Experiência de Uso de um Processo de Software para Sistemas KDD no Contexto de Instituições Públicas

Adriana Herden[1], Leandro Araújo[1], Brandon de Almeida[1], Adriano Bessa Albuquerque[2], Pedro Porfírio Muniz Farias[2]

[1]Departamento de Informática, Universidade Tecnológica Federal do Paraná
Cornélio Procópio, Paraná, Brasil

herden@utfpr.edu.br; leoaraujo21@gmail.com; brandonalmeida@hotmail.com

[2] Programa de Pós-Graduação em Informática Aplicada, Universidade de Fortaleza
Fortaleza, Ceará, Brasil

adrianoba@unifor.br; porfirio@unifor.br

Resumo. *Este artigo descreve a aplicação de um processo de software para aplicações analíticas em organizações públicas. O processo escolhido é chamado de UPKDD (Unified Process for Knowledge Discovery in Database), e foi utilizado para apoiar a tomada de decisão gerencial. Primeiramente, optou-se pela utilização do processo na empresa Sanepar, e depois na Prefeitura Municipal de Ibiporã. Além das lições aprendidas, os resultados mostraram que o UPKDD apoia os tomadores de decisão, especialmente em como explorar os dados tornando-os úteis, e em como mapear as expectativas dos tomadores de decisão em situações realísticas.*

Abstract. *This article describes the application of a software process for analytical applications in public organizations. The process chosen is called UPKDD (Unified Process for Knowledge Discovery in Database), and was used to support decision making by the management departments. First, we chose to use the process in the company Sanepar, and then the government of the City of Ibiporã. In addition to the lessons learned, the results showed that the UPKDD supports decision makers, especially in how to exploit the data making them useful, and how to map the expectations of decision makers in realistic situations.*

1. Introdução

A busca por conhecimentos úteis em dados estruturados dá suporte a decisões estratégicas e gerenciais da maioria das empresas de médio e grande porte, sejam elas públicas ou privadas. Fayyad, Piatetsky-Shapiro e Smyth (1996) definem *Knowledge-Discovery in Databases* (KDD) como "*o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em conjuntos de dados*". Igualmente, afirmam a necessidade emergente da criação de ferramentas computacionais que auxiliem o ser humano na extração rápida de conhecimento.

Os processos de KDD propostos por diversos autores, como Fayyad, Piatetsky–Shapiro e Smyth (1996), Han e Kamber (2001), Brachman e Anand (1996) e Reinartz (1999), identificam as principais etapas para obter conhecimento podendo ser divididos em três passos básicos que são: pré-processamento, mineração de dados e pós-processamento. Por conseguinte, não são abordados nos processos de KDD artefatos, responsabilidades, e atividades que transformem a estrutura de dados organizacionais

em dados que favoreçam o uso de tecnologias analíticas, como *data warehouse* processamento analítico *on-line*, modelos multidimensionais e mineração de dados.

Neste sentido, observou-se a carência de um processo de software, que poderia ser aplicado em cada fase do desenvolvimento de um sistema de apoio à decisão, uma vez que a maioria dos processos propostos pela engenharia de software atua sobre sistemas transacionais. Conforme Dias (2001), a ordenação rigorosa de atividades para a descoberta de conhecimento, diminui a característica de indeterminismo dos sistemas.

Alguns trabalhos já foram desenvolvidos propondo adaptação de processos tomando como base o Processo Unificado de desenvolvimento de software (UP), proposto por Jacobson, Booch e Rumbaugh (1999), porém nenhum destes voltado para o desenvolvimento de sistemas de apoio à decisão. A adaptação mais conhecida pelo meio acadêmico e industrial é o RUP. Nesta adaptação, conforme descrito por Kruchten (2003), são inseridos três *workflows* gerenciais. Tais *workflows* como gerência de projeto e controle do ambiente envolvendo atividades de treinamento foram incluídos no Processo Unificado para atender mais adequadamente a realidade empresarial.

Este artigo tem como objetivo relatar a experiência da aplicação do UPKDD em organizações públicas e está assim organizado: a seção 2 apresenta o processo UPKDD; na seção 3 as experiências de uso; na seção 4 estão os resultados e, na seção 5 são apresentadas as conclusões sobre o trabalho.

2. O Processo UPKDD

O processo UPKDD esta descrito em Herden (2007) e Herden et.al. (2011), e é caracterizado pela permanência dos aspectos dinâmicos e estáticos do UP, além da agregação de duas perspectivas, que são: (i) condução do processo de KDD, (ii) análise da estrutura de dados existentes.

Para a modelagem do processo de software UPKDD, optou-se pelo metamodelo *Software Process Engineering Metamodel Specification* SPEM [OMG 2005]. A figura 1(A) mostra um diagrama de pacotes da visão geral do UPKDD. Nota-se que nesta figura houve a divisão do processo em Disciplinas, que são: **Requisitos, Análise e Projeto**.

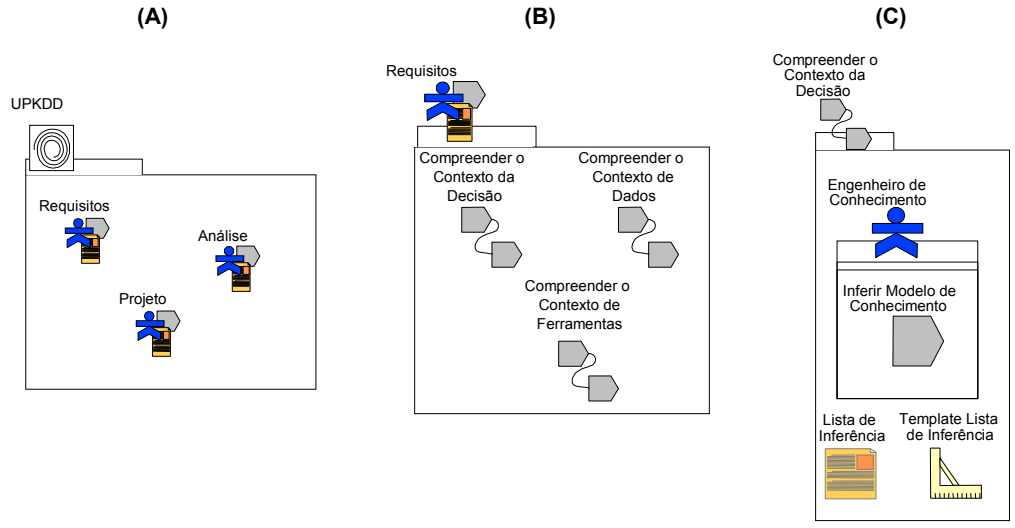

Figura 1: Diagramas Estruturais do UPKDD

Posteriormente, na figura 1 (B), a disciplina de **Requisitos** foi detalhada em três Definições de Trabalho, para compreender o contexto: da decisão, de dados, e ferramentas. Depois tais Definições de Trabalho, de cada Disciplina, foram representadas em outros diagramas, como os de pacotes, caso de uso, classe e atividades [Herden 2007]. A figura 1(C) mostra os elementos da Definição de Trabalho **Compreender o Contexto da Decisão,** e a figura 2 mostra os papéis que participam desta Definição de Trabalho**,** a qual captura as expectativas dos tomadores de decisão, que por sua vez são formalizadas na lista de inferência.

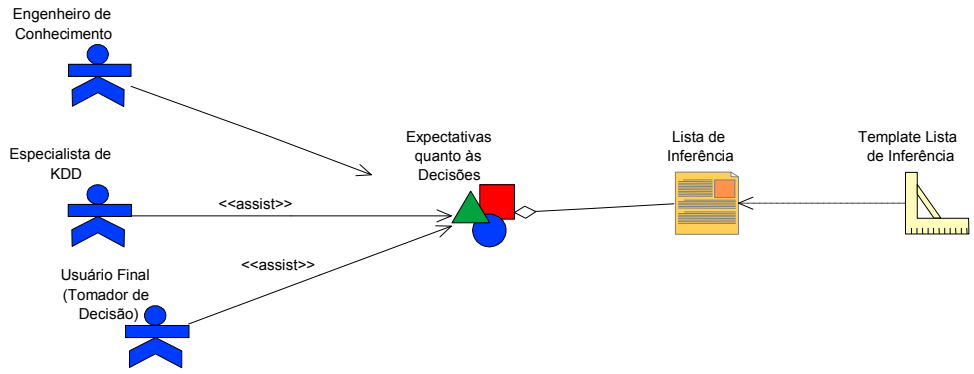

Figura 2: Diagrama de Classes – Compreender o Contexto da Decisão

No modelo do processo os Elementos-Chave (EC) da área de KDD foram integrados ao processo de software tradicional, e definidos em três princípios, que são:

(i) Artefatos: Lista de Inferência, Descrição da Base de Dados Existente, Caracterização de Ferramentas Existentes, Lista de Informações de Negócio, Matriz de Barramento, Modelo Dimensional.

(ii) Responsabilidades: Engenheiro do Conhecimento, Especialista de KDD, Usuário Final (Tomador de Decisão).

(iii) Atividades: Analisar o Contexto de Dados, Analisar o Contexto de Ferramentas, Gerar Informações de Negócio, Projetar Arquitetura de *Datawarehouse*, Construir a Modelagem de Dados.

Após a definição do UPKDD optou-se pela realização de um estudo de caso, a fim de verificar se as mudanças propostas pela utilização deste auxiliam no desenvolvimento de sistemas de apoio à decisão. Avaliação experimental (instanciação do UPKDD) durou aproximadamente cinco meses, envolveu doze desenvolvedores e passou por quatro fases, que foram: definição, planejamento, operação e interpretação, conforme *framework* experimental de Basili, Selby e Hutchens (1985). O experimento utilizou a abordagem GQM, proposta pelo mesmo autor, para o estabelecimento de objetivos mensuráveis, a fim de quantificar variáveis que mostram o efeito da mudança como a utilização, utilidade e adequação textual de cada EC.

Então, observou-se que o UPKDD é adequado, pois os resultados quantitativos do estudo de caso realizado mostram ***alta utilização dos EC's propostos***. No acompanhamento dos grupos de participantes, durante o exercício de desenvolvimento de uma solução KDD para um laboratório de análises clínicas, comprovou-se a necessidade de um processo realmente focado em características deste tipo de aplicação como o UPKDD.

3. Experiências de Uso

Após a avaliação do processo foram escolhidas duas instituições públicas para observar se a utilização de um processo para sistemas KDD é eficiente para os tomadores de decisão em ambientes reais. Houve treinamento em ambas as organizações, antes da sua adoção pelos desenvolvedores. Após o treinamento, já durante a aplicação do processo, as equipes passaram pelas fases de Concepção e Elaboração do UPKDD. Tais equipes iniciaram os trabalhos construindo o artefato Lista de Inferência, e posteriormente construíram os demais artefatos propostos pelo modelo.

Uma das instituições escolhida foi a Companhia de Saneamento do Paraná – Sanepar, que tem como compromisso universalizar o acesso ao saneamento, levando os serviços de fornecimento de água tratada, coleta e tratamento de esgoto sanitário para paranaenses. Nesta experiência de uso buscava-se conhecer os motivos que causam ***perda de água*** durante o abastecimento de usuários [Araújo 2009].

A outra instituição escolhida foi a Prefeitura Municipal de Ibiporã (PMI) e sua administração pública, principalmente o poder executivo. Esta experiência de uso foi conduzida para conhecer a ***previsão de arrecadação pública***, com ênfase nos impostos municipais, como: imposto predial e territorial urbano (IPTU), imposto sobre serviços de qualquer natureza (ISSQN), fundo de participação dos municípios (FPM), cota parte do imposto de circulação de mercadorias e serviços (ICMS) e outros [Almeida 2009].

3.1. Projeto para Sanepar

O projeto abordou o desenvolvimento de um software com dois módulos analíticos para a Sanepar. A principal finalidade do software foi descobrir os motivos que causam perda de água durante o abastecimento de usuários. A busca foi realizada essencialmente na base de dados já existente, que é composta por histórico de leituras e hidrômetros instalados na cidade de Cornélio Procópio e região, com mais de 45.000 clientes.

No primeiro **Módulo (OLAP)** foram investigadas características do perfil das ligações de água, além do volume medido e suas propriedades em determinado período. As tabelas do banco de dados utilizadas foram a de *Leitura*, *Hidrômetro*, *Instalação*, *Categoria* e *Clientes*, todas relacionadas dentro do período de 2006 a 2009. Então, a Lista de Inferência (proposições a serem provadas pelas investigações na base de dados e elaboradas pelo tomador de decisão) continha as seguintes perguntas:

a) Quais são as características das trocas de hidrômetro, no decorrer do tempo, para determinado local?

b) Quais são as características (Marca, Ano, Tamanho, e n^{o}. de habitantes por ligação de água) da ligação de água em determinado período?

c) Quais são as ligações de água com menor consumo por habitante? (ligações acima de 4 pessoas e com consumo baixo).

Vale ressaltar que a escolha destas inferências partiu do princípio, segundo a Sanepar e a Organização Mundial de Saúde (OMS), que uma pessoa necessita em média de 2,5 m³ de água por mês, sendo assim ligações com consumo menor que isso pode indicar o desgaste do hidrômetro.

Após a realização de um ensaio com 10 ligações de água, tendo como referência o período para avaliação os meses de março e abril de 2009, obteve-se o seguinte resultado: (i) 7 ligações apresentaram aumento no consumo medido e (ii) 3 ligações

apresentaram o mesmo consumo medido ou sua diminuição. Com base neste resultado, foi realizada a substituição de 10 hidrômetros.

Já no segundo **Módulo (Mineração de Dados)** foram identificados padrões de dados, especialmente pela aplicação de algoritmos. Optou-se pela utilização de métodos baseados em árvores de decisão. Neste caso, para compreender o contexto da decisão a Lista de Inferência elaborada foi:

a) Porque algumas substituições de hidrômetro aumentam o volume medido?

b) Porque algumas substituições de hidrômetro mantém volume medido?

c) Quais são as características (Marca, Ano Fabricação, Tamanho) da ligação de água que tiveram aumento de volume medido após a substituição do hidrômetro?

d) Porque hidrômetros com determinadas características apresentam aumento de consumo após a troca?

Para a preparação da amostra de dados foi considerada a evolução do consumo, após a troca de hidrômetro (período de cinco meses). Depois, foi executado o algoritmo de mineração de dados J48 [Goldschmidt e Passos 2005]. Este é uma implementação do algoritmo C4.5, que por sua vez gera uma árvore de decisão com base em um conjunto de dados, presente em uma API do *Weka Data Mining Software* [Hall et. al. 2009].

Logo, o resultado caracterizou medidores que apresentaram problemas no seu funcionamento, indicando grande chance de serem trocados. Para a aplicação do algoritmo foram selecionadas todas as trocas de hidrômetro, e as seguintes situações foram observadas:

(i) após a substituição os hidrômetros que possuem código de marca 4, e com ano de fabricação depois de 1992 e antes que 1993 inclusive, apresentam faixa *média* de volume medido após sua substituição

(ii) após a substituição os hidrômetros que possuem código de marca 4, e com ano de fabricação depois de 1993 e antes de 1997 inclusive, apresentam faixa *grande* de volume medido após sua substituição.

Então foi realizada a substituição de 15 hidrômetros com código de marca 4, e o período de avaliação foram os meses de março e abril de 2009, e obteve-se o seguinte resultado: (a) 10 ligações apresentaram aumento no consumo medido, e (b) 5 ligações apresentaram mesmo consumo medido ou sua diminuição.

3.2. Projeto Prefeitura Municipal de Ibiporã (PMI)

Neste projeto foi abordado o desenvolvimento de um software que utilizasse tecnologias analíticas de descoberta de conhecimento em banco de dados para a PMI, a fim de disponibilizar uma previsão orçamentária aos tomadores de decisão de administração pública, especialmente na área de arrecadação de receitas. Considerando como base de dados relevante o histórico da arrecadação anual e outros dados históricos e correlatos. Neste caso foram utilizados algoritmos de regressão linear, pois estes oferecem resultados de previsão contínuos e discretos.

No primeiro **Módulo (OLAP)**, o objetivo foi entender o contexto da decisão, então foi elaborado o artefato Lista de Inferência. E o período escolhido para análise foi o 1º bimestre do ano, pois apresenta aproximadamente 20% da arrecadação em diversos exercícios financeiros diferentes. Também foi considerada uma base de dados que

possui informações desde janeiro de 2005 até 2009. E o tomador de decisão escolheu determinadas receitas, especialmente as de maior retorno, e as de menor prazo de realização. Então, a inferência escolhida foi:

a) Quais as características dos imóveis da área central?

Depois de conhecer um pouco sobre imóveis e arrecadações foi elaborada uma Matriz de Barramento, a fim de conhecer os processos de negócio. No processo de preparação e seleção de amostras foi necessária à eliminação de detalhamentos das informações individuais de cada contribuinte no município. Ainda foi necessário trabalhar com exportação da tabela [Arrecadacao_Analitica] para a tabela Fato, do banco dimensional [Fato_Arrecadacao_Analitica].

Após as consultas OLAP o resultado mostrou que a maioria dos imóveis é do tipo *residencial*, quando não estes são do tipo *outros*, sendo edificadas em áreas planas e sem acabamento.

O segundo **Módulo (Mineração de Dados)** foi desenvolvido com a finalidade de prever as receitas orçamentárias municipais, por meio de regressão linear em séries temporais, a partir das inferências nomeadas na investigação de padrões. Nesta fase o tomador de decisão necessitava das previsões orçamentárias anuais. Então a investigação foi norteada pelas perguntas:

a) Porque existe aumento da arrecadação de "Cota Parte de IPVA" de 2005 a 2009?

b) Estimar a arrecadação do tributo de "Cota Parte de IPVA" para o exercício financeiro de 2008, e efetuar comparação da metodologia empregada atualmente.

A configuração da série temporal foi realizada da seguinte forma: (i) Os cálculos deverão prever a arrecadação para o ano de 2008, de forma a demonstrar como a regressão se porta em relação à praticada atualmente; (ii) A primeira série é composta dos meses de janeiro de 2005 a dezembro de 2007; (iii) A segunda série é composta da mesma forma que a metodologia atualmente adotada no município para a previsão de arrecadação, iniciando-se em outubro de 2006 a setembro de 2007. Ainda, para a elaboração da série temporal a ser utilizada, o banco de dados foi submetido à classificação em ordem cronológica crescente, utilizando-se de consulta SQL, a qual retorna para a aplicação uma sequência com os campos do exercício financeiro, mês de competência, código da receita orçamentária e total mensal.

Em relação ao módulo de mineração de dados, com foco em predição de arrecadação de receitas públicas, observou-se volumosa quantidade de técnicas que possibilitam efetuar este processo. Desta forma optou-se por utilizar conceitos de econometria, neste caso a regressão linear em séries temporais.

4. Resultados

No projeto Sanepar nota-se que o principal conhecimento obtido mostrou a diferença existente na medição do consumo entre marcas diferentes de hidrômetros instalados. A partir dos resultados obtidos, por meio de consultas OLAP, foram solicitadas ***trocas preventivas*** em ligações de água com hidrômetros que apresentaram *menor consumo por habitante*. E, com os resultados obtidos a partir de mineração de dados, também foram solicitadas ***trocas preventivas*** para ligações com hidrômetros que apresentaram *aumento de volume medido após a troca*. Após a experiência de uso foi possível observar que 70% das trocas apresentaram aumento no consumo medido. Portanto, o

UPKDD apoiou a avaliação de dados no controle de perdas de água, desde a representação do modelo de conhecimento idealizado pelo tomador de decisão, por meio da Lista de Inferência, até a implementação do sistema. Além de sugerir a criação de uma base de conhecimento compartilhada sobre as trocas preventivas.

No projeto da PMI foram observados os imóveis e suas arrecadações, e também foram realizadas várias simulações de estimativas de arrecadação para o ano de 2008. Cada uma com três cálculos de séries temporais diferentes, que foram: três, dois e um ano de base (exercício financeiro). A partir dos resultados obtidos foi observado que houve ***bastante diferença entre previsto e arrecadado*** para alguns tributos, que são: [IPTU – oitenta e oito mil reais]; [Cota Parte FPM – quinhentos e vinte e cinco mil reais] e [Cota Parte ICMS – um milhão e setecentos mil reais]. Já no caso do tributo Cota Parte IPVA o valor foi de sessenta mil reais, considerado como ***pouca diferença entre previsto e arrecadado***. Logo, o UPKDD apoiou a predição da arrecadação de receita pública da PMI, pois melhorou a estimativa de origem de recursos a serem recebidos, e melhorou indiretamente a aplicação desses recursos em despesas previamente definidas.

5. Conclusões

A alta utilização dos elementos criados pelo UPKDD trouxe benefícios para as instituições públicas, principalmente quanto ao direcionamento do processo de descoberta de conhecimento, que anteriormente era realizado sem mapear as expectativas dos usuários, muitas vezes, trazendo conhecimento inválido para estes.

Outros benefícios que podem ser destacados são o embasamento para tomada de decisões como foi na troca de hidrômetros, o controle sobre os fatores que causam a perda de água, a transformação dos dados estruturados em um modelo dimensional para futuras consultas de conhecimento. Já, o benefício para a PMI foi permitir a execução de melhores políticas públicas para o município, a partir dos valores previstos.

Outros fatores que causam perda de água foram identificados, porém não foram analisados por este trabalho. Como por exemplo, informações a respeito do clima (temperatura, umidade relativa do ar, volume de chuvas) associadas a variação de consumo dos clientes. Assim, durante a aplicação do processo não houve nenhum outro processo ou alteração do ambiente.

Alguns pontos negativos podem ser lembrados como a falta de explicação na definição do UPKDD dos detalhes para a implementação de OLAP. Outros pontos são: em geral existem muitos dados para análise (funções de ETL), equipe pequena de desenvolvedores, falta de acompanhamento do processo de KDD por parte dos tomadores de decisão, e pouco tempo dedicado ao treinamento.

Vale ressaltar a importância de adaptar processos de software às realidades empresariais, e ainda as particularidades dos tipos de sistemas como é o caso do KDD. Esta experiência de uso mostrou que a aplicabilidade de um processo de software enriquece a definição do seu modelo proposto.

Referências

Almeida, B. (2009) "*Sistema de descoberta de conhecimento em banco de dados em área pública: arrecadação de receita pública*". 224 f. Trabalho de Conclusão de

Curso (Graduação) de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná (UTFPR),Cornélio Procópio.

Araújo, L. (2009) *"Sistema de apoio à avaliação de dados no controle de perdas de água"*. 180 f. Trabalho de Conclusão de Curso (Graduação) de Tecnologia em Análise e Desenvolvimento de Sistemas, Universidade Tecnológica Federal do Paraná (UTFPR),Cornélio Procópio.

Basili, V. R.; Selby, R. W. Jr.; Hutchens, D. H. (1985) *"Experimentation in software engineering*", Relatório TR1575. Universidade de Maryland, USA.

Brachman, R. J.; Anand, T. (1996) *"The process of knowledge discovery in databases: a human-centered approach*", In: Fayyad, U. M.; Piatetsky-Shapiro,G.; Smyth, P.; Uthurusamy, R., (editores). Advances in Knowledge Discovery and Data Mining. Menlo Park, Calif.: American Association for Artificial Intelligence (AAAI)/MIT Press.

Dias, M. M. (2001) *"Um modelo de formalização do processo de desenvolvimento de sistemas de descoberta de conhecimento em banco de dados*". 197f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina (UFSC), Florianópolis.

Fayyad, U.; Piatetsky–Shapiro, G.; Smyth, P. (1996) *"From Data Mining To Knowledge Discovery in Databases"*, IA MAGAZINE, American Association for Artificial Intelligence, Menlo Park.

Goldschmidt, R.; Passos, E. (2005). "*Data mining"*. Rio de Janeiro: Elsevier.

Hall, M. et. al. (2009). "*The WEKA Data Mining Software: An Update*. SIGKDD Explorations, Volume 11, Issue 1.

Han, J.; Kamber, M. (2001) *"Data mining: concepts and techniques"*, In: Gray, Jim (ed.). USA: The Morgan Kaufmann Series. 500p. (Series in Data Management Systems).

Herden, A. (2007) *"UPKDD: um processo para desenvolvimento de sistemas de descoberta de conhecimento em banco de dados"*. 171f. Dissertação (Mestrado em Ciência da Computação) – Programa de Pós-Graduação em Ciência da Computação, Universidade Estadual de Maringá (UEM), Maringá.

Herden, A. et.al. (2011) *"UPKDD: Processo de Software para Aplicações de Tecnologias Analíticas e Centradas em Objetivos de Descoberta de Conhecimento"*. In: X Simpósio Brasileiro de Qualidade de Software X SBQS 2011. Curitiba, Paraná.

Jacobson, I.; Booch, G.; Rumbaugh, J. (1999) *"The unified software development process"*. 2.ed. Canadá: Addison-Wesley. 463p. (Object Technology Series).

Kruchten, P. (2003) *"Introdução ao RUP"*. Rio de Janeiro: Ciência Moderna.

OMG Object Management Group. (2005) "*Software process engineering metamodel specification (SPEM)*", Relatório Técnico - OMG document number formal/05-01-06), disponível em: <http://www.omg.org>.

Reinartz, T. (1999) *"Focusing solutions for data mining: analytical studies and experimental results in real-world domains"*, In: Siekmann, J; Carbonell, J. G. Lecture Notes in Artificial Intelligence (LNAI). New York: Springer-Verlag.