produzir um plano de projeto. Cada tarefa pode gerar uma saída, geralmente um relatório sobre o

resultado da tarefa. Embora CRISP-DM apresente as tarefas e aparentemente controle as sequências de execução, a decisão sobre quais tarefas realizar é ainda dependente do especialista responsável pelo projeto. Observamos, na Figura 3, que há interação entre as etapas de entendimento do negócio e entendimento dos dados, mostrando que pode haver necessidade de retorno à etapa anterior, caracterizando um ciclo de refinamento. De forma semelhante acontece entre as etapas de preparação dos dados e modelagem. Na etapa de avaliação pode-se verificar a possibilidade de retornar à etapa de entendimento do negócio. Porém, cada retorno envolve custos (aspectos financeiros e tempo gasto), o que pode impactar negativamente, inclusive inviabilizando a continuidade do projeto.

Por ser um processo representado por etapas genéricas ou “em alto nível”, o CRISP-DM busca a independência do domínio, o qual pode ser comercial, financeiro, de recursos humanos, produção industrial, etc. Como exemplificação do uso da metodologia CRISP-DM podemos citar as seguintes referências: (GONZALEZ-ARANDA; MENASALVAS; *et al.*, 2008), (PAN, 2009) e (ZENG; PAN, 2010).

A metodologia CRISP-DM foca a responsabilidade de execução do processo no conhecimento tácito do especialista em mineração de dados, o que torna difícil sua condução por profissionais menos experientes, já que não há um maior detalhamento do que é necessário ser feito ou um fluxo que conduz o processo de forma pedagógica.

SEMMA, por definição da empresa que a mantém - *SAS Enterprise Miner* - não é em si uma metodologia para aplicação de KDD (SAS INSTITUTE INC., 2010). Na realidade, é uma organização lógica/funcional das tarefas de Mineração de Dados a serem utilizadas por sua ferramenta.

A metodologia SEMMA começa pela obtenção de uma amostra estatisticamente representativa da base de dados para iniciar um processo piloto (experimental) de extração de conhecimento. Conforme mostrado na Figura 4, as etapas do SEMMA são as seguintes: seleção da amostra, exploração de dados, preparação dos dados, modelagem e avaliação. Embora a metodologia SEMMA não apresente os retornos a etapas anteriores como a metodologia CRISP-DM, espera-se que isso aconteça por conta do conhecimento tácito do especialista.

Figura 4 - Metodologia SEMMA

Fonte: (SAS INSTITUTE AND USING SAS AND ENTERPRISE MINER SOFTWARE, 1998)

Em síntese, foi possível observar que a metodologia SEMMA pode ser aplicada em diferentes domínios, porém a sua condução depende da abordagem e do profissionalismo do responsável pelo projeto. A metodologia CRISP-DM exige um grande nível de especialização e experiência do responsável pelo projeto, pois a condução do mesmo se dá baseado em seu conhecimento tácito. Neste trabalho apresentamos uma proposta de um método para processo KDD com características de uma metodologia pedagógica para conduzir o desenvolvimento de processo de descoberta de conhecimento. O método proposto, denominado PICTOREA, propõe um fluxo para execução de tarefas típicas identificadas à partir de conhecimentos tanto tácitos como explícitos extraídos através do método científico interpretativista.

De acordo com GONZALEZ-ARANDA *et al.* (2008), existe hoje a necessidade de uma metodologia padrão para a aplicação de KDD. Há metodologias como o CRISP-DM, SEMMA. Porém, atualmente, os projetos estão sendo desenvolvidos mais como arte do que ciência, transformando o projeto em algo difícil de entender,

de acompanhar, validar, comparar resultados e reaproveitamento, pois não há uma metodologia padrão.

Na implantação de um processo KDD no mundo dos negócios, encontramos uma grande variedade de cenários, fatores organizacionais, necessidades e preferências dos usuários. Entretanto, muitos processos de KDD se resumem a uma mera aplicação de técnicas de Mineração de Dados e de ferramentas que geralmente pecam ao entregar resultados que satisfazem apenas às 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.

Segundo Cao (2008), atualmente existe uma grande preocupação com processos KDD bem feitos e que sejam aderentes à realidade e necessidade das organizações. Diante dessa necessidade, surge a Mineração de Dados Orientada ao Domínio *(Domain-Driven Data Mining - D3M)* que tem por objetivo reduzir esse conflito, tornando os aspectos e demandas das organizações relevantes para o processo KDD. D3M pode ajudar na construção de uma nova geração de metodologias para aplicação do processo KDD. Com o D3M, cada etapa do processo KDD é acompanhada e validada por um especialista de negócio. Segundo o autor, com a colaboração entre especialistas de TI, especialistas em KDD e especialistas do domínio do problema, o processo de KDD pode se tornar mais aderente e útil na tomada de decisão.

Frente ao cenário da difusão significante de metodologias próprias, da necessidade da correta aplicação de um processo KDD, e levando em conta as necessidades da organização, abre-se um espaço potencialmente fértil para a proposta de padronização de processos de KDD. O uso de um processo padronizado, que leve em conta os princípios de D3M, pode orientar melhor os especialistas de domínio e de mineração de dados no processo de KDD, conduzindo o projeto de forma organizada e viabilizando o controle de interações, iterações e alterações às quais todo projeto de TI está sujeito.

Neste trabalho, como já mencionado, apresentaremos um método denominado PICTOREA, que preferimos definir 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.

Para a construção do método PICTOREA, foi utilizada uma metodologia híbrida utilizando BPM e SPEM. A metodologia BPM (*Business Process Management*) foi utilizada para modelagem dos fluxos das etapas do método PICTOREA, e como metodologia para modelagem de processos foi utilizado o SPEM (*Software Process Engineering Metamodel*).

O processo de KDD elaborado nesse projeto é composto das seguintes etapas: Exploração do Espaço Problema, Definição do Espaço Solução, Entendimento do Domínio do Problema, Caracterização do Problema Através de Atributos, Montagem da Base de Dados, Exploração dos Dados, Preparação dos Atributos, Redução da Dimensionalidade e Seleção de Amostra, Pré-Processamento, Mineração de Dados, Descoberta de Padrões, Validação Estatística e Visualização. Ao final são apresentadas duas avaliações experimentais que mostrarão a viabilidade do método PICTOREA. O método possui um fluxo pré-estabelecido de etapas com documentação detalhada, o que permite um melhor acompanhamento por parte de um superior, facilitando o trabalho em equipe com profissionais menos experientes.

É importante destacar que o projeto está restrito a bancos de dados convencionais e estruturados, excluindo, assim, bancos de dados de multimídia e outros formatos que podem demandar diferentes etapas.

1.2 Objetivos

1.2.1 Objetivo Geral

O objetivo desse trabalho é propor um novo método com caráter pedagógico, denominado PICTOREA, para desenvolvimento, acompanhamento e documentação das etapas e atividades de um projeto de KDD.

1.2.2 Objetivos Específicos

Os objetivos específicos desse trabalho são:

a) identificar quais são os requisitos aos quais um método para o processo KDD deve atender. O conhecimento será adquirido pelo estudo de metodologias existentes, de *frameworks* e de projetos KDD em diferentes domínios.
b) modelar o fluxo das etapas necessárias ao PICTOREA, bem como os atores responsáveis por cada uma delas.
c) formalizar processos utilizando o SPEM como metodologia para modelagem. Especificamente neste trabalho serão formalizadas as etapas de Exploração do Espaço Problema e Definição do Espaço Solução.
d) detalhar a documentação necessária para as etapas selecionadas.
e) construir um protótipo para o método PICTOREA como resultado ao uso do SPEM.
f) aplicar o método proposto em dois estudos de caso.

1.3 Justificativas

Em um projeto de KDD, cada etapa requer análise específica, bem como um tratamento específico.

Conforme observamos na Figura 2, nos anos de 2002, 2004 e 2007, a utilização de metodologia própria foi a segunda opção mais utilizada para projetos KDD. Isso demonstra que as especificidades de um processo KDD podem não ser

satisfeitas pelas metodologias existentes. Dessa forma o especialista define a sua própria metodologia. Isso leva a uma aplicação de um processo KDD sem um padrão definido.

A utilização de metodologias próprias tem boa representatividade, porém não há padrões para essas aplicações. Cada organização ou especialista desenvolve a própria forma de aplicação de um processo KDD.

Diante desse cenário, uma metodologia com uma documentação concisa pode contribuir para a otimização das aplicações de KDD, além de contribuir para a sua difusão.

1.4 Contribuição esperada

Proposta de um novo método para descoberta de conhecimento em banco de dados convencionais que possa oferecer um padrão para os especialistas que utilizam uma metodologia própria.

Espera-se também, uma contribuição para a disseminação da aplicação de descoberta de conhecimento nas organizações, pois observa-se que os principais projetos KDD são restritos ao âmbito científico.

Além disso, seu caráter pedagógico poderá contribuir para o aprendizado de profissionais menos experientes.

1.5 Organização do trabalho

Este trabalho está organizado em cinco seções. A seção 2 apresenta a revisão bibliográfica e a fundamentação teórica necessárias para o desenvolvimento do método PICTOREA. Na seção 3, a metodologia de trabalho é apresentada. Na seção 4, o desenvolvido do método PICTOREA é apresentado. Na seção 5 são apresentadas as avaliações experimentais para validação do método. E, finalmente, as considerações finais na seção 6.

2 REVISÃO BIBLIOGRÁFICA E FUNDAMENTAÇÃO TEÓRICA

2.1 Considerações iniciais

Segundo Yang e Wu (2006), um dos principais desafios envolvendo Descoberta de Conhecimento em Banco de Dados é encontrar uma teoria canônica que possa gerar uma metodologia unificada capaz de orientar melhor as pesquisas de forma a minimizar a ocorrência de erros comuns no processo. Ainda segundo os autores, há uma necessidade de metodologia unificada para etapas que geralmente não são descritas pelas metodologias existentes, como a etapa de limpeza de dados. Atualmente, é possível construir modelos e encontrar padrões de forma rápida através das diferentes ferramentas disponíveis. Porém, um processo de descoberta de conhecimento executado criteriosamente requer maior tempo. Segundo os autores, aproximadamente 90% dos custos do projeto normalmente são provenientes do pré-processamento (integração de dados, limpeza de dados, análise de *outliers*, etc).

Esta revisão bibliográfica foi organizada de forma a apresentar uma descrição das etapas típicas de um processo KDD, uma revisão de metodologias e *frameworks* aplicados, uma revisão sobre a aplicação do processo KDD em diferentes domínios, um estudo sobre D3M e uma revisão sobre modelagem de processos. Todos estes temas ajudaram na construção do método PICTOREA.

2.2 Etapas típicas de um processo KDD

Em sua concepção, como já mencionado, o processo KDD estabelece 5 etapas principais. As metodologias que foram sendo propostas para o processo obedecem ou adaptam essas 5 etapas propostas por (FAYYAD; PIATETSKY-SHAPIRO; PADHRAIC, 1996). Essas etapas são: Seleção de Dados, Pré-processamento, Transformação dos Dados, Mineração de Dados e Interpretação.

Na etapa de seleção de dados define-se o domínio do problema a ser explorado e seleciona-se o conjunto de dados que se relaciona com o domínio definido. Nessa etapa já se apresenta o desafio de selecionar o conjunto de dados que de alguma forma possa "explicar" o domínio do problema.

O pré-processamento consiste em preparar os dados para a aplicação do processo KDD, isto é, além de migrar os dados selecionados de fontes externas para um repositório local, é necessário limpar esses dados, eliminando erros, inconsistências e dados ausentes.

A etapa de Transformação de Dados é necessária basicamente para adequar os dados selecionados de forma que possam servir de entrada para os algoritmos de Mineração de Dados. Além disso, são aplicadas sumarização e categorizações, etc.

A etapa de Mineração de Dados consiste na extração de padrões através da execução de algoritmos de inteligência artificial ou de aprendizado de máquina que utilizam, dentre outras funções, a classificação, a regressão, o agrupamento, etc.

A etapa de interpretação consiste na apresentação dos padrões de forma a serem úteis. A análise da relevância e a qualidade dos padrões encontrados é verificada e interpretada.

2.3 Revisão de metodologias, modelos e *framework* aplicados ao processo KDD

Como pode ser observado através da literatura, alguns autores propuseram diversas metodologias para aplicação do processo KDD, muitas delas com alto nível de abstração e que geralmente são focadas sobre um domínio de problema específico. Entender com qual intuito essas metodologias foram propostas e identificar nelas seus pontos positivos e falhos é importante para o método proposto neste trabalho.

Como já mencionado anteriormente, de acordo com Gonzalez-Aranda, Menasalvas, *et al.* (2008), há a necessidade de uma metodologia padrão para a aplicação de KDD. Há padrões como o CRISP-DM, SEMMA, PMML. Porém, atualmente, os projetos estão sendo desenvolvidos mais como arte do que ciência, transformando o projeto em algo difícil de entender, de validar e comparar resultados. Os autores defendem a utilização de CRISP-DM e se concentram na fase de concepção do projeto para determinar um planejamento correto. A fase de concepção do projeto descrita em CRISP-DM compreende a etapa de Entendimento do Problema com as tarefas de Determinar os Objetivos do Negócio, Avaliação da Situação e Produzir um Plano de Projeto. Os autores defendem que essa etapa

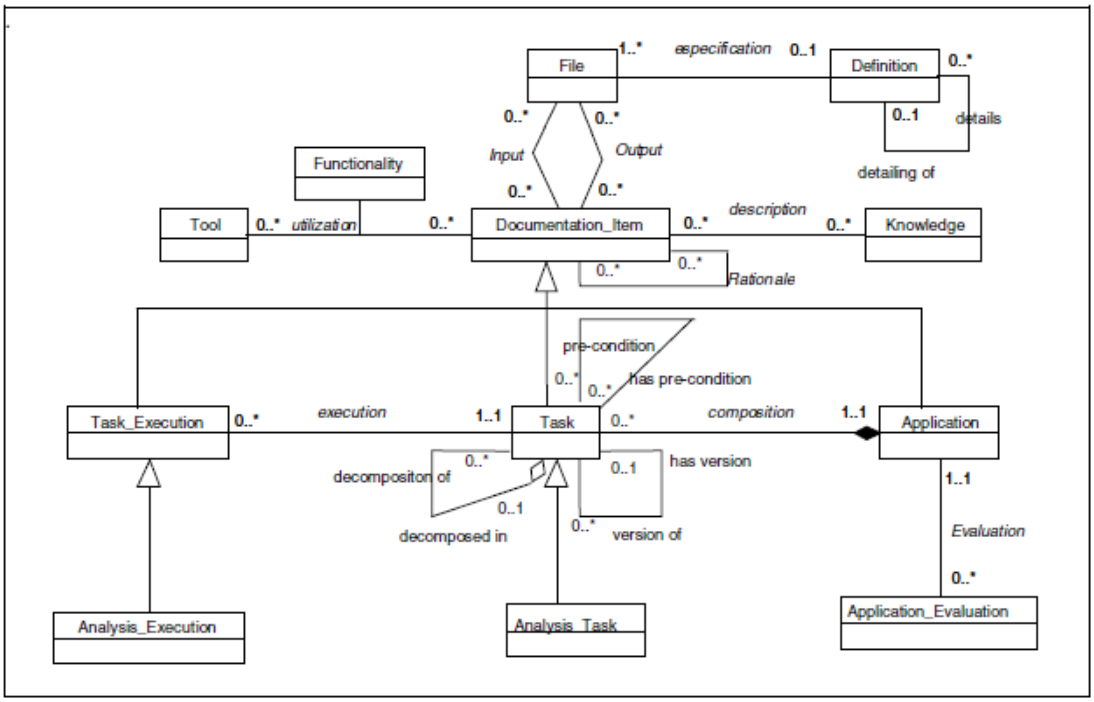
pode determinar a qualidade e o atendimento às expectativas referentes a um processo KDD.

Dessa forma, para o PICTOREA, essa fase inicial de concepção deverá receber grande atenção e a documentação deverá ser detalhada para que as etapas seguintes não se distanciem do escopo-problema-expectativa, inicialmente identificado.

Ghedini e Becker (2001) apresentaram um modelo de documentação para estruturar e organizar as informações necessárias para gerenciar uma aplicação KDD, baseado na premissa de que a documentação é importante não apenas para um melhor gerenciamento de recursos e resultados de um projeto KDD, mas também para guardar e reusar as experiências adquiridas.

O modelo apresentado gera uma documentação simples das atividades, ajudando o analista a rastrear tarefas, recursos e resultados. Os autores não mostram o tipo de documentação e os formalismos adotados na elaboração dos mesmos. Entretanto, fica evidente que para uma metodologia de qualidade é necessário que exista uma documentação detalhada para que o processo se torne gerenciável.

Figura 5 - Diagrama de classes do modelo



Fonte: Ghedini e Becker, 2001

Em Boente, Goldschmidt e Estrela (2006), os autores mostram que a complexidade inerente ao processo KDD decorre diretamente da enorme diversidade de alternativas de ação que surgem ao longo do processo. O artigo tem como objetivo descrever a modelagem de um sistema de informação que viabilize a aplicação da metodologia para realização de KDD. Os autores propõem um sistema, já desenvolvido, que vem sendo utilizado com sucesso como apoio ao processo de ensino de KDD em disciplinas oferecidas pelos autores tanto em nível de graduação quanto em nível de pós-graduação. Para modelagem foram utilizados recursos da linguagem de modelagem unificada UML. Os modelos de casos de uso e de classes do sistema foram discutidos. Observamos através desse artigo, que a UML tem importante papel na modelagem e entendimento de processos e descrição de *frameworks* para aplicações KDD. Dessa forma, para o PICTOREA, a UML será utilizada como notação formal.

Não relacionada diretamente com a proposta de metodologia para processos KDD, a seleção de ferramentas traz informação acerca dos requisitos demandados pelo especialista na descoberta de conhecimento.

Em Britos, Merlino, *et al.* (2006), os autores propõem critérios de avaliação para escolher a ferramenta de Mineração de Dados, dentre as oferecidas no mercado, que mais se adeque à organização. A metodologia é dividida em 9 fases, sendo que para o desenvolvimento de uma metodologia de processo KDD, devemos analisar as duas primeiras fases. As fases 3 a 9 estão diretamente relacionadas com busca e entrevistas com os fornecedores das ferramentas. Na fase 1, os autores descrevem a importância de se conhecer a própria organização e seus processos, para isso são documentados informações sobre departamentos, funções e os objetivos a serem alcançados pela organização com a adoção de uma ferramenta para mineração de dados. Percebe-se neste item, que o primeiro passo para um processo KDD está no bom entendimento do domínio do problema. Na fase 2, são analisadas as necessidades da organização que deverão ser satisfeitas pela ferramenta de mineração de dados, bem como os objetivos da mesma no sentido de tornar a tecnologia em questão, uma parceira estratégica. Segundo os autores, essas duas primeiras fases definirão o tipo de busca e análise que farão no mercado. Pode-se observar a busca do mercado por ferramentas aderentes às necessidades das organizações. O estudo deste artigo nos ajudou a entender os requisitos a serem observados para a aquisição de uma ferramenta de Mineração de

Dados que permita o desenvolvimento de uma metodologia que atenda às necessidades reais de uma organização. Dentre os principais requisitos descritos pelos autores destacamos:

a) documentação concisa
b) documentação de uma etapa de entendimento do negócio
c) documentação das necessidades e expectativas
d) necessidade da ferramenta ser utilizada com uma metodologia que acompanhe todo o ciclo de vida da solução
e) interface amigável
f) geração de relatórios e gráficos

Em Soundararajan, M, *et al.* (2004) os autores fazem uma análise das principais ferramentas e técnicas de KDD para gerenciamento de conhecimento em uma biblioteca digital. Ao final são dadas sugestões para a melhoria de performance das ferramentas KDD. Algumas das sugestões para otimização de performance de sistemas Mineração de Dados são:

a) redução do tamanho de *dataset* considerado na descoberta de conhecimento
b) descarte de descobertas inconsistentes, redundância e conhecimento trivial
c) definição do algoritmo mais eficiente de Mineração de Dados
d) empregar sistemas de alta performance para melhorar o processo de descoberta de conhecimento

Estas sugestões são importantes para a documentação gerada pelo PICTOREA. Os requisitos de desempenho deverão ser documentados e discutidos para até mesmo uma análise de viabilidade por parte do responsável pelo processo.

2.4 Uma revisão crítica sobre a aplicabilidade de processos KDD em diferentes domínios

Entender como os especialistas vêm aplicando o processo KDD em diferentes domínios de problema e identificar quais são os principais requisitos que estão sendo atendidos foi crucial para a metodologia PICTOREA. Observamos que embora aplicados em distintos domínios, o processo KDD possui aspectos comuns a todos eles. Identificar e documentar esses aspectos fará com que o PICTOREA possa ser melhor aplicado.

Em Goebel e Gruenwald (1999), os autores fazem uma análise das tarefas típicas em KDD e as abordagens que devem ser utilizadas para essas tarefas. Propõem um esquema de classificação que pode ser utilizado no estudo de softwares de Mineração de Dados. Esta classificação é baseada em características comuns dos softwares, conectividade de banco de dados e características de Mineração de Dados. Os autores analisaram 43 *softwares* comerciais. Os principais aspectos analisados que definem a classificação e vistos como mais importantes pelos autores são:

a) conectividade - Capacidade de se conectar em diferentes fontes de dados como arquivos ASCII, Oracle , MS Excel, etc.
b) características diretamente relacionadas com mineração de dados - tarefas de mineração (classificação, regressão, etc), algoritmos (Redes neurais artificiais, algoritmos genéticos, etc).

Goebel e Gruenwald (1999) nos mostram quais aspectos devem ser observados e analisados no momento da escolha de uma ferramenta que dê suporte ao método PICTOREA.

Em Lee, Stolfo e Mok (1999), é proposto um *framework* para aplicação de processo KDD com o objetivo de detecção de invasões em sistemas computacionais. O modelo é executado sobre dados de auditoria de sistema com o intuito de encontrar padrões e informações que caracterizam uma invasão ao sistema. O *framework* apresenta as fases de pré-processamento, classificação e detecção de anomalia. Observamos que o autor não utiliza um método padrão para a aplicação do processo KDD. Por isso, o processo é feito em sequências de

tentativa e erro. Dessa forma, acreditamos que o método PICTOREA poderia contribuir para aplicação apresentada pelo autor.

Em Elsila e Roning (2002) foi descrita a aplicação do processo KDD em uma indústria de extração de aço. O foco principal é na limpeza e enriquecimento dos dados de acordo com o tipo de dado disponível. A forma como a fase de pré-processamento é abordada no artigo mostra a importância da boa definição de estratégias de limpeza e enriquecimento dos dados, além da necessidade de documentação detalhada das estratégias utilizadas e para que se possa avaliar a etapa visando a qualidade do processo.

De acordo com Freitas, Brazdil e Pereira (2005), em banco de dados de hospitais há um grande acumulo de dados heterogêneos e esses dados não são analisados como deveriam para uma tomada de decisão.

A proposta dos autores é ajudar a suprir essas dificuldades e limitações em hospitais e departamentos de sistemas de informação de saúde, facilitando o acesso de gerentes e administradores à informação através da análise de seus bancos de dados. Os autores citam como trabalhos futuros a validação dos resultados pelos especialistas de cada área. O processo KDD pode trazer positivas contribuições nas tomadas de decisão feitas por profissionais da saúde e, consequentemente, o melhor gerenciamento de pacientes e da organização. Além do estudo de aplicação KDD na área médica, o artigo evidencia a necessidade de uma documentação detalhada das etapas e tarefas do processo KDD, uma vez que os autores citam dificuldades enfrentadas devido à falta de documentação, tais como retrabalho, falta de medição de esforços e custos, etc. Possivelmente, esses problemas poderiam ser evitados caso houvesse um documento das etapas de forma que o processo pudesse ser melhor gerenciado.

2.5 Mineração de Dados Orientada ao Domínio – D3M

Em Cao (2008), o autor detalha como é feita a Mineração de Dados Orientada ao Domínio - *Domain-Driven Data Mining (D3M)*.

Na implantação da Mineração de Dados no mundo real dos negócios, nós temos uma grande variedade de cenários, fatores organizacionais, preferências dos usuários e necessidades. Entretanto, os atuais algoritmos de Mineração de Dados e ferramentas geralmente pecam ao entregar resultados 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 mercado. Para reduzir este conflito, tornar os fatores do mundo real relevantes à Mineração de Dados e tornar a Mineração de Dados mais útil ao suporte de decisão no mundo real, o autor propõe a metodologia de Mineração de Dados Orientada ao Domínio (D3M).

Cada etapa do processo KDD é acompanhada e validada por um especialista de negócio. Com a colaboração entre especialistas de TI, especialistas em KDD e especialistas do domínio do problema, o processo de KDD se torna mais aderente e útil na tomada de decisão.

No Quadro 1, adaptado de Cao (2010), há um comparativo entre o processo tradicional de KDD e a utilização de D3M.

Quadro 1 - Comparação entre KDD tradicional e D3M

Aspectos	Método Tradicional	D3M
Geral	Os dados orientam o processo	Dados e especialistas participam da criação de soluções para resolução de problemas
Objetivo	Inovação e melhoria de algoritmos	Efetivamente solucionar o problema
Processo	Sequencia contínua de passos	Múltiplos passos, iteração e interação em cada tarefa
Mecanismo	Automatizado	Centrado no ser humano ou cooperação entre ser humano e mineração
Infraestrutura	Padrões fechados de sistemas de numeração	Interações em sistemas de resolução de problemas em um ambiente aberto
Usabilidade	Modelos e processos pré-definidos	Ad-hoc, dinâmica e modelos e processos customizados
Entregáveis	Padrões	Análises de negócios para apoio na tomada de decisões
Desenvolvimento	Sólida validação	Bem fundamentada na resolução do problema

Fonte: Adaptado de CAO, 2010

Observa-se, através da análise do Quadro 1, que uma das principais características de D3M é centrar o processo no atendimento das necessidades e expectativas do especialista de domínio. Na proposta deste trabalho, durante o processo KDD, há pontos de avaliação constantes de forma que o especialista de domínio norteie e participe na decisão dos passos e configurações necessárias à continuidade do processo.

2.6 Modelagem de processos

Modelar processos é uma preocupação antiga. Há um século pesquisadores tentam estabelecer padrões de modelagem de processos na tentativa de melhora-los. Frederick W. Taylor (1856 - 1915) estudou os processos sobre o prisma de tempos e movimentos, o que permitia aumentar a produtividade tornando os processos mais eficientes.

Segundo Graham (1999), no início do século passado, Frank Bunker Gilbreth (1868 - 1924), cujo trabalho ajudou na criação dos princípios da administração científica, procurou encontrar a melhor forma de modelar processos tendo como visão a fadiga e o movimento, os quais, segundo o autor, estão relacionados. Todo movimento que fosse reduzido diminuiria a fadiga. Ele propôs um conjunto de ferramentas com potencial para melhorar processos, sejam eles industriais ou de negócios. Esse conjunto de ferramentas foi chamada por ele de “Simplificação de trabalho”. Em 1940, Ben S. Graham trouxe métodos de fábrica para o escritório e criou o *Graham Process Chart* que consiste em uma metodologia para múltiplos fluxos de trabalho. Dessa forma, ele desenvolveu uma metodologia com uma abordagem de equipe para melhoramentos de processos de negócios.

Graham (1999) argumenta que para toda modelagem de processos, um dos grandes desafios é gerenciar a resistência que os colaboradores geralmente têm a respeito de qualquer mudança. Essa resistência, geralmente se baseia em medo de perder funções ou até mesmo o emprego. Para o responsável pela modelagem de processo é importante deixar claro que em toda a modelagem as pessoas devem ser reconhecidas como recursos e não como despesas que precisam ser cortadas.

Pensando nessa abordagem, Graham (1999) definiu alguns aspectos que toda modelagem de processos de negócios deve considerar. Decidimos caracteriza-las em dois grandes grupos para facilitar o entendimento conforme a seguir:

1. Documentação
    a) todos os detalhes devem ser documentados.
    b) documentar o que foi realizado em cada etapa do processo.
    c) documentar onde o processo inicia e o fluxo de etapas que ele segue.
    d) documentar quem é responsável pela tarefa e cada pessoa que participa de qualquer mudança.

e) organizar as informações por meio de um diagrama de fluxo de processos.

2. Método de trabalho
   a) não procurar detectar falhas no processo. Somente se preocupar em representar as atividades.
   b) representar o ciclo de trabalho NORMAL, sem dar ênfase nas exceções.
   c) ser metódico - seguir um fluxo e lista de passos em ordem.
   d) sempre esclarecer dúvidas em relação às etapas do processo.
   e) entrevistar a pessoa que executa a tarefa, não a pessoa responsável pelo gerenciamento.
   f) entrevistar os envolvidos na área onde uma tarefa específica acontece.
   g) primeiramente procurar entender fatos através da observação - as pessoas podem mostrar melhor os detalhes do que elas podem descreve-los.
   h) respeitar calendários e cronogramas de trabalho e possíveis interrupções.
   i) deixar claro o que você está fazendo e porque eles estão envolvidos.

O autor explica que um fluxo de tarefas é capaz de conduzir os envolvidos além de tornar mais fácil o entendimento do processo, servindo como um guia para a condução do mesmo processo. Dessa forma, os benefícios de se ter um processo de negócio modelado vão desde a melhora de produtividade da equipe até à diminuição de esforço e tempo necessário para o aprendizado de um novo integrante da equipe.

Verificamos que um processo KDD possui características semelhantes a um processo de negócios, principalmente no que concerne à papéis de usuários, necessidade de documentação, possibilidade de retornos e possibilidade de evolução.

Ainda em Graham (1999) são definidas alguns aspectos para se construir diagramas para representação de processos que justificam sua utilização, dentre elas quebrar processos em passos, para facilitar o entendimento, e uso de símbolos para representação de cada tarefa. Conexões e sequencias complexas são mais

facilmente representadas, além disso, através de um diagrama, qualquer parte do processo pode ser localizada imediatamente.

Segundo o Ben Graham, um diagrama de fluxo de processo deve ter os seguintes requisitos:

a) os itens utilizados para representar processos devem ser representados em linhas horizontais.
b) cada linha deve possuir uma identificação que represente o item.
c) cada passo de trabalho deve ser identificado por um símbolo que represente o que acontece ao item em determinado ponto do processo.
d) os símbolos devem ficar posicionados em sequência ao longo das linhas de itens.

Ainda segundo o autor, para modelagem de processos, a experiência é fator importante.

Recker, Indulska, *et al.* (2005), afirma que atualmente estamos vivendo a “padronização das organizações”. Nas últimas três décadas foram desenvolvidas diversas técnicas, métodos e ferramentas para a modelagem de gerenciamento de processos. Segundo o autor, uma tentativa de consolidar essas diversas opções é a Metodologia para Modelagem de Processos de Negócios (*Business Process Management* - BPM).

Devido ao crescimento da popularidade da modelagem de processos, cresceu muito o número de técnicas e ferramentas para este fim. Hommes (2004), em sua tese de doutorado, pesquisou as metodologias e ferramentas existentes e desenvolveu um método para validação da qualidade das metodologias para a construção de processos existentes. Para isso definiu quais são os requisitos necessários para uma modelagem de qualidade. Primeiramente a metodologia para modelagem deve ser capaz de identificar conceitualmente os aspectos dos objetos do mundo real e representar o relacionamento existente entre estes objetos. Nesse ponto, Hommes (2004) explica as diferenças entre técnica de modelagem e ferramenta de modelagem. A técnica de modelagem tem por objetivo oferecer um caminho certo a ser seguido na conceituação e documentação da realidade. A ferramenta de modelagem é um sistema automatizado que aplica as técnicas de modelagem que a suporta.

Conforme descrito por Hommes (2004), no que concerne à avaliação de qualidade de modelos, os critérios mais importantes identificados na literatura são discutidos abaixo:

a) integralidade - Na medida em que o modelo possui todas as etapas e instruções que podem ser expressas sobre o domínio modelado.
b) exatidão
    - sintaxe correta - Na medida em que as instruções expressas no modelo estão em conformidade com a sintaxe da linguagem de modelagem utilizada.
    - semântica correta - Na medida em que as instruções expressas no modelo estão em conformidade com o domínio modelado.
c) consistência - Na medida em que as instruções expressas no modelo não contradizem umas às outras.
d) compreensibilidade - A facilidade com que conceitos e pensamentos no modelo podem ser compreendidos pelos usuários do modelo.

De acordo com Hommes (2004), não há uma só metodologia de pesquisa para todas as pesquisas científicas e são as características individuais de cada projeto de pesquisa que determinam a metodologia mais adequada para o projeto.

Hommes (2004) relaciona 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 é sistematicamente adquirir conhecimento objetivo sobre o fenômeno conhecido. Em sua tese, Hommes (2004) aponta problemas relacionados a essa abordagem. O autor argumenta que dessa forma há uma observação individual da realidade e assim não há garantia que a imagem percebida da realidade seja realmente a realidade, uma vez que até os filósofos esclarecem que temos limitações na percepção da 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. De toda forma, analisando a tradição interpretativista, a pesquisa científica contribui para o entendimento do mundo através de construções de teorias científicas e artefatos que nos ajudam a conceituar e perceber esse mundo. (HOMMES, 2004).

De acordo com Mingers (2001), para o desenvolvimento de qualquer metodologia de pesquisa, deve-se considerar as diferentes dimensões da situação real, material, social e pessoal nas tarefas que envolvem os diferentes estágios da pesquisa e seu contexto.

O método PICTOREA foi concebido levando em consideração a tradição interpretativista. Neste trabalho, foi criado um método para aplicação de processo KDD baseado em conhecimentos adquiridos pelo pesquisador através de entrevistas com um especialista e definições dos principais requisitos necessários para um processo de KDD de qualidade.

Segundo Jorgensen e Anniken (2004), sistemas de processos de negócios foram construídos para automatizar os procedimentos de rotina. Automação exige bom entendimento do domínio, dos processos repetitivos, organização clara de papéis, uma terminologia estabelecida e planos pré-definidos. Em relação à conhecimento de trabalho, planos de conhecimento de processos são elaborados e reinterpretados conforme o trabalho progride. Modelos de processos interativos são criados e atualizados pelos participantes do projeto de acordo com a evolução dos planos.

Um sistema de processo interativo deve conter as seguintes características:

a) permitir a modelagem pelos usuários.
b) integrar o apoio para o trabalho *ad-hoc* e de rotinas.
c) dinamicamente personalizar a funcionalidade e interfaces.
d) integrar a aprendizagem e gestão do conhecimento.

Com o advento da computação, a área de desenvolvimento de softwares cresceu. Os sistemas surgem para automatizar atividades do mundo real e procuram atender às necessidades em sua completude. Diversas metodologias foram desenvolvidas para padronizar o processo de desenvolvimento de software,

uma vez que a busca por qualidade e métodos de avaliação de softwares crescem com a utilização dos mesmos em diversos setores.

Um modelo de processo de software é uma representação das atividades do mundo real de um processo de produção de software. (GENVIGIR, SANTANNA, *et al.*, 2003). Um modelo de processo de software é desenvolvido, analisado, refinado, transformado e/ou representado dentro de um meta-processo. Assim, qualquer modelo de processo de software deve modelar adequadamente o processo do mundo real e deve atender às exigências específicas de cada fase do meta-processo.

Em Genvigir, Santanna, *et al.* (2003), o autor salienta que os processos precisam ser formalizados ou padronizados de forma a entendermos o seu funcionamento. Isso possibilita o melhor treinamento, as propostas de melhoria dentre outros fatores. Um modelo de processo deve especificar os pré-requisitos e consequências de cada tarefa, bem como a sincronização entre essas tarefas.

Em Humphrey (1989), o autor define os principais objetivos para se modelar um processo:

a) possibilitar a comunicação e o entendimento efetivo do processo.
b) facilitar a reutilização do processo (padronização).
c) apoiar a evolução do processo.
d) facilitar o gerenciamento do processo.

Ainda segundo Humphrey (1989), as principais razões que levam à padronização de processos organizacionais são:

a) permitir treinamento, gerenciamento, revisões e ferramentas de suporte.
b) utilizando-se processos padronizados, cada experiência de projeto pode contribuir para a melhoria dos processos na organização.
c) um processo padronizado fornece uma base estrutural para medição.
d) definições de processo levam tempo e esforço, o que torna impraticável novas definições de processo para cada projeto.

Em meio à essa necessidade surge o SPEM - *Softwares and Systems Process Engineering Meta-model*, que é um metamodelo proposto pela *Object Management Group* - OMG, para descrição de um processo concreto de desenvolvimento de software ou uma família relacionada de processos de desenvolvimento de software. SPEM aparece como uma proposta de unificação entre as diferentes metodologias propostas para modelagem de processos.

Dentre suas características, destacam-se a utilização da orientação a objeto para modelar uma família relacionada de processos de software e a utilização de *Unified Modeling Language* - UML como notação.

A modelagem de processos de software iniciou através da utilização de técnicas de análise de sistemas para representar um processo. As linguagens de modelagem de processos (PML) surgiram como forma de agregar diversos elementos utilizados por técnicas de análise de sistemas. Nos últimos 10 anos, muitas PML's foram propostas e em maio de 2000, a OMG (*Object Management Group*) publicou UPM (*Unified Process Model*) como uma proposta de unificação entre as diferentes metodologias para modelagem de processos. Em 2002, a mesma organização publicou a versão 1.0 do SPEM, que é o metamodelo para definição de processos e seus componentes.

Diferente da maioria das PMLs, o SPEM aparece como uma proposta de unificação entre as diferentes metodologias propostas para modelagem de processos. Essa necessidade de unificação surgiu, pois as organizações, pesquisadores, universidades criam suas metodologias e nomeiam os elementos dos processos com diferentes termos causando grande dificuldade para o desenvolvimento de produtos e pesquisas. (GENVIGIR; SANTANNA; *et al.*, 2003).

A modelagem de um processo deve ser conduzida de modo a possibilitar o entendimento e a padronização do processo. Através de processos padronizados, cada experiência de projeto pode contribuir para a melhoria dos processos como um todo. Além disso, fornece uma base estrutural para medição. (GENVIGIR; SANTANNA; *et al.*, 2003).

2.7 Considerações finais

Para a construção do método PICTOREA, os fundamentos da modelagem de processos, juntamente com os fundamentos D3M e as aplicações em diferentes domínios, foram analisados em paralelo ao conhecimento tácito do especialista em processos KDD.

3 METODOLOGIA

3.1 Considerações iniciais

Como estabelecido, este trabalho tem como objeto-problema o desenvolvimento de um método para modelagem de processos KDD. Este método envolve dois procedimentos: o primeiro procedimento visa o entendimento do objeto-problema e o segundo procedimento visa a representação do mesmo.

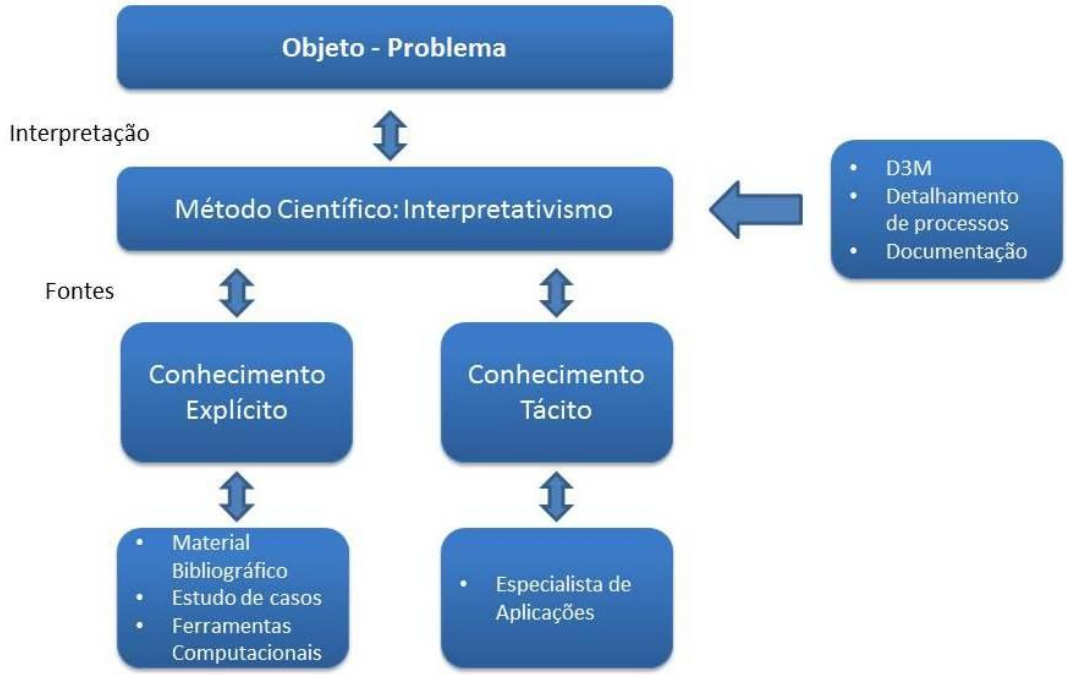
Neste capítulo são apresentados os procedimentos adotados na construção do método para aplicação de processos KDD.

3.2 Entendimento do objeto-problema

Como já mencionado anteriormente, as metodologias mais utilizadas no mercado são dependentes de ferramentas, como é o caso de CRISP-DM e SEMMA. Em outras situações, observa-se que o uso de metodologias próprias, definida pelo especialista envolvido no processo KDD, carecem de um padrão de execução e documentação. Um método que estabeleça os procedimentos a serem adotados para o desenvolvimento de processos KDD, pode contribuir para o grupo de especialistas que utilizam um método próprio. Além disso, o método PICTOREA permite um bom gerenciamento de todas as etapas de um processo KDD, com o intuito de permitir um gerente de projetos mais experiente coordenar um ou mais projetos executados por profissionais menos experientes.

A proposta metodológica foi construída através do método científico interpretativista, demonstrado na Figura 6, e este será descrito a seguir.

Figura 6 - Procedimento para entendimento do objeto-problema



Fonte: Elaborado pelo autor

3.2.1 Método interpretativista

Segundo o método interpretativista, a percepção e interpretação da possível realidade não podem ser separadas do pesquisador. O conhecimento depende das interpretações do fenômeno ou da realidade. (HOMMES, 2004).

Dessa forma, e de acordo com o princípio interpretativista, neste trabalho, o conhecimento acerca de descoberta de conhecimento em bancos de dados foi construído ao longo do processo de desenvolvimento do método PICTOREA.

Seguindo o método citado, houve interação entre o pesquisador e o especialista em KDD através de entrevistas, nas quais procurou-se alinhar o conhecimento tácito com o conhecimento pesquisado na literatura (conhecimento explícito) para a definição de um método que fosse aderente às necessidades tanto acadêmicas quanto de mercado. Como resultado, foi criado um método unindo o empírico ao teórico e com a ideia de aproximação para uma teoria canônica.

3.2.2 Captura de conhecimento tácito

Como já mencionado, procurou-se analisar o objeto-problema através do conhecimento tácito de um especialista em processo KDD concomitantemente à pesquisa (explícita) de fundamentos acerca de modelagem de processos, mineração de dados orientada ao domínio - D3M e aplicações de processos KDD em diferentes domínios. Através da interação entre o especialista em processos KDD e o pesquisador, foram identificados os aspectos e etapas principais para a condução de um processo KDD de forma a garantir sua qualidade e aderência a uma necessidade de mercado. Nestas interações, as informações encontradas na literatura foram confrontadas com a experiência do especialista para a definição de uma melhor abordagem das etapas do novo método sendo proposto. Para a captura do conhecimento tácito foi utilizado o método interpretativista apresentado na seção 3.2.1.

3.2.3 Captura de conhecimento explícito

Como visto no capítulo anterior, foram definidos três grandes áreas para captação de conhecimento explícito: Mineração de Dados Orientada ao Domínio -

D3M, aplicações de projetos KDD em diferentes domínios e estudo de *framework* para aplicação de projetos KDD.

3.2.3.1 Mineração de Dados Orientada ao Domínio – D3M

Como já mencionado no capítulo 2, D3M surgiu com a necessidade de tornar processos KDD mais aderentes à realidade dos negócios. Os estudos deste tema foram utilizados durante as reuniões com o especialista KDD para que fossem absorvidos no método PICTOREA. Durante todo o processo de desenvolvimento os aspectos D3M foram considerados, tais como processo centrado no ser humano, foco na solução do problema, especialistas de negócio e de KDD participam na criação de soluções para resolução do problema, dinamicidade e expectativa de negócios (cliente). Estes aspectos estão descritos no quadro 1 presente no capítulo 2.

3.2.3.2 Aplicações do processo KDD em diferentes domínios

Neste trabalho, através da literatura, foram identificadas as principais etapas da aplicação de processos de KDD em diferentes domínios. O foco foi verificar os aspectos comuns a estas aplicações de domínio.

Nas publicações referentes à aplicação de KDD, encontrou-se uma limitação: os autores, por vezes, não descrevem detalhadamente os passos utilizados no processo. Geralmente, eles focam em descrever a técnica e/ou algoritmo de mineração de dados utilizada e seus resultados, não informando por exemplo, quais critérios foram utilizados para a seleção de dados.

Embora esta limitação estivesse presente, verificou-se em algumas publicações algumas considerações relevantes para o método PICTOREA, como a importância de uma documentação dos objetivos e resultados de cada interação, bem como quais informações precisam ser documentados.

As informações obtidas através da revisão de literatura foram discutidas nas entrevistas com o especialista KDD para que fossem também absorvidas pelo método PICTOREA.

3.3 Representação do problema

Como já apresentado anteriormente, BPM (*Business Process Management*) é uma metodologia para modelagem de processos. Verificamos que a representação do método proposto neste trabalho constitui-se de um processo, ou fluxo de passos principal. Este fluxo será apresentado no próximo capítulo através de BPMN (*Business Process Management Notation*).

O fluxo principal do método PICTOREA, detalhado no próximo capítulo, e as etapas do fluxo principal (descritas no capítulo 4) foram modelados utilizando a ferramenta para modelagem de processos em notação BPMN chamada BizAgi Process Modeler (BIZAGI, 2011). Trata-se de uma ferramenta gratuita que dispõe os conceitos em caixas e as relações entre eles são especificadas através de frases de ligação, que unem cada um dos conceitos, como requer a metodologia BPMN.

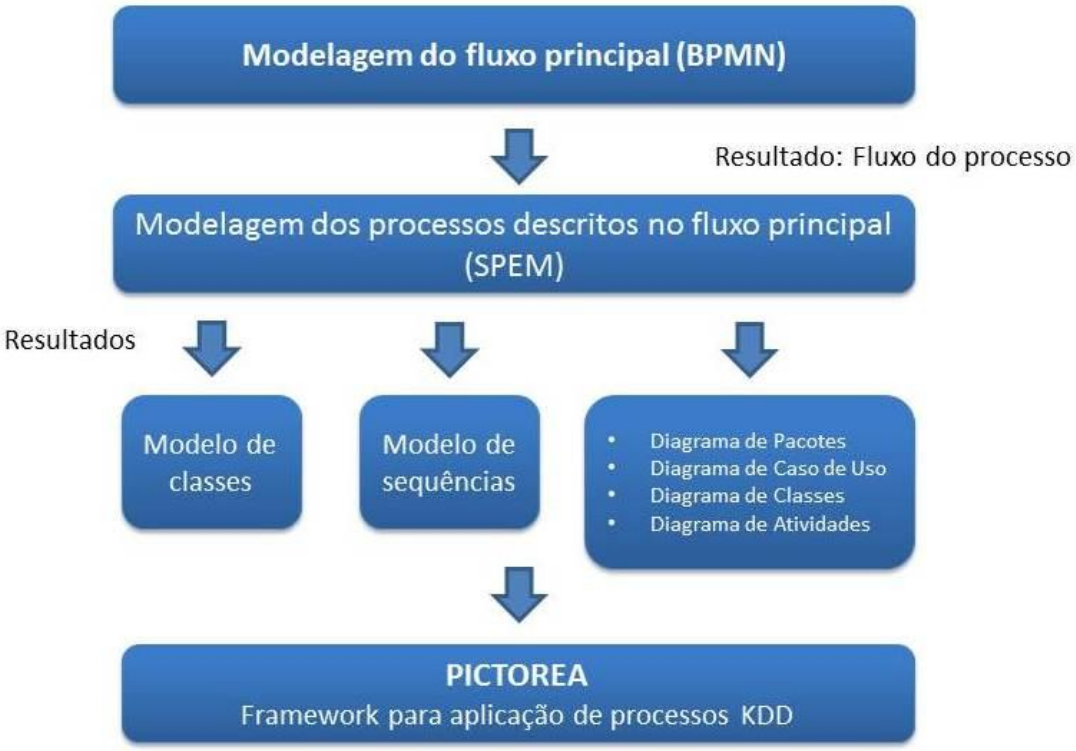
Uma modelagem de processo deve ser conduzida de modo a possibilitar o entendimento e a padronização do processo. Neste trabalho optou-se pela utilização de SPEM para modelagem dos processos do método PICTOREA, pois dentre outros fatores já discutidos no capítulo 2, destacamos a possibilidade de comunicação e entendimento efetivo do processo, permitir a reutilização do processo (padronização), apoio para a sua evolução e melhor possibilidade de gerenciamento do método.

Neste projeto, foram utilizados alguns diagramas do SPEM, tais como:

a) diagrama de Caso de Uso - Para representação detalhada das atividades do processo e a relação entre cada responsável por cada atividade.
b) diagrama de Atividades - Para representação da sequência e da ordem das atividades executadas pelo processo (entendimento do problema), bem como quem é o seu responsável.

A Figura 7 mostra os elementos e o fluxo para a representação do objeto-problema.

Figura 7 - Procedimento para representação do objeto-problema



Fonte: Elaborado pelo autor

4 DESENVOLVIMENTO DO MÉTODO PICTOREA

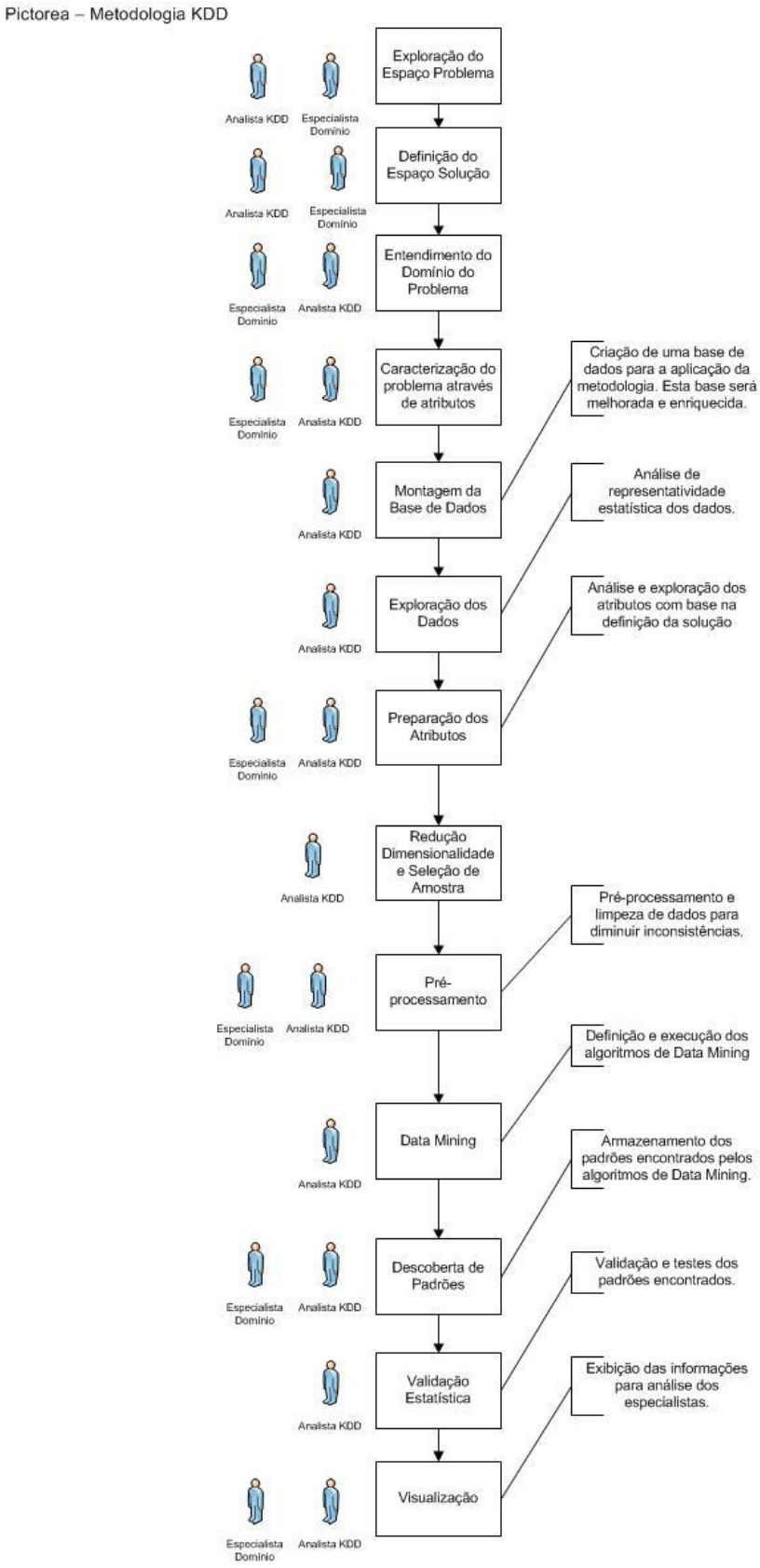
4.1 Considerações iniciais

Como mencionado no capítulo 2, para a construção do método PICTOREA foi utilizada uma metodologia híbrida utilizando BPMN e SPEM. A notação BPMN (*Business Process Management Notation*) foi utilizada para modelagem do fluxo principal e etapas do método PICTOREA, e o detalhamento das etapas foi feito utilizando o SPEM.

4.2 Fluxo principal em BPMN

Cada etapa representada no fluxo principal demonstrado na Figura 8, segue um processo conforme descrito nesta seção.

Figura 8 - Fluxo principal de informação do método PICTOREA



Fonte: Elaborado pelo autor

4.2.1 Exploração do Espaço Problema

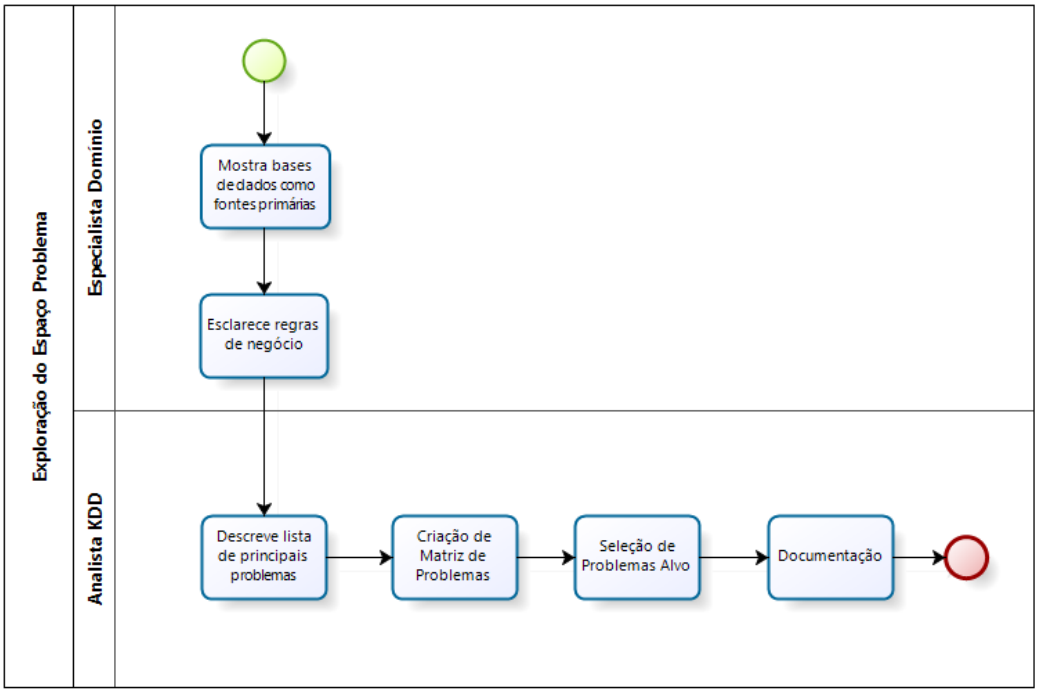
É necessário um entendimento acerca dos domínios de problema que definem o espaço problema pelo Analista KDD. O conhecimento sobre um domínio, bem como as regras de negócio envolvidas, são transmitidos ao Analista KDD pelo Especialista de Domínio.

A participação do especialista de domínio é imprescindível para a definição e listagem dos possíveis problemas. O especialista de domínio contribui atribuindo peso aos possíveis problemas, visto que a resposta do processo para determinados problemas pode gerar mais valor para os negócios do outras respostas.

Segundo Pyle (1999), uma matriz de problemas, *pairwise*, deve conter os campos, problema, importância, dificuldade, retorno. Cada problema deverá ser inserido na matriz com o valor de ponderação por exemplo entre 0 a 5 para os campos Importância, Dificuldade e Retorno. O resultado para o problema de mais alta importância será obtido pela multiplicação das colunas mais a soma. A multiplicação é definida da seguinte forma: para a coluna Importância é definido por padrão um peso de 0.5 e para as colunas Dificuldade e Retorno é definido um peso de 0.25. Esses valores devem ser multiplicados pelo valor da célula correspondente.

Como resultado desta etapa, é definido o domínio de problema a ser investigado.

Figura 9 - Exploração do espaço problema

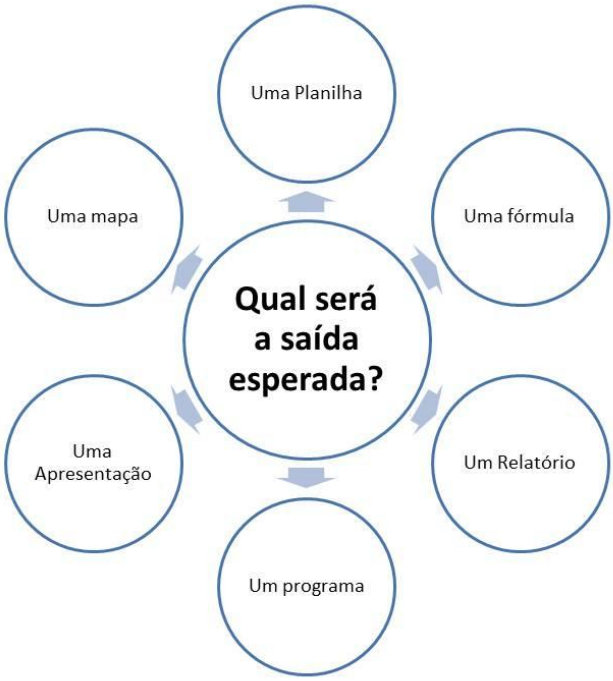


Fonte: Elaborado pelo autor

4.2.2 Exploração do Espaço Solução

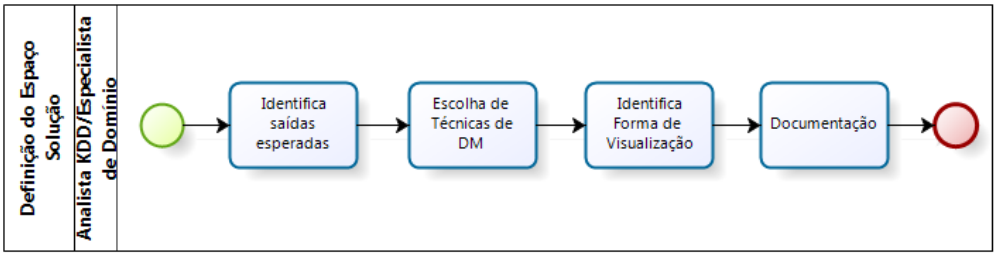
Após a definição e priorização dos problemas levantados, o Analista KDD, com o auxílio do Especialista de Domínio escolhido, definirá as expectativas sobre o resultado e as saídas esperadas. Dessa forma, deverão ser definidas as técnicas mineração de dados e visualização de forma que sejam claras e que correspondam às expectativas de informações do Especialista de Domínio. Para cada problema e sua específica tarefa de Mineração de Dados, haverá um tipo de visualização que será definida na etapa de Visualização. Como mostra a Figura 10, cada domínio de problema pode ter uma visualização ou saída esperada pelo especialista de domínio, dentre elas um relatório, um gráfico, um programa, uma lista simples de dados, uma fórmula estatística, dentre outras.

Figura 10 - Escolha das saídas esperadas de um processo KDD



Fonte: Adaptado de Pyle (1999)

Figura 11 - Definição do espaço solução



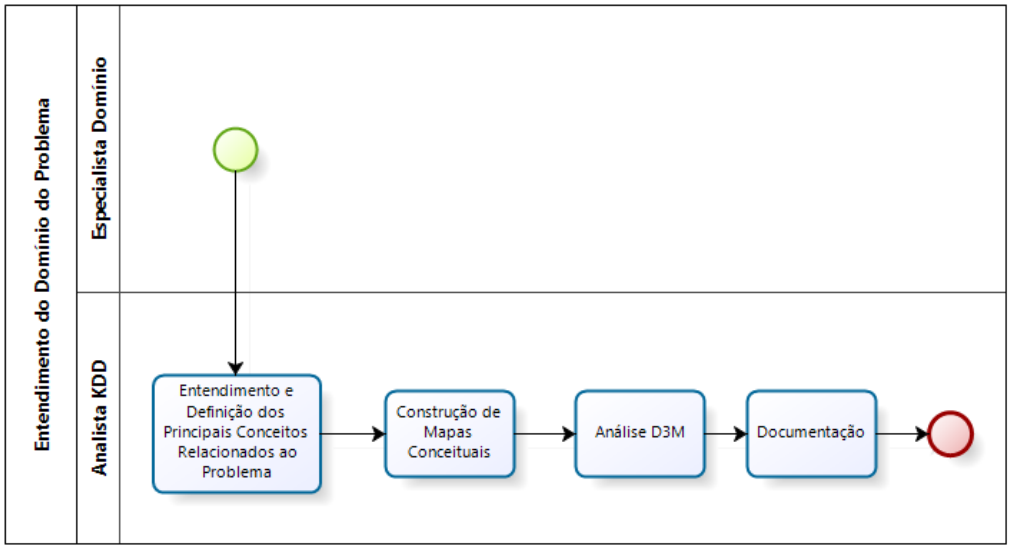
Fonte: Elaborado pelo autor

4.2.3 Entendimento do domínio do problema

Nesta etapa o Analista KDD deverá conhecer o domínio do problema e caracterizá-lo utilizando Mapas Conceituais e sob a orientação do Especialista do Domínio. A utilização de Mapas Conceituais para representação oferece uma visão clara dos conceitos que envolvem o domínio de problema. Mapa conceitual é uma imagem física do que é percebido como objeto em um domínio do problema, juntamente com suas interconexões e interações entre variáveis destes objetos.

Outras técnicas que podem ser empregadas são Análise Convergente/Divergente, Pró-Contra-e-fixação e mapas cognitivos. (PYLE, 1999).

Figura 12 - Entendimento do domínio do problema



Fonte: Elaborado pelo autor

4.2.4 Caracterização do problema através de atributos

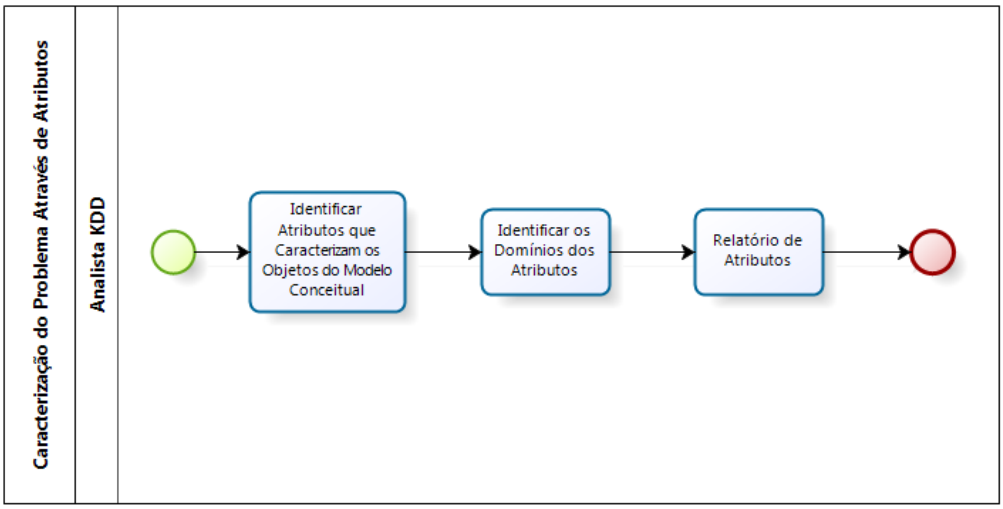
Após o entendimento do domínio do problema, das expectativas e resultados esperados, é necessário identificar os atributos relevantes para a descoberta de conhecimento.

Cada atributo será avaliado pelo analista KDD e selecionado de acordo com a sua relevância em relação ao problema definido. Será documentado cada atributo, seu tipo, faixa de valor e a sua relevância em relação ao domínio do problema.

Na etapa anterior, através de um mapa conceitual, caracteriza-se o domínio de problema identificando os objetos que o representam. Nesta etapa, o principal desafio é identificar quais serão os atributos que representam e definem os objetos identificados na etapa anterior. Por exemplo, podemos definir para um objeto identificado como “automóvel”, os seguintes atributos: fabricante, cor, número de portas e capacidade de passageiros.

Isto será útil para a montagem do banco de dados sobre a qual será extraído conhecimento.

Figura 13 - Caracterização do problema através de atributos



Fonte: Elaborado pelo autor

4.2.5 Montagem do banco de dados

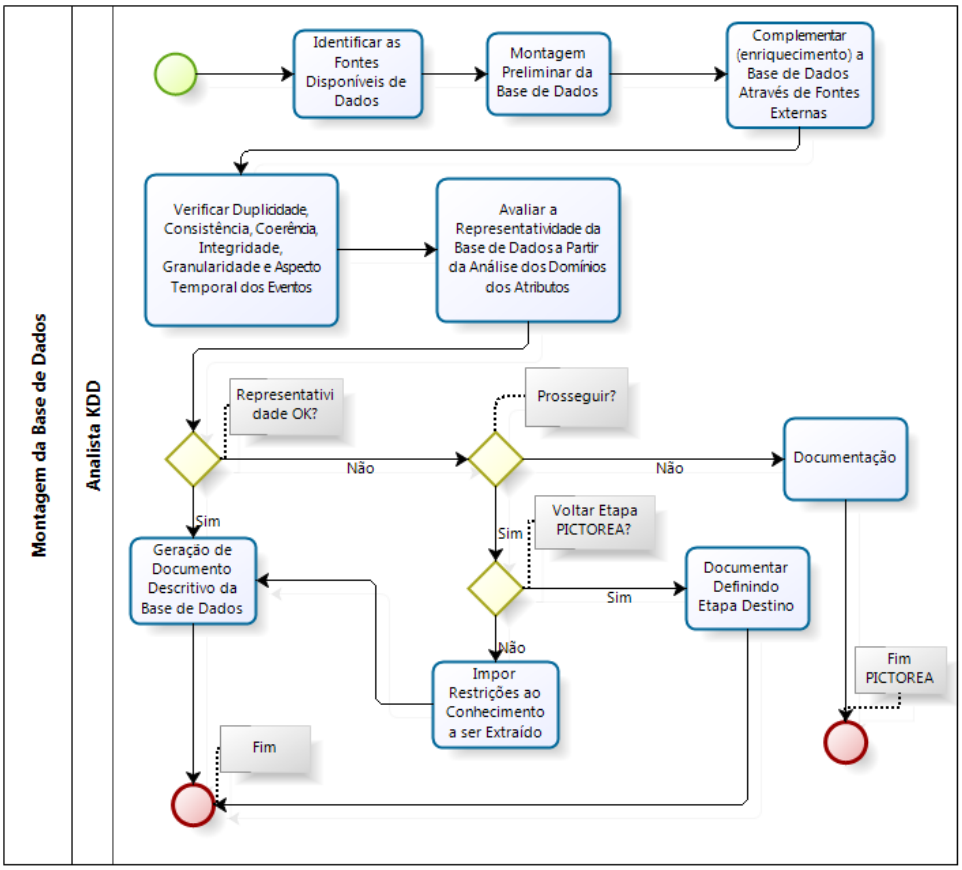
Com os atributos selecionados na etapa anterior será criado um banco de dados para o projeto. O valor de alguns atributos poderá ser combinado entre valores de outros atributos. Além disso, há atributos que podem conter informações similares para determinado problema. Nesse caso, o Analista KDD deverá definir uma estratégia para lidar com esse tipo de atributo, como por exemplo remove-lo ou combiná-lo com outra variável. (PYLE, 1999).

Cria-se então uma base de dados preliminar. É comum o Analista KDD identificar a necessidade de buscar em fontes de dados externas os atributos identificados como essenciais durante o entendimento e caracterização do problema. Além disso, pode ser identificada a necessidade de se aumentar a granularidade de algum atributo (Melhoramento de Atributos). Deverão ser verificadas a consistência e coerência dos atributos, das instâncias, a presença de poluição nos dados, integridade e duplicidade de instâncias.

Ao final desta etapa, é feita uma avaliação da representatividade do banco de dados criado a partir da análise dos domínios dos atributos. Entendemos por representatividade termos dados suficientes para descrever o domínio do problema. Caso o banco de dados criado não seja representativa o suficiente para a

descoberta de conhecimento, mas o Analista de KDD decidir prosseguir, pode-se voltar a alguma etapa anterior do método ou impor restrições ao conhecimento a ser extraído. Caso o Analista KDD opte por não prosseguir, os motivos são documentados e o processo de descoberta de conhecimento é finalizado.

Figura 14 - Montagem do banco de dados



Fonte: Elaborado pelo autor

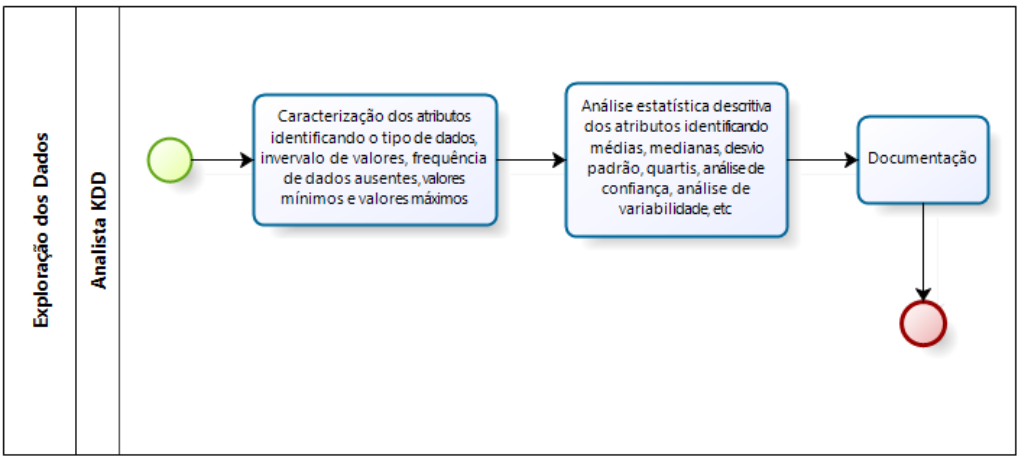
4.2.6 Exploração dos dados

Nesta etapa será feita a caracterização dos atributos identificando o tipo de dado, intervalo de valores, frequência de dados ausentes, valores mínimos e valores máximos. O Analista KDD precisa conhecer e explorar os dados que serão utilizados na descoberta de conhecimento.

Será feita uma análise estatística descritiva dos atributos com o intuito de identificar médias, medianas, desvio padrão, quartis, análise de confiança e análise de variabilidade.

Ao final da etapa, tem-se um documento descritivo da exploração dos dados.

Figura 15 - Exploração dos dados



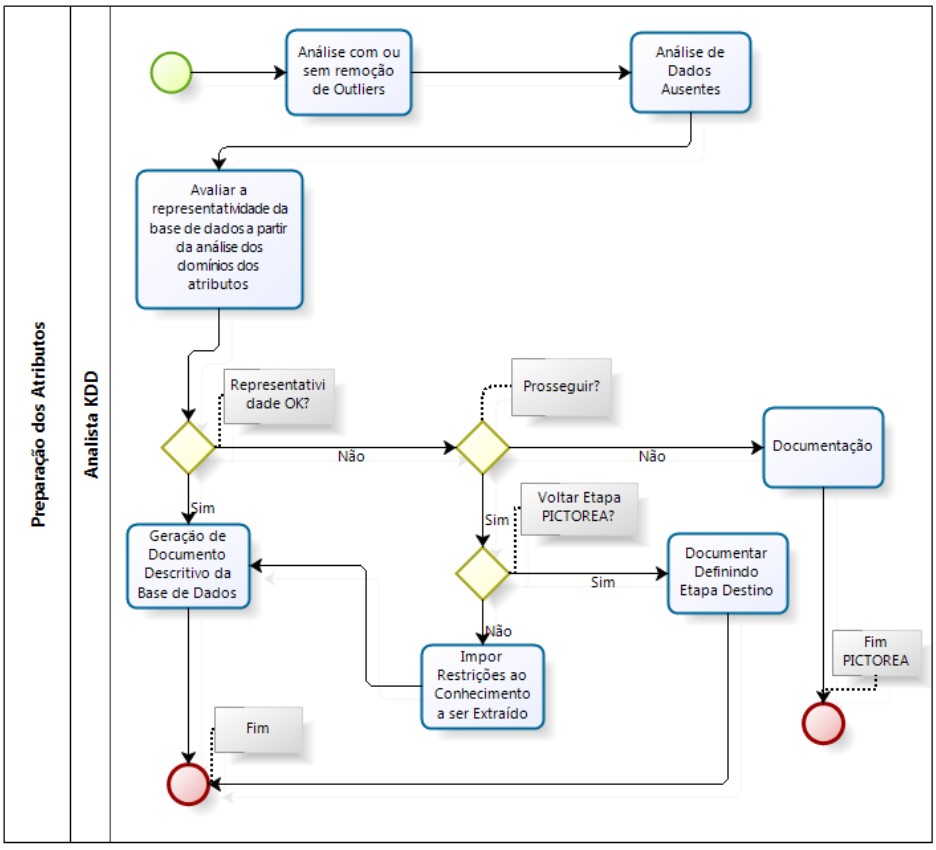
Fonte: Elaborado pelo autor

4.2.7 Preparação dos atributos

Nesta etapa será feita a análise de *Outliers*. *Outliers* são dados com comportamento muito diferente dos demais. Dados “discrepantes”, que fogem ao padrão da base de dados. Estes dados precisam ser analisados, pois possivelmente tratam-se de erros na base de dados. (PYLE, 1999).

Será feita também uma análise de dados ausentes com o intuito de verificar o impacto que esta ocorrência terá na descoberta de conhecimento. Cabe ao analista KDD definir uma estratégia para tratar *outliers* e dados ausentes. Nesta etapa ainda, será feita uma análise de representatividade dos dados a partir da análise dos domínios dos atributos. Se a base de dados não for representativa, o analista KDD poderá decidir entre voltar em qualquer etapa do processo ou impor restrições ao conhecimento a ser extraído.

Figura 16 - Preparação dos atributos



Fonte: Elaborado pelo autor

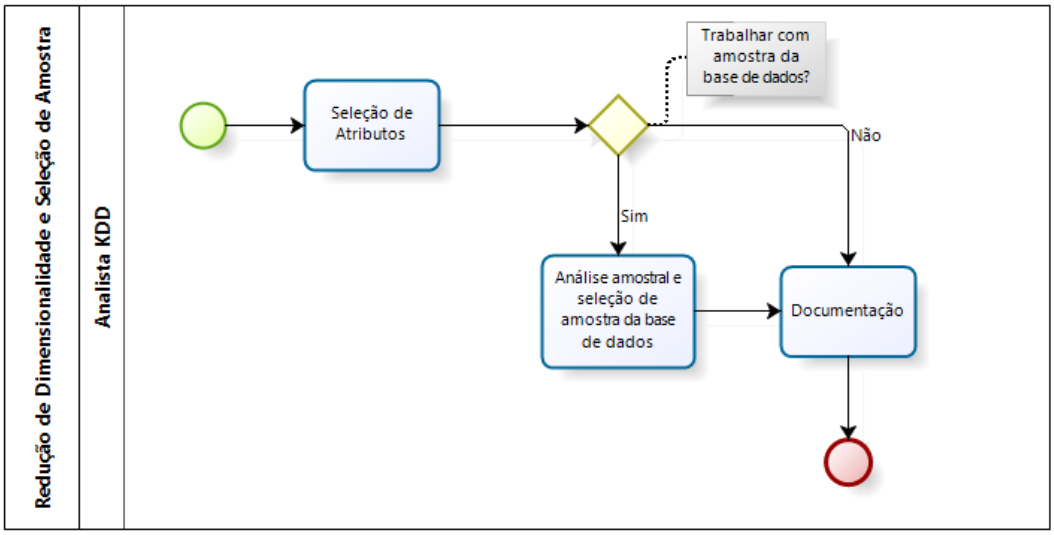
4.2.8 Redução da dimensionalidade e seleção de amostra

Nesta etapa serão avaliados os atributos pelo conceito de entropia.

O analista KDD poderá aplicar técnicas de redução de dimensionalidade como análise de componentes principais, análise de correlação entre atributos, análise de dependência de atributos e análise de compressão de dimensionalidade de atributos. (PYLE, 1999).

Ainda nesta etapa, o Analista KDD deverá decidir se trabalhará com uma amostra da base de dados. Caso faça essa opção, será feita uma análise e seleção da amostra da base de dados. Ao final da etapa as análises e decisões tomadas serão documentadas.

Figura 17 - Redução da dimensionalidade e seleção de amostra



Fonte: Elaborado pelo autor

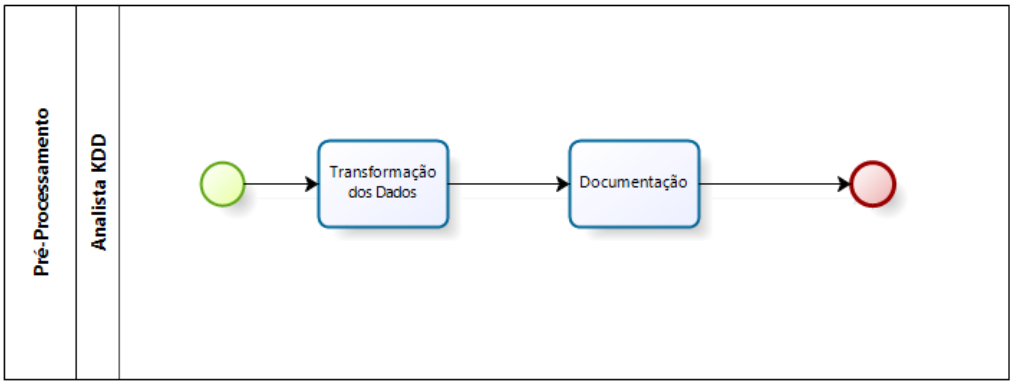
4.2.9 Pré-Processamento

Os dados precisam ser transformados de forma que possam servir de entrada para os algoritmos de Mineração de Dados, na etapa seguinte. Normalmente os algoritmos de Mineração de Dados aceitam somente quantidades numéricas. Dessa forma, se faz necessária a transformação dos dados ou mudanças de escala sobre os dados sem perder as características do valor original.

Segundo Pyle (1999), a melhor forma de definir o pré-processamento é verificar quais requisitos a solução precisa atender e quais são os requisitos que a técnica de mineração de dados precisa.

Ao final da etapa, tem-se um documento descritivo.

Figura 18 - Pré-Processamento



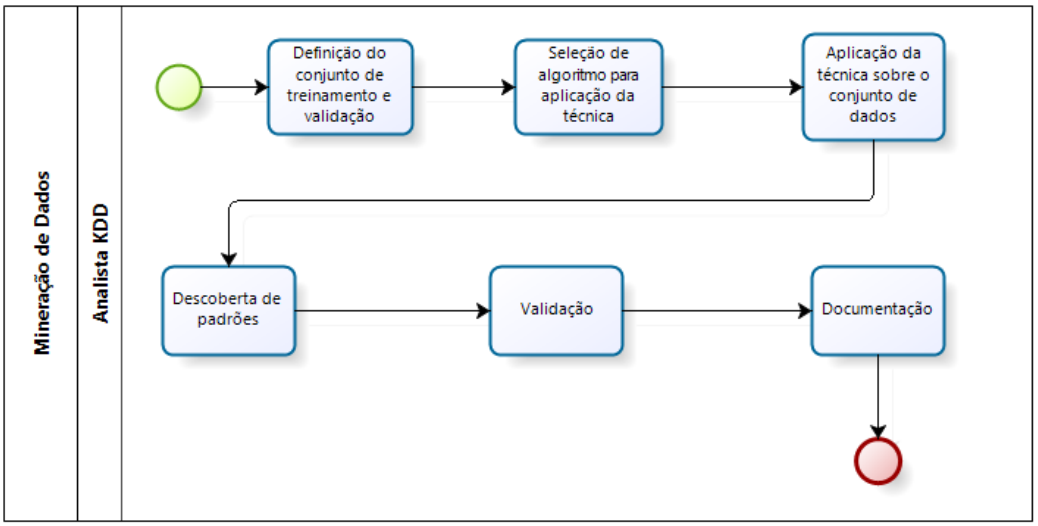
Fonte: Elaborado pelo autor

4.2.10 Mineração de dados

O analista KDD deverá definir o conjunto de dados de treinamento e validação. Deverá identificar a técnica apropriada para a extração de conhecimentos que atendam as expectativas documentadas. A escolha da técnica de mineração de dados depende da análise das necessidades do especialista de domínio juntamente com as características da técnica. Algumas técnicas têm melhor performance e melhor resposta quando utilizadas sobre dados categóricos, ao passo que outras têm melhor resultado com dados contínuos.

Como mencionado no capítulo 2, as ferramentas para mineração de dados existentes no mercado, implementam variações de algoritmos “clássicos” que executam técnicas de mineração de dados. Por isso, não fixaremos aqui um software específico para execução de uma técnica de mineração de dados. Definida a técnica, esta deverá ser aplicada em dados de amostra para posteriormente ser aplicada na base de dados.

Figura 19 - Mineração de dados

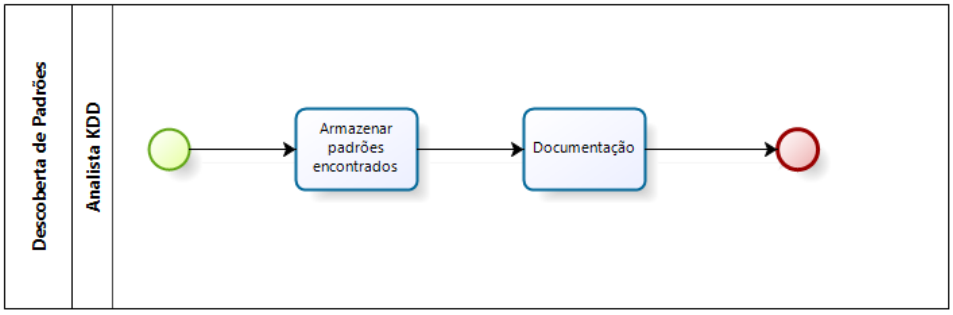


Fonte: Elaborado pelo autor

4.2.11 Descoberta de padrões

Após a aplicação dos algoritmos de Mineração de Dados, os padrões encontrados deverão ser armazenados para a análise da qualidade do conhecimento.

Figura 20 - Descoberta de padrões



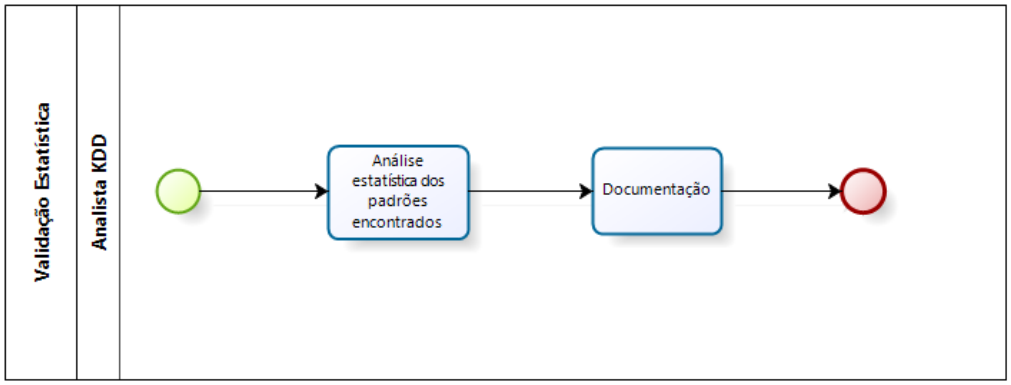
Fonte: Elaborado pelo autor

4.2.12 Validação estatística

Cross-Validation é um método estatístico geralmente utilizado para validação e comparação de algoritmos de aprendizado de máquina. Basicamente os dados são divididos em dois grupos: um usado para dados de treinamento e outro usado para validação.

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.* (REFAEILZADEH; TANG; LIU, 2009).

Figura 21 - Validação estatística



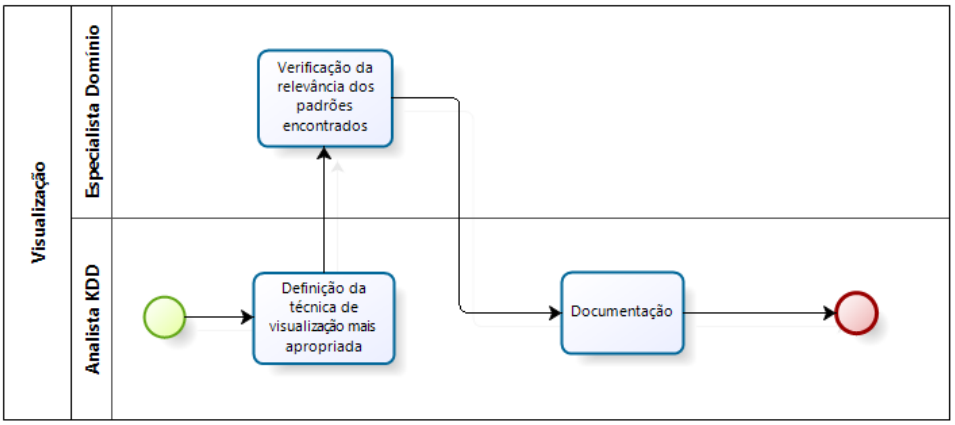
Fonte: Elaborado pelo autor

4.2.13 Visualização

De acordo com a técnica de Mineração de Dados utilizada, o Analista KDD deverá selecionar a técnica de visualização mais apropriada para a exibição do padrão encontrado.

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 Analista KDD deverá verificar os resultados de cada etapa e poderá refazer o processo a partir da etapa que decidir.

Figura 22 - Visualização



Fonte: Elaborado pelo autor

4.3 Método PICTOREA formalizado através do SPEM

Foi desenvolvido um protótipo (MONTEVECCHI, 2011), para descrição das etapas do método.

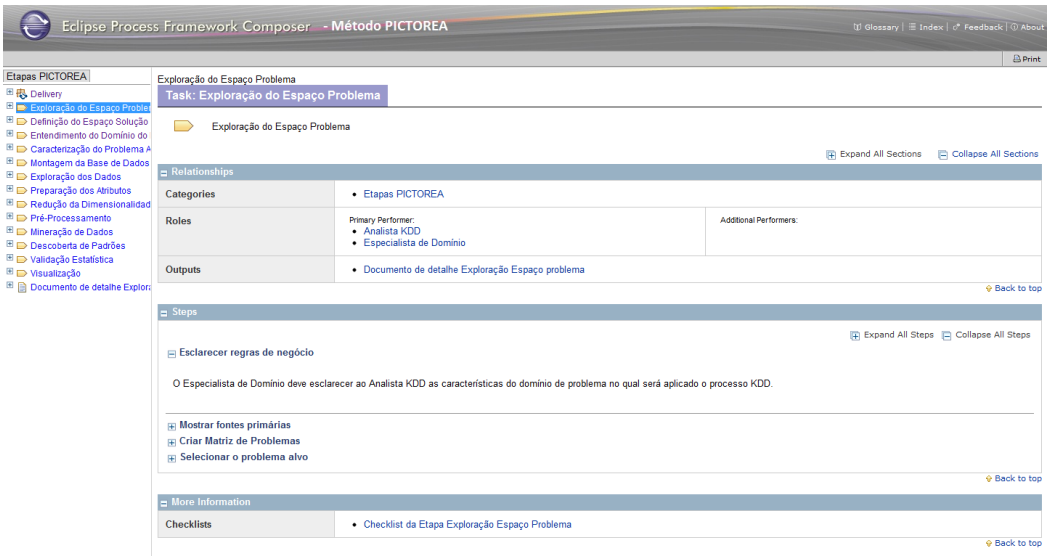
Conforme visto anteriormente, os processos precisam ser formalizados ou padronizados de forma a entendermos o seu funcionamento. Dessa forma, consegue-se melhor treinamento e possibilidade de melhorias contínuas. (GENVIGIR; SANTANNA; *et al.*, 2003).

Utilizando o SPEM, neste trabalho foi formalizado o método utilizando a ferramenta *Eclipse Process Framework Composer*, foi possível detalhar cada etapa, com seus pré-requisitos e saídas esperadas, bem como seus autores.

Em cada etapa do método PICTOREA, demonstrado na Figura 8, tem-se uma sequência de passos, seus atores, suas saídas e suas validações.

Na Figura 23 verifica-se a representação da etapa de Exploração do Espaço Problema do método PICTOREA. Estão formalizados os papéis (*Roles*), a saída da etapa (*Outputs*), os passos da etapa (*Steps*) e o *Checklist* criado para validação da etapa.

Figura 23 - Página do Método PICTOREA - Exploração do Espaço Problema

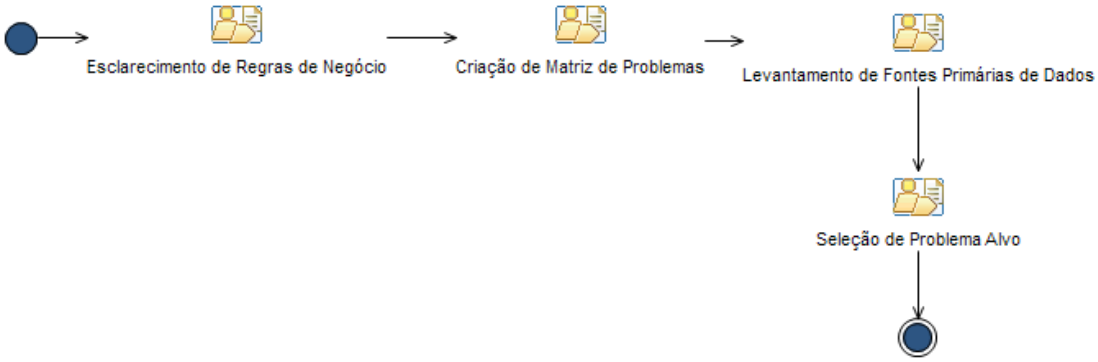


Fonte: Montevecchi, 2011

Conseguimos assim, com que as informações sejam compartilhadas entre uma equipe de trabalho, de forma que seus membros possam consultar informações do método PICTOREA e saber o seu papel no método.

A sequência para cada etapa está detalhada na formalização gerada, conforme exemplificado na Figura 24 que representa as atividades da etapa de Exploração do Espaço Problema.

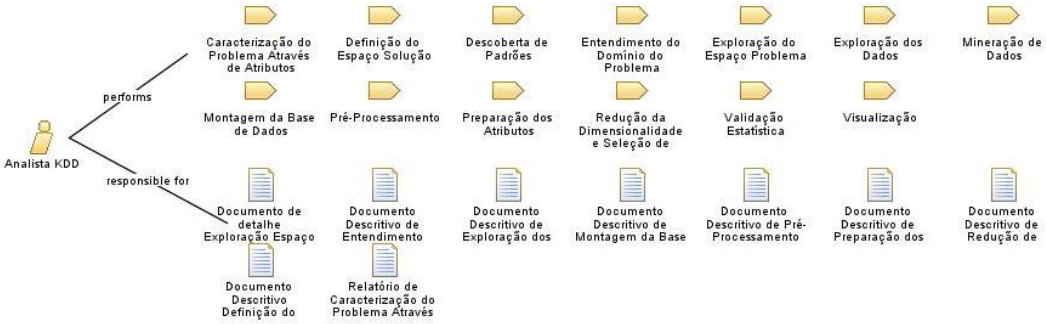
Figura 24 - Diagrama de Sequencia de Atividades da Etapa de Exploração do Espaço Problema



Fonte: Montevecchi, 2011

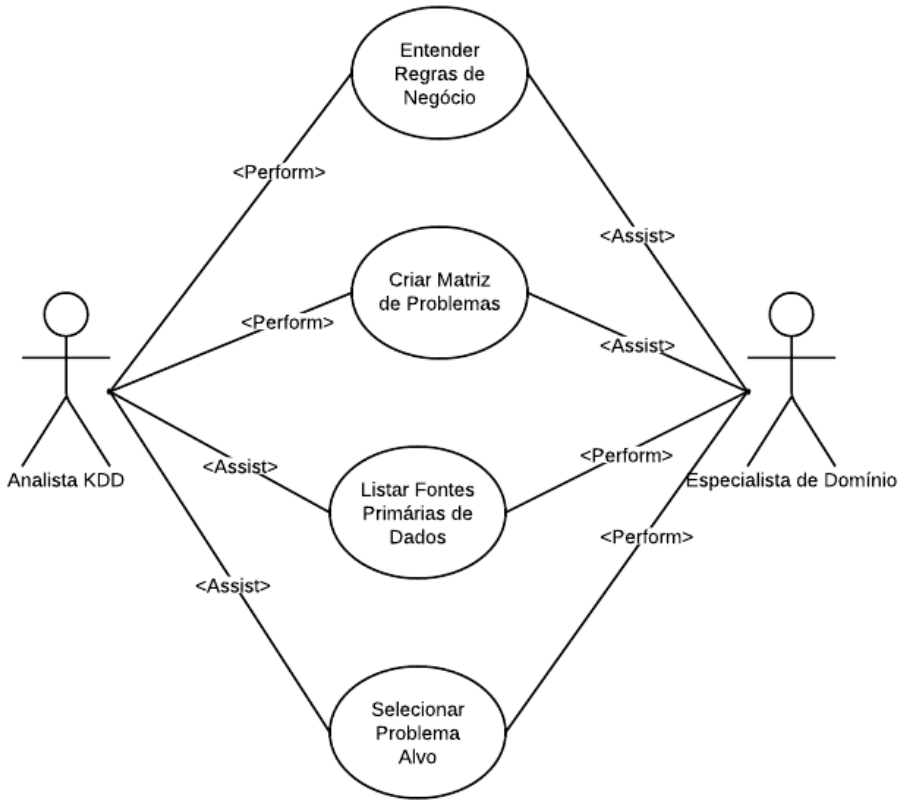
Cada papel possui sua participação nas etapas do método PICTOREA e os documentos pelos quais é responsável. Na Figura 25 observa-se as atividades e responsabilidades do Analista KDD dentro do método PICTOREA.

Figura 25 - Diagrama de Atividades do Papel Analista KDD



Fonte: Montevecchi, 2011

Figura 26 - Diagrama de Caso de Uso da Etapa de Exploração do Espaço Problema



Fonte: Elaborado pelo autor

Na Figura 27 observa-se as atividades e responsabilidades do Especialista de Domínio dentro do método PICTOREA.

Figura 27 - Diagrama de Atividades do Papel Especialista de Domínio

Fonte: Montevecchi, 2011

Na Figura 28, vemos a formalização e detalhamento do documento a ser gerado como artefato da etapa de Exploração do Espaço Problema.

Figura 28 - Página de formalização da saída da etapa de Exploração do Espaço Problema

Fonte: Montevecchi, 2011

Ainda analisando a Figura 28, encontramos um link para o modelo de documento (*template*) para registro das informações da etapa. Este documento encontra-se no apêndice A deste trabalho.

Dentre os benefícios do SPEM para a construção do Método PICTOREA podemos citar:

a) Acessibilidade via navegador Web.
b) Software livre.
c) Extensa documentação disponível.
d) Ferramenta intuitiva.

5 AVALIAÇÃO EXPERIMENTAL

5.1 Considerações iniciais

O método PICTOREA foi avaliado através de dois estudos de caso aplicados em organizações de médio-grande porte. Ambas as empresas solicitaram sigilo quanto ao seu nome, pois no método são levantadas e discutidas questões estratégicas. Os estudos de caso trazem informações reais, estratégicas, porém sem mencionar o nome das organizações.

A avaliação experimental está limitada às etapas de Exploração do Espaço Problema e Definição do Espaço Solução.

5.2 Aplicações

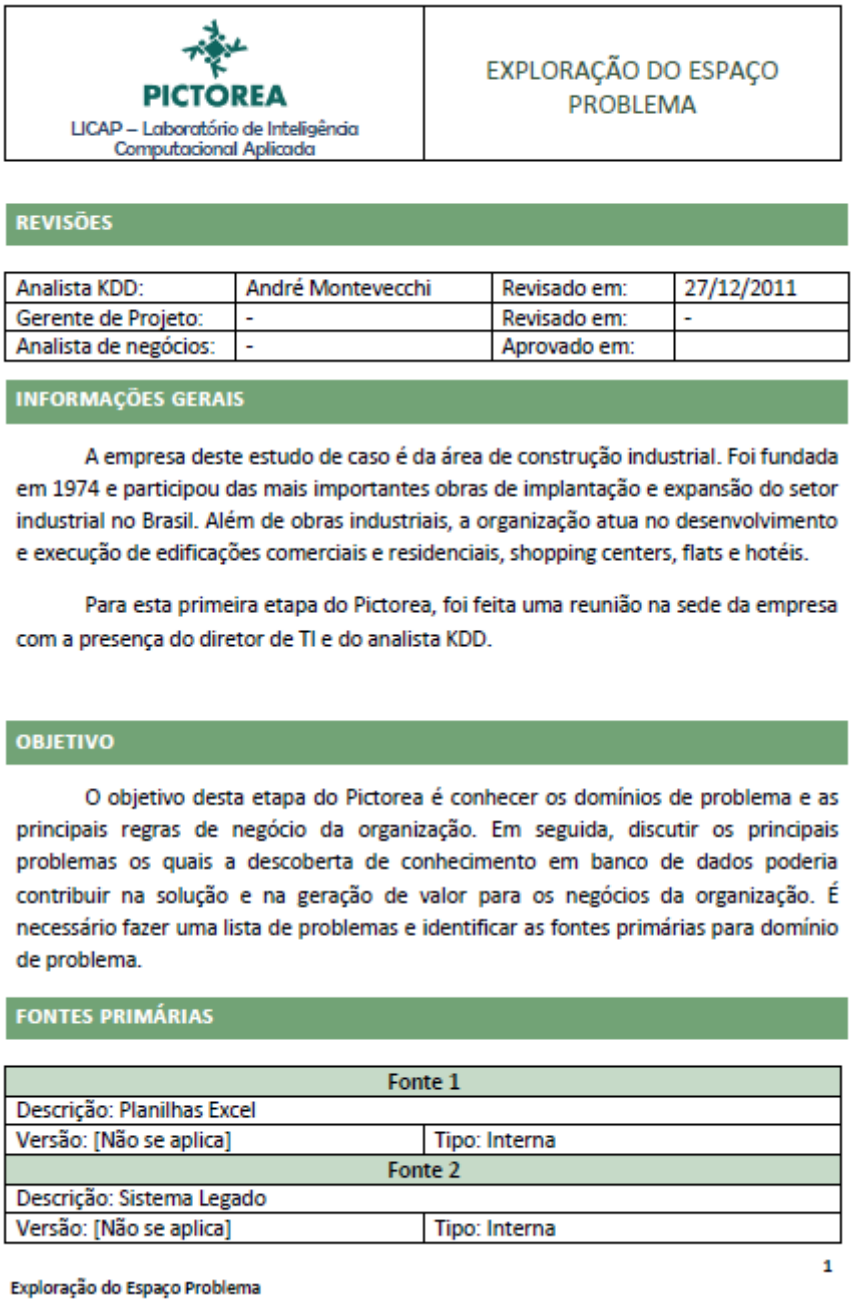
5.2.1 Empresa do ramo de construção industrial

O primeiro estudo de caso foi aplicado em uma organização da área de construção industrial. A organização foi fundada em 1974 e participou de importantes obras de implantação e expansão do setor industrial no Brasil. Além de obras industriais, ela atua no desenvolvimento e execução de edificações comerciais e residenciais, shopping centers, flats e hotéis.

5.2.1.1 Exploração do espaço problema

Através das Figuras 29 e 30, está representado o documento gerado na etapa de Exploração do Espaço Problema.

Figura 29 - Documento de Exploração do Espaço Problema - Página 1



PICTOREA
LICAP – Laboratório de Inteligência Computacional Aplicada

EXPLORAÇÃO DO ESPAÇO PROBLEMA

REVISÕES

Analista KDD:	André Montevecchi	Revisado em:	27/12/2011
Gerente de Projeto:	-	Revisado em:	-
Analista de negócios:	-	Aprovado em:	

INFORMAÇÕES GERAIS

A empresa deste estudo de caso é da área de construção industrial. Foi fundada em 1974 e participou das mais importantes obras de implantação e expansão do setor industrial no Brasil. Além de obras industriais, a organização atua no desenvolvimento e execução de edificações comerciais e residenciais, shopping centers, flats e hotéis.

Para esta primeira etapa do Pictorea, foi feita uma reunião na sede da empresa com a presença do diretor de TI e do analista KDD.

OBJETIVO

O objetivo desta etapa do Pictorea é conhecer os domínios de problema e as principais regras de negócio da organização. Em seguida, discutir os principais problemas os quais a descoberta de conhecimento em banco de dados poderia contribuir na solução e na geração de valor para os negócios da organização. É necessário fazer uma lista de problemas e identificar as fontes primárias para domínio de problema.

FONTES PRIMÁRIAS

Fonte 1	
Descrição: Planilhas Excel	
Versão: [Não se aplica]	Tipo: Interna
Fonte 2	
Descrição: Sistema Legado	
Versão: [Não se aplica]	Tipo: Interna

1

Exploração do Espaço Problema

Fonte: Elaborado pelo autor

Figura 30 - Documento de Exploração do Espaço Problema - Página 2

PICTOREA
LICAP – Laboratório de Inteligência Computacional Aplicada

EXPLORAÇÃO DO ESPAÇO PROBLEMA

MATRIZ DE PROBLEMAS (PAIRWIZE)

Problema	Importância	Facilidade	Retorno	Total
Assertividade entre orçamento e execução	3	0	3	2,25
Desempenho de fornecedor em diferentes obras	0	3	0	0,75
Conhecer perfil de investidores	1	2	1	1,25
Identificação de causas de acidentes de trabalho	2	1	2	1, 5

RANKING

Abaixo estão relacionados os problemas identificados ordenados por pontuação no *pairwise*.

1. Assertividade entre orçamento e execução
2. Desempenho de fornecedor em diferentes obras
3. Conhecer perfil de investidores
4. Identificação de causas de acidentes de trabalho

CONSIDERAÇÕES OU RESTRIÇÕES ADICIONAIS

Foi identificado como melhoria desta primeira etapa do método Pictorea, a prévia apresentação sobre as possibilidades de descoberta de conhecimento em bases de dados e sobre as técnicas de mineração de dados.

2

Exploração do Espaço Problema

Fonte: Elaborado pelo autor

Como mencionado anteriormente, na etapa de Exploração do Espaço Problema busca-se conhecer os domínios de problema, bem como as regras de negócio da organização. Em seguida, deve-se atribuir valor e peso aos domínios para que seja escolhido um alvo para a descoberta de conhecimento. Esta etapa compreende também a identificação e documentação das fontes primárias dos domínios de problema.

Em reunião com o diretor de TI da organização, foram identificados os seguintes domínios de problema:

a) discrepância entre orçamento e execução.
b) desempenho de fornecedor em diferentes obras.
c) conhecer perfil de investidores.
d) identificação de causas de acidentes de trabalho.

Após identificação dos domínios de problema, foi utilizada a técnicas de *pairwise* (PYLE, 1999) para criação de um *ranking* com o intuito de identificar o domínio de problema alvo de acordo com sua importância. O resultado do *pairwise* está representado na Tabela 1.

Segundo Pyle (1999), na análise de *pairwise*, a coluna total deve ser calculada usando a seguinte métrica:

Total = (Importância * 0,50) + (Facilidade * 0,25) + (Retorno * 0,25)

Observa-se que a coluna intitulada “Importância” deve possuir peso maior por ter uma relevância maior para as estratégias de negócios da organização.

Tabela 1 - Pairwise do Estudo de Caso 1 - Construção Industrial

Problema	Importância	Facilidade	Retorno	Total
Discrepância entre orçamento e execução	3	0	3	**2,25**
Desempenho de fornecedor em diferentes obras	0	3	0	**0,75**
Conhecer o perfil de investidores	1	2	1	**1,25**
Identificação de causas de acidentes de trabalho	2	1	2	**1,5**

Fonte: Elaborado pelo autor

Através da análise de *pairwise* representada da Tabela 1, entendemos que o domínio do problema alvo foi **Discrepância entre orçamento e execução**. Este domínio de problema possui maior importância para os negócios da organização e maior retorno, porém, como mostra a coluna intitulada “Facilidade”, não é um problema simples de ser implementado no método PICTOREA.

5.2.1.2 Definição do espaço solução

Através das Figuras 31 e 32 está representado o documento gerado na etapa de Definição do Espaço Solução.

Figura 31 - Documento de Definição do Espaço Solução - Página 1

PICTOREA LICAP – Laboratório de Inteligência Computacional Aplicada	DEFINIÇÃO DO ESPAÇO SOLUÇÃO

REVISÕES

Analista KDD:	André Montevecchi	Revisado em:	27/12/2011
Gerente de Projeto:		Revisado em:	
Analista de negócios:		Aprovado em:	

INFORMAÇÕES GERAIS

A empresa deste estudo de caso é da área de construção industrial. Foi fundada em 1974 e participou das mais importantes obras de implantação e expansão do setor industrial no Brasil. Além de obras industriais, a organização atua no desenvolvimento e execução de edificações comerciais e residenciais, shopping centers, flats e hotéis.

Para esta primeira etapa do Pictorea, foi feita uma reunião na sede da empresa com a presença do diretor de TI e do analista KDD.

OBJETIVO

O objetivo desta etapa do Pictorea é definir as expectativas sobre o resultado e as saídas esperadas. Em seguida, deverão ser definidas as técnicas de mineração de dados, e a forma de visualização dos resultados.

SAÍDAS ESPERADAS

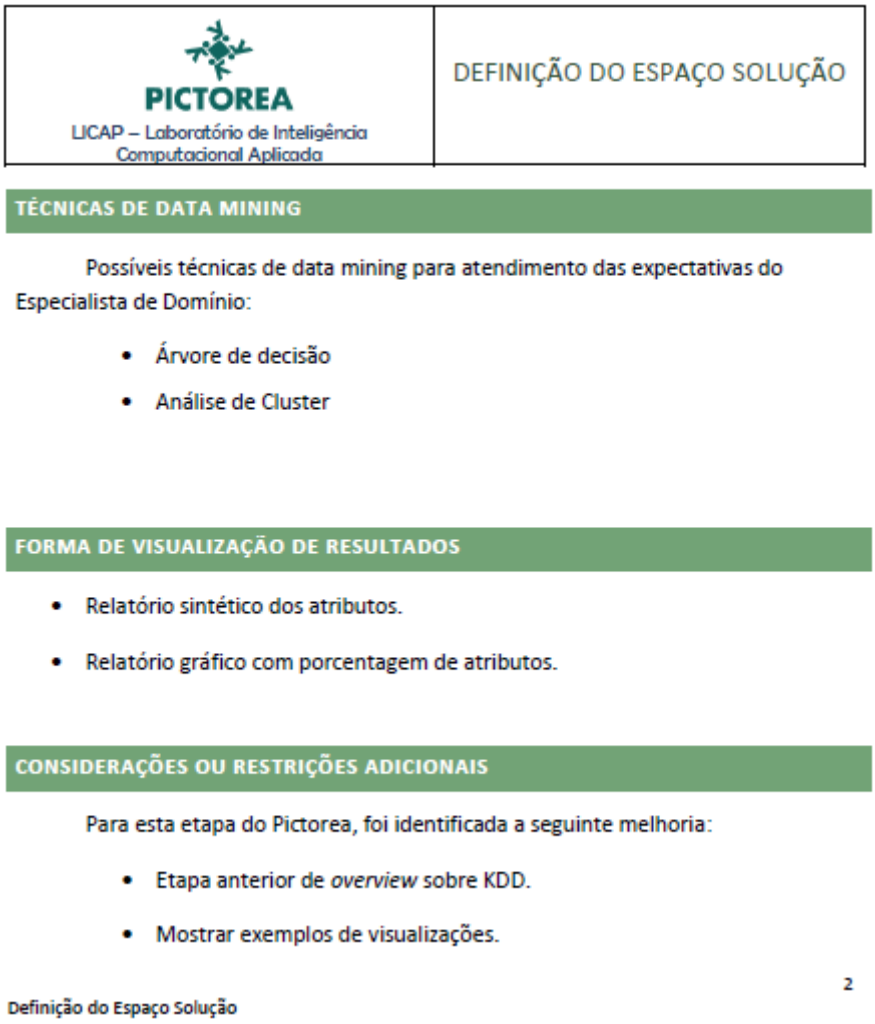
O Especialista de Domínio espera como saída um relatório sintético com os tipos de itens/categorias que podem influenciar na diferença que ocorre entre orçado e executado.

Exemplos de categorias:

- Mão de obra direta
- Mão de obra indireta
- Equipamentos
- Despesas indiretas
- Subempreiteiros

Fonte: Elaborado pelo autor

Figura 32 - Documento de Definição do Espaço Solução - Página 2



PICTOREA
LICAP – Laboratório de Inteligência Computacional Aplicada

DEFINIÇÃO DO ESPAÇO SOLUÇÃO

TÉCNICAS DE DATA MINING

Possíveis técnicas de data mining para atendimento das expectativas do Especialista de Domínio:

- Árvore de decisão
- Análise de Cluster

FORMA DE VISUALIZAÇÃO DE RESULTADOS

- Relatório sintético dos atributos.
- Relatório gráfico com porcentagem de atributos.

CONSIDERAÇÕES OU RESTRIÇÕES ADICIONAIS

Para esta etapa do Pictorea, foi identificada a seguinte melhoria:

- Etapa anterior de *overview* sobre KDD.
- Mostrar exemplos de visualizações.

2

Definição do Espaço Solução

Fonte: Elaborado pelo autor

Como já mencionado, os objetivos da etapa de Definição do Espaço Solução são definir as expectativas sobre o resultado e as saídas esperadas.

Conforme explicado pelo Diretor de TI da organização, ele espera que seja exibido um relatório sintético com os tipos de itens/categorias que podem de alguma forma influenciar na ocorrência de discrepância entre o orçado e o executado.

De acordo com as expectativas sobre o resultado, entendemos que as técnicas de mineração de dados que mais se adequam são: árvore de decisão e análise de cluster.

5.2.1.3 Melhorias identificadas do método PICTOREA

Na aplicação da etapa de Definição do Espaço Solução do Método PICTOREA, identificou-se a necessidade de se ter uma introdução à mineração de dados e exemplos de visualizações de resultados.

5.2.2 Empresa do ramo da Tecnologia da Informação

O segundo estudo de caso foi aplicado em uma organização da área de Tecnologia da Informação. A organização tem aproximadamente 16 anos de atuação no mercado atendendo clientes de médio a grande porte. Em sua sede em Belo Horizonte estão alocados aproximadamente 150 colaboradores. Além de uma fábrica de software, a organização possui uma área de treinamentos oficiais Microsoft e uma área de consultoria, sendo esta, nas áreas de Business Intelligence - BI, gerenciamento de projetos, infraestrutura de TI e produtos Microsoft.

5.2.2.1 Exploração do espaço problema

Através das Figuras 33, 34 e 35, está representado o documento gerado na etapa de Exploração do Espaço Problema.

Figura 33 - Documento de Exploração do Espaço Problema - Página 1

PICTOREA LICAP – Laboratório de Inteligência Computacional Aplicada	EXPLORAÇÃO DO ESPAÇO PROBLEMA

REVISÕES

Analista KDD:	André Montevecchi	Revisado em:	20/01/2012
Gerente de Projeto:	-	Revisado em:	-
Analista de negócios:	-	Aprovado em:	

INFORMAÇÕES GERAIS

A empresa deste estudo de caso é da área de TI e tem aproximadamente 16 anos de atuação no mercado. Em sua sede em Belo Horizonte possui aproximadamente 150 colaboradores. Além de uma fábrica de software, a empresa possui uma área de treinamentos oficiais Microsoft e uma área de consultoria, sendo esta última, nas áreas de Business Intelligence – BI, gerenciamento de projetos, infraestrutura e produtos Microsoft.

Para esta primeira etapa do Pictorea, foi feita uma reunião na sede da empresa com a presença do diretor executivo, da gerente de RH e do analista KDD.

OBJETIVO

O objetivo desta etapa do Pictorea é conhecer os domínios de problema e as principais regras de negócio da organização. Em seguida, discutir os principais problemas os quais a descoberta de conhecimento em banco de dados poderia contribuir na solução e na geração de valor para os negócios da organização. É necessário fazer uma lista de problemas e identificar as fontes primárias para domínio de problema.

FONTES PRIMÁRIAS

Fonte 1	
Descrição: Exames técnicos	
Versão: [Não se aplica]	Tipo: Interna
Fonte 2	
Descrição: Exames de conhecimento	
Versão: [Não se aplica]	Tipo: Interna

Fonte: Elaborado pelo autor

Figura 34 - Documento de Exploração do Espaço Problema - Página 2

PICTOREA LICAP – Laboratório de Inteligência Computacional Aplicada	EXPLORAÇÃO DO ESPAÇO PROBLEMA

Fonte 3	
Descrição: Exames psicológicos	
Versão: [Não se aplica]	Tipo: Interna
Fonte 4	
Descrição: Documentação de entrevista	
Versão: [Não se aplica]	Tipo: Interna
Fonte 5	
Descrição: Exames de conhecimento de segundo idioma	
Versão: [Não se aplica]	Tipo: Interna
Fonte 6	
Descrição: Entrevista de desligamento	
Versão: [Não se aplica]	Tipo: Interna
Fonte 7	
Descrição: Base de dados do Sistema de RH	
Versão: [Não se aplica]	Tipo: Interna
Fonte 8	
Descrição: Base de dados do Project Server	
Versão: [Não se aplica]	Tipo: Interna
Fonte 9	
Descrição: Base de dados com relação de clientes que compraram Office 365	
Versão: [Não se aplica]	Tipo: Externa
Fonte 10	
Descrição: Entrevista de desligamento	
Versão: [Não se aplica]	Tipo: Interna
Fonte 11	
Descrição: Resultados de exames para cargo de analista de sistema	
Versão: [Não se aplica]	Tipo: Interna

MATRIZ DE PROBLEMAS (PAIRWIZE)

Problema	Importância	Facilidade	Retorno	Total
Assertividade na contratação de colaboradores	3	2	3	2,75
Retenção de colaboradores	4	1	2	2,75
Perfil de clientes potenciais para compra de Office 365	1	4	2	2
Assertividade na estimativa de projetos	2	0	5	2,25

2

Exploração do Espaço Problema

Fonte: Elaborado pelo autor

Figura 35 - Documento de Exploração do Espaço Problema - Página 3

PICTOREA LICAP – Laboratório de Inteligência Computacional Aplicada	EXPLORAÇÃO DO ESPAÇO PROBLEMA

Identificação em colaboradores com cargo desenvolvedor o perfil de analista de sistemas	0	3	0	0,75

RANKING

Abaixo estão relacionado os problemas identificados ordenados por pontuação no *pairwise*.

1. Retenção de colaboradores
2. Assertividade na contratação de colaboradores
3. Assertividade na estimativa de projetos
4. Perfil de clientes potenciais para compra de Office 365
5. Identificação de colaboradores com cargo desenvolvedor o perfil de analista de sistemas

CONSIDERAÇÕES OU RESTRIÇÕES ADICIONAIS

Após discutirmos os problemas na reunião, entendemos que o alvo do Pictorea deve ser Retenção de colaboradores. Foi considerado o mais importante, e, de maneira geral, a solução deste problema gerará grande valor para a organização.

Fonte: Elaborado pelo autor

Como mencionado anteriormente, na etapa de Exploração do Espaço Problema busca-se conhecer os domínios de problema, bem como as regras de negócio da organização. Em seguida, deve-se atribuir valor e peso aos domínios para que seja escolhido um alvo para a descoberta de conhecimento. Esta etapa compreende também a identificação e documentação das fontes primárias dos domínios de problema.

Em reunião com o diretor de TI da organização e do Gerente de Recursos Humanos, foram identificados os seguintes domínios de problema:

a) assertividade na contratação de colaboradores.
b) retenção de colaboradores.
c) perfil de clientes para compra de Office 365.
d) assertividade na estimativa de projetos
e) identificação em colaboradores com cargo desenvolvedor o perfil de analista de sistemas.

Após identificação dos domínios de problema, foi utilizada a técnicas de *pairwise* (PYLE, 1999) para criação de um *ranking* com o intuito de identificar o domínio de problema alvo de acordo com sua importância. O resultado do *pairwise* está representado na Tabela 2.

Tabela 2 - Pairwise do Estudo de Caso 1 - Tecnologia da Informação

Problema	Importância	Facilidade	Retorno	Total
Assertividade na contratação de colaboradores	3	2	3	**2,75**
Retenção de colaboradores	4	1	2	**2,75**
Perfil de clientes potenciais para compra de Office 365	1	4	2	**2**
Assertividade na estimativa de projetos	2	0	5	**2,25**
Identificação em colaboradores com cargo desenvolvedor o perfil analista de sistemas	0	3	0	**0,75**

Fonte: Elaborado pelo autor

Através da análise de *pairwise* representada da Tabela 2, verificamos que houve um empate entre “Assertividade na contratação de colaboradores” e “Retenção de colaboradores”. Após avaliações e discussões sobre os dois domínios de problema, consideramos que o domínio de problema mais importante é **Retenção de colaboradores** e, de maneira geral, a solução deste problema gerará grande valor para a organização.

5.2.2.2 Definição do espaço solução

Através das Figuras 36 e 37, está representado o documento gerado na etapa de Definição do Espaço Solução.

Figura 36 - Documento de Definição do Espaço Solução - Página 1

PICTOREA LICAP – Laboratório de Inteligência Computacional Aplicada	DEFINIÇÃO DO ESPAÇO SOLUÇÃO

REVISÕES

Analista KDD:	André Montevecchi	Revisado em:	20/01/2012
Gerente de Projeto:		Revisado em:	
Analista de negócios:		Aprovado em:	

INFORMAÇÕES GERAIS

A empresa deste estudo de caso é da área de TI e tem aproximadamente 16 anos de atuação no mercado. Em sua sede em Belo Horizonte possui aproximadamente 150 colaboradores. Além de uma fábrica de software, a empresa possui uma área de treinamentos oficiais Microsoft e uma área de consultoria, sendo esta última, nas áreas de Business Intelligence – BI, gerenciamento de projetos, infraestrutura e produtos Microsoft.

Para esta segunda etapa do Pictorea, foi feita uma reunião na sede da empresa com a presença do diretor executivo, da gerente de RH e do analista KDD.

OBJETIVO

O objetivo desta etapa do Pictorea, é definir as expectativas sobre o resultado e as saídas esperadas. Em seguida, deverão ser definidas as técnicas de mineração de dados, e a forma de visualização dos resultados.

SAÍDAS ESPERADAS

O Especialista de Domínio espera como saída um relatório de importância de atributos que são decisivos para que um colaborador continue na empresa e não procure outras oportunidades fora da organização. De forma inversa, o Especialista de Domínio espera um relatório contendo os atributos que influenciam a decisão de um colaborador sair da organização.

Fonte: Elaborado pelo autor

Figura 37 - Documento de Definição do Espaço Solução - Página 2

PICTOREA LICAP – Laboratório de Inteligência Computacional Aplicada	DEFINIÇÃO DO ESPAÇO SOLUÇÃO

TÉCNICAS DE DATA MINING

Possíveis técnicas de data mining para atendimento das expectativas do Especialista de Domínio:

- Análise de associação de atributos.
- Árvore de decisão
- Análise de Cluster

FORMA DE VISUALIZAÇÃO DE RESULTADOS

- Relatório detalhado dos atributos.
- Relatório gráfico com porcentagem de atributos.

CONSIDERAÇÕES OU RESTRIÇÕES ADICIONAIS

Para esta etapa do Pictorea, foi identificada a seguinte melhoria:

- Mostrar exemplos de saídas esperadas e formas de visualização.

Fonte: Elaborado pelo autor

Como já mencionado, os objetivos da etapa de Definição do Espaço Solução são definir as expectativas sobre o resultado e as saídas esperadas.

Conforme explicado pelo Diretor de TI da organização, ele espera que seja exibido um relatório sintético com os tipos de itens/categorias que podem de alguma forma influenciar na ocorrência de discrepância entre o orçado e o executado.

De acordo com as expectativas sobre o resultado, entendemos que as técnicas de mineração de dados que mais se adequam são árvore de decisão e análise de cluster.

5.2.2.3 Melhorias identificadas do método PICTOREA

Na aplicação da etapa de Definição do Espaço Solução do Método PICTOREA, identificou-se a necessidade de no *pairwise* alterar o nome da coluna de “Dificuldade” para “Facilidade” com intuito de ficar mais fácil o entendimento e análise do especialista de domínio.

Outra melhoria identificada neste estudo de caso foi a alteração da ordem do levantamento das fontes primárias de dados. A princípio, o levantamento das fontes primárias de dados era feito antes da identificação dos domínios de problema e do *pairwise*. Fazendo a identificação de domínios de problema e *pairwise* antes, somente é necessário analisar as fontes primárias de dados do domínio de problema escolhido como alvo, melhorando assim, a produtividade da etapa.

6 CONSIDERAÇÕES FINAIS

Conforme exposto neste trabalho, foi criado um método para aplicação de descoberta de conhecimento em bancos de dados convencionais. Foi descrito o desafio e necessidade em encontrar uma teoria canônica para projetos KDD garantindo qualidade, reuso, padronização e diminuição de custos. Diante desta necessidade e da falta de um padrão, utilizamos o conhecimento de um especialista (tácito) através da metodologia interpretativista aliado às pesquisas científicas (explícito) para fundamentar as etapas necessárias para a correta condução e aplicação de um projeto KDD.

Posteriormente, foi descrito que um método para descoberta de conhecimento em bancos de dados segue uma sequência de etapas (processo) e para a notação de processos optou-se pelo BPMN, já consolidado tanto na academia quanto no mercado. Utilizamos neste trabalho, o BPMN para modelagem do processo principal e das subetapas do método PICTOREA.

Verificamos que há semelhanças entre um projeto KDD e metodologias para desenvolvimento de software. Com isso, utilizamos o SPEM juntamente com BPMN para a modelagem do PICTOREA, criando assim um processo híbrido de construção do método. Através do SPEM, foi gerada uma formalização que cumpre a função de auxílio na condução e aprendizagem na aplicação de descoberta de conhecimento em bancos de dados. Um ganho dessa abordagem, é que o método PICTOREA pode ser conduzido por profissionais menos experientes sob a supervisão de um especialista, além de ser um facilitador no aprendizado do tema, e por isso, chamamos o PICTOREA de um método pedagógico.

Foram realizados dois estudos de caso restritos às duas primeira etapas do método PICTOREA (Exploração do Espaço Problema e Definição do Espaço Solução), em área diferentes, com o intuito de validar a metodologia. Estes foram conduzidos de acordo com o conceito D3M, destacado neste trabalho. As etapas abordadas nos estudos de caso tiveram como foco principal atender a necessidade do usuário e a geração de valor para os negócios das organizações. Na aplicação dos estudos de caso, foram identificados pontos de melhoria, tais como demonstração prévia das possibilidade da descoberta de conhecimento em bancos de dados e demonstração das visualizações possíveis.

Foi possível verificar que o método PICTOREA deve passar por aprimoramentos na sua sequência de etapas, como observado nos estudos de caso.

Observou-se através dos estudos de caso, que o método PICTOREA possui capacidade de generalização, podendo ser aplicado em diferentes áreas ou domínios de problema. Porém, em cada um dos estudos de caso, encontramos pontos de melhoria diferentes o que nos aponta ser necessário novas validações. É importante ressaltar que as melhorias não são no nível estrutural do método PICTOREA, mas sim em nível que facilite a comunicação entre o analista KDD e o especialista de domínio.

A integração de BPMN com SPEM atendeu às necessidades de detalhamento e documentação do método. A representação gerada pelo SPEM e disponibilizada em Montevecchi (2011) pode contribuir para o aprendizado e treinamento de equipe através da intuitiva navegação entre os elementos do método.

Como contribuição, vale destacar o caráter pedagógico para o apoio à aprendizagem de KDD e a possibilidade de padronização com intuito de garantir maior qualidade, controle e métricas de avaliação em projetos de descoberta de conhecimento em bancos de dados.

Observamos o potencial do método para trabalho em grupo. A representação gerada através do SPEM, além de padronizar e formalizar os processos, auxilia o aprendizado de equipe sendo um ponto de consulta e compartilhamento da evolução do método, além de tornar possível a execução de projetos KDD por profissionais menos experientes.

Baseado na análise dos resultados deste trabalho, pode-se sugerir as seguintes linhas de investigação como trabalhos futuros:

a) validação do método PICTOREA com outros especialistas KDD.
b) continuação do detalhamento e implementação das demais etapas do PICTOREA.
c) aplicação de todas as etapas do método PICTOREA em um estudo de caso.
d) melhoria e tradução da interface gerada pelo SPEM.

REFERÊNCIAS

BIZAGI. Bizagi. **Bizagi Process Modeler,** 5 novembro 2011. Disponivel em: <http://www.bizagi.com/index.php?option=com_content&view=article&id=27&catid=5&Itemid=98/>. Acesso em: 08 fev. 2011.

BOENTE, Alfredo N. P.; GOLDSCHMIDT, Ronaldo, R.; ESTRELA, Vânia V. Uma Metodologia para Apoio e Realização do Processo de Descoberta de Conhecimento em Bases de Dados. In: WORKSHOP DE COMPUTAÇÃO CIENTÍFICA DA UENF, 2, 2006, Rio de Janeiro. **Anais…** Rio de Janeiro: UENF, 2006.

BRITOS, Paola. et al. Tool Selection Methodology in Data Mining. In: IBERO-AMERICAN SYMPOSIUM ON SOFTWARE ENGINEERING AND KNOWLEDGE ENGINEERING, 6, 2006, Puebla. **Anais…** Puebla: JIISIC, 2006. p. 85-90.

BRUSSE, Bart; WENNING, Rigo. **Standardization guidelines for IST research projects interfacing with ICT standards organizations**, Brussels: Cooperation Platform for Research and Standards, 2005. Disponível em: <http://www.w3.org/2004/copras/docu/D15.html>. Acesso em: 08 fev. 2011.

CAO, Longbing. Domain Driven Data Mining (D3M). In: INTERNATIONAL CONFERENCE ON DATA MINING WORKSHOPS, 6, 2008, Pisa. **Anais…** Pisa: IEEE Computer Society, 2008. p. 74-76.

CAO, Longbing. Domain Driven Data Mining: Challenges and Prospects. **IEEE Transactions on Knowledge and Data Engineering**, v. 99, n. 6, p. 755-769, 2010.

CHAPMAN, P. et al. **CRISP-DM 1.0:** Step-by-Step Data Mining Guide. SPSS, 2000. Disponível em: < http://www.whitepapercentral.com/browse/marketing/crisp-dm-1-0-step-by-step-data-mining-guide/ >. Acesso em: 09 fev. 2011.

ELSILA, Ulla; RONING, Juha. Knowledge Discovery in Steel Industry Measurements. In: STARTING ARTIFICIAL INTELLIGENCE RESEARCHERS SYMPOSIUM. , 2, 2002, Berlin. **Anais…** Berlin: STAIRS, 2002. Disponível em: <http://www.ee.oulu.fi/research/isg/files/pdf/pdf_369.pdf>. Acesso em: 13 fev. 2011.

FAYYAD, Usama.; PIATETSKY-SHAPIRO, Gregory.; PADHRAIC, Smyth. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, New York, v. 39, n. 11, p. 27-34, 1996.

FREITAS, Alberto.; BRAZDIL, Pavel.; PEREIRA, Altamiro. D. C. Mining Hospital Databases for Management Support. In: IADIS VIRTUAL MULTI CONFERENCE ON COMPUTER SCIENCE AND INFORMATION SYSTEMS, 1, 2005, Porto. **Anais…** Porto: IADIS Press, 2005. p. 207-212.

GENVIGIR, Elias. C; Filho, Luiz, F., B. 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, 2003, São José dos Campos. **Anais...** São José dos Campos: Instituto Nacional de Pesquisas Espaciais. 2003. p. 85-90.

GHEDINI, Cinara; BECKER, Karin. A documentation model for KDD application management support. In: INTERNATINAL CONFERENCE OF THE CHILEAN COMPUTER SCIENCE SOCIETY, 11, 2001, Punta Arenas. **IEEE Computer Society.** Punta Arenas: IEEE Computer Society, 2001. p. 105-114.

GOEBEL, Michael; GRUENWALD, Le. A Survey of Data Mining and Knowledge Discovery Software tools. In: ACM SPECIAL INTEREST GROUP ON KNOWLEDGE DISCOVERY AND DATA MINING,13, 1999, New York. **ACM SIGKDD Explorations Newsletter.** New York: ACM, 1999. p. 20-23.

GONZALEZ-ARANDA, P. et al. Towards a Methodology for Data Mining Project Development: The Importance of Abstraction. In: GONZALEZ-ARANDA, P. et al. **Data Mining: Foundations and Practice**: Studies in Computational Intelligence. 118. Springer Berlin: Springer, 2008. p. 165-178.

GRAHAM, Ben. **Business Process Improvement Methodology**. Ohio: The Ben Graham Corporation, 1999, 69 p.

HOMMES, L. J. **The Evaluation of Business Process Modeling Techniques**. 2004. 277f. Tese (Doutorado) - Delft University of Technology, Electrical Engineering, Mathematics and Computer Science, Delft.

HUMPHREY, Watts. S. **Managing the software process**. Boston: Addison-Wesley Longman,1989, 89 p.

JORGENSEN, Havard. D. **Interactive Process Models**. 2004. 304f.Tese (Doutorado) - Norwegian University of Science and Technology, Department of Computer and Information Science, Trondheim.

KDNUGGETS. **What main methodology are you using for data mining?**, 2002. Disponivel em: <http://www.kdnuggets.com/polls/2002/methodology.htm>. Acesso em: 20 mar. 2010.

KDNUGGETS. **Data Mining Methodology 2004**, 2004. Disponivel em: <http://www.kdnuggets.com/polls/2004/data_mining_methodology.htm>. Acesso em: 20 mar. 2010.

KDNUGGETS. **Data Mining Methodology 2007**, 2007. Disponivel em: <http://www.kdnuggets.com/polls/2007/data_mining_methodology.htm>. Acesso em: 20 mar. 2010.

LEE, Wenke.; STOLFO, Salvatore. J.; MOK, Kui. W. **A data mining framework for building intrusion detection models**. In: IEEE SYMPOSIUM ON SECURITY AND PRIVACY, [s.n.], 1999, Oakland: IEEE Computer Society, 1999. p. 120-132.

MINGERS, John. Combining IS Research Methods: Towards a Pluralist Methodology. **Information Systems Research**, Maryland, v. 12, n. 3, p. 240-259, 2001.

MONTEVECCHI, A. **Eclipse Process Framework Composer - Método PICTOREA**, 2011. Disponivel em: <http://www.montevecchi.com.br/SPEM>. Acesso em: 12 dez. 2011.

PAN, Ding. A formal framework for Data Mining process model. In: ASIA-PACIFIC CONFERENCE, 2009, Wuhan. **Computational Intelligence and Industrial Applications.** Wuhan**:** IEEE Computer Society, 2009. v. 2, p. 126-129.

PYLE, Dorian. **Data preparation for data mining**. San Francisco: Morgan Kaufmann, 1999, 466 p.

RECKER, Jan. C. et al. Do Process Modelling Techniques Get Better? A Comparative Ontological Analysis of BPMN. In: AUSTRALASIAN CONFERENCE ON INFORMATION SYSTEMS, 16, 2005, Sidney. **Australasian Chapter of the Association for Information Systems.** Sidney:CiteSeer, 2005. Disponível em: < http://eprints.qut.edu.au/2879/1/Recker_et_al-ACIS2005b.pdf>. Acesso em: 11 fev. 2011.

REFAEILZADEH, Payam.; TANG, Lei.; LIU, Huan. **Encyclopedia of Database Systems - Cross-Validation**. [S.l.]: Springer, 2009.

SAS INSTITUTE AND USING SAS AND ENTERPRISE MINER SOFTWARE. **Data Mining and the Case for Sampling - A SAS Institute Best Practices Paper Solving Business Problems Using SAS**, 1998. Disponível em: <http://nas.cl.uh.edu/boetticher/ml_datamining/sas-semma.pdf>. Acesso em: 02 ago. 2011.

SAS INSTITUTE INC. **SAS Enterprise Miner**, 2010. Disponivel em: <http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html>. Acesso em: 08 fev. 2011.

SOUNDARARAJAN, Ezekiel. et al. Knowledge Discovery Tools and Techniques. **Indira Gandhi Center for Atomic Research**, Kalpakkam, 02 set. 2004. Disponível em: <http://library.igcar.gov.in/readit-2005/conpro/km/s3-1.pdf>. Acesso em: 12 fev. 2011.

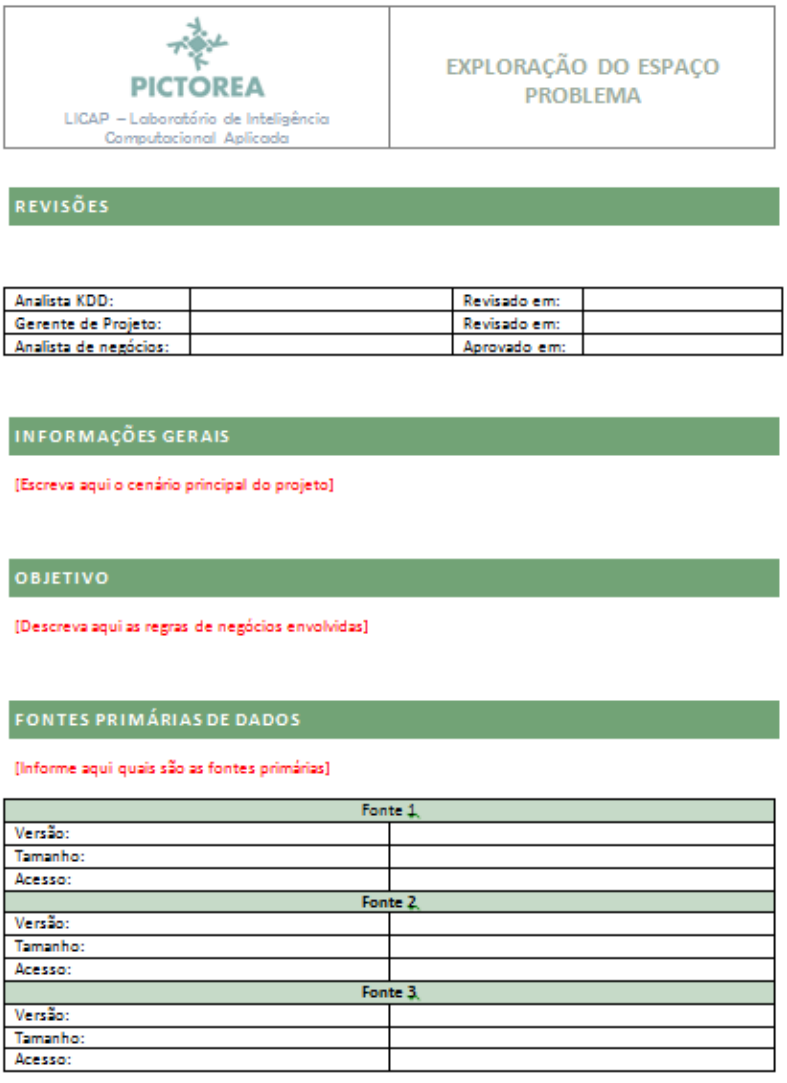
YANG, Qiang.; WU, Xindong. 10 Challenging Problems in Data Mining Research. **International Journal of Information Technology & Decision Making**, New Jersey, v. 5, p. 597-604, set. 2006.

ZENG, Huifang.; PAN, Ding. A knowledge discovery and data mining process model in E-marketing. In: INTELLIGENT CONTROL AND AUTOMATION (WCICA), 8, 2010, **Anais…**. Jinan: Springer, 2010. p. 3960-3964.

APÊNDICE A – *TEMPLATES* DA ETAPA DE EXPLORAÇÃO DO ESPAÇO PROBLEMA

A.1 Template de Exploração do Espaço Problema – Página 1

Figura 38 - *Template* de Exploração do Espaço Problema - Página 1



PICTOREA
LICAP – Laboratório de Inteligência Computacional Aplicada

EXPLORAÇÃO DO ESPAÇO PROBLEMA

REVISÕES

Analista KDD:		Revisado em:	
Gerente de Projeto:		Revisado em:	
Analista de negócios:		Aprovado em:	

INFORMAÇÕES GERAIS

[Escreva aqui o cenário principal do projeto]

OBJETIVO

[Descreva aqui as regras de negócios envolvidas]

FONTES PRIMÁRIAS DE DADOS

[Informe aqui quais são as fontes primárias]

Fonte 1	
Versão:	
Tamanho:	
Acesso:	
Fonte 2	
Versão:	
Tamanho:	
Acesso:	
Fonte 3	
Versão:	
Tamanho:	
Acesso:	

1

Exploração do Espaço Problema

Fonte: Elaborado pelo autor

A.2 *Template* de Exploração do Espaço Problema – Página 2

Figura 39 - Template de Exploração do Espaço Problema - Página 2

PICTOREA

LICAP – Laboratório de Inteligência Computacional Aplicada

EXPLORAÇÃO DO ESPAÇO PROBLEMA

MATRIZ DE PROBLEMAS (PAIRWIZE)

Problema	Importância	Dificuldade	Retorno	Total

RANKING

[Liste aqui o ranking de problemas obtivo através da matriz de problema]

CONSIDERAÇÕES OU RESTRIÇÕES ADICIONAIS

[Descreva aqui as considerações e restrições que achar importante]

2

Exploração do Espaço Problema

Fonte: Elaborado pelo autor