UPKDD: Processo de Software para Aplicações de Tecnologias Analíticas e Centradas em Objetivos de Descoberta de Conhecimento

Adriana Herden[1], Maria Madalena Dias[2], Sheila Reinehr[3], Andreia Malucelli[3]

[1]Departamento de Informática, Universidade Tecnológica Federal do Paraná
Cornélio Procópio, Paraná, Brasil
herden@utfpr.edu.br

[2] Departamento de Informática, Universidade Estadual de Maringá
Av. Colombo, 5790, CEP 87020-900, Maringá, Paraná, Brasil
mmdias@din.uem.br

[3] Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná
Rua Imaculada Conceição 1155, CEP 80215-901, Curitiba, Paraná, Brasil
sheila.reinehr@pucpr.br; malu@ppgia.pucpr.br

Resumo. *Cada vez mais os processos de desenvolvimento de engenharia de software abordam aspectos relacionados à interação com o usuário, aos mecanismos que aumentem a produtividade e que ajudem a estimar de maneira mais realista orçamentos e prazos. O processo unificado de desenvolvimento de software (Unified Process - UP), além de propor soluções para esses problemas, também é considerado um framework de processo para ser personalizado, conforme necessidades da aplicação. Assim sendo, este artigo apresenta uma adaptação do UP, em uma perspectiva transformacional inerente ao UP, com ênfase nas necessidades específicas do desenvolvimento de sistemas de descoberta de conhecimento em banco de dados. O processo proposto, denominado de UPKDD, foi avaliado utilizando o método de avaliação experimental.*

Abstract. *Increasingly the processes of software engineering development addresses aspects related to user interaction, the mechanisms that increase productivity and help to more realistically estimate budgets and timelines. The unified process for software development (Unified Process - UP) proposes solutions to these problems and it is considered as a framework of process to be customized, according to application needs. Therefore, this paper presents an adaptation of the UP in a transformational perspective inherent to UP, with emphasis on the specific needs of developing systems for knowledge discovery in database. The proposed process, named UPKDD was evaluated using the method of experimental evaluation.*

1. Introdução

Sistemas de apoio à decisão tiveram sua origem nas décadas de 40 e 50 e partem do princípio básico de analisar o comportamento do negócio com base na busca de dados operacionais a fim de modificar comportamentos da empresa de maneira adequada. Nas décadas de 60 e 70 promoveu-se a utilização de computadores durante o processo de

tomada de decisões. No contexto histórico de Date (2003) a forma de utilização dos computadores era apenas para gerar relatórios implementados pelas linguagens de consulta da época. Por meio do avanço tecnológico, surgiram os bancos de dados relacionais nos anos 80, que incentivaram abandonar o uso de arquivos simples e estudos de técnicas na área de apoio à decisão. Atualmente, a maioria dos sistemas gerenciadores de bancos de dados possibilita o uso de tecnologia de apoio à tomada de decisão, tais como *data warehouse*, processamento analítico *on-line* (OLAP), modelos multidimensionais e mineração de dados.

De modo geral os sistemas de processamento de transação geram grande quantidade de informações, dessa forma, a busca por soluções mais eficientes de análise exploratória dos dados tornou-se indispensável, levando ao desenvolvimento de estratégias que auxiliem o tomador de decisões nessa tarefa. Uma dessas soluções é o processo de Descoberta do Conhecimento em Bases de Dados, (KDD, do inglês *Knowledge Discovery in Databases*). KDD é um processo que busca, por meio da aplicação de algoritmos e etapas específicas, descobrir conhecimentos válidos, novos e úteis em base de dados (Fayyad, Piatetsky–Shapiro e Smyth 1996). O processo de KDD é separado em três etapas, sendo composto tradicionalmente por: pré-processamento, mineração de dados e pós-processamento (Rezende 2005).

As etapas do processo de KDD podem ser vistas como requisitos do domínio[1], para um sistema de apoio a decisão. Neste caso o domínio envolve as funções de ETL (*Extraction/Transformation/Loading*), a execução de algoritmos, e as funções de compreensibilidade do conhecimento descoberto. Logo, pré-processamento, mineração de dados e pós-processamento não definem um processo de software para o desenvolvimento de um sistema de apoio à decisão. Desta forma, 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), definem e contribuem positivamente para identificação das principais funções de um sistema de apoio à decisão para a obtenção de conhecimento em banco de dados.

Já a definição dos processos de engenharia de software possui artefatos, responsabilidades e atividades bem definidas, possibilitando adequação às necessidades específicas. Esses processos são geralmente voltados ao desenvolvimento de sistemas de processamento de transação, não atendendo totalmente às necessidades de sistemas de apoio à decisão, que envolvem muitas vezes diferentes processos, técnicas e ferramentas.

Neste contexto, percebe-se a inexistência de um processo de software, que possa 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. Este processo deve conter a definição de: (i) artefatos que capturem e associem os objetivos de descoberta de conhecimento aos dados nas várias fontes de origem, (ii) responsabilidades que interajam direta ou indiretamente na tomada de decisão e, (iii) atividades que transformem a estrutura de dados organizacional em dados que favoreçam o uso de tecnologias analíticas. É interessante que este seja baseado no Processo Unificado de desenvolvimento de software (UP), proposto por Jacobson, Booch e Rumbaugh (1999), pelo fato de ser considerado um processo estabelecido e que se adapta de acordo com o domínio da aplicação.

Os benefícios esperados pelo uso do UPKDD são principalmente para apoiar o desenvolvedor na elaboração de uma estrutura de dados formatada para suportar

[1] Requisitos do Domínio são requisitos que se originam do domínio de aplicação do sistema e refletem características desse domínio (Sommerville 2003).

aplicações analíticas, e no controle para a condução da descoberta de conhecimento, especialmente em banco de dados.

Alguns trabalhos já foram desenvolvidos propondo uma adaptação de processos tomando como base o Processo Unificado, porém nenhum destes voltado para o desenvolvimento de sistemas de apoio à decisão.

Sousa (2004) propôs uma abordagem para separação de preocupações transversais, desde o início do processo de desenvolvimento, considerando o paradigma orientado a aspectos e as fases e *workflows* do UP. As contribuições deste trabalho apresentam melhoria no reuso, manutenção e compreensão dos artefatos gerados no processo de desenvolvimento.

Existe uma abordagem de adaptação do RUP para o domínio de Jogos Móveis proposto por Almeida (2006). Este trabalho propõe adequações que partem do UP discutindo papéis, artefatos e fases envolvidos na produção de jogos móveis. Como contribuição deste trabalho é a solução de processo para o domínio de jogos móveis.

Em Álvares (2001) foi definido um processo para aplicações Web, especificamente para o sistema e-Merci. O trabalho é caracterizado por personalizar o processo Práxis para a realidade de aplicações Web, denominado WebPraxis. Houve inserção de fluxos principalmente relacionados à usabilidade, e mostrou-se correto quanto a ordem temporal e lógica das atividades.

A adaptação mais conhecida pelo meio acadêmico e industrial é o RUP. Nesta adaptação, conforme descrito por Kruchten (2003), é inserido três *workflows* gerenciais, que atendem adequadamente a realidade empresarial como gerência de projeto e controle do ambiente envolvendo atividades de treinamento.

A intenção de formalizar o processo de desenvolvimento de aplicações KDD, investigada por Dias (2001), prova que a ordenação rigorosa de atividades para a descoberta de conhecimento diminui a característica de indeterminismo desses sistemas.

Este artigo tem como objetivo apresentar um processo para sistematização do desenvolvimento de aplicações que utilizam o KDD, denominado UPKDD (*Unified Process for Knowledge Discovery in Database*).

Este artigo está assim organizado: a seção 2 apresenta o processo proposto UPKDD; na seção 3 são apresentadas considerações sobre a avaliação realizada do processo proposto; na seção 4 estão os resultados e discussão e, na Seção 5 são apresentadas as conclusões sobre o trabalho.

2. O Processo UPKDD

A caracterização do UPKDD é dada pela permanência das estruturas dinâmicas e estáticas do UP, assim como a agregação de dois aspectos, a saber: condução do processo de KDD e análise da estrutura de dados empresariais já existentes. Assume-se como premissa que o UPKDD é dirigido por Objetivos de Descoberta de Conhecimento e, também, centrado em Arquitetura e Tecnologias Analíticas.

A adaptação proposta por este modelo destaca-se porque os objetivos da descoberta de conhecimento são mapeados desde o início do detalhamento do processo de KDD, sendo estes acompanhados em uma perspectiva transformacional até a base de conhecimento final fornecida pela aplicação, assim como ocorre com os requisitos funcionais nos sistemas de processamento de informação. Por sua vez, o uso das tecnologias analíticas como o *Data Warehouse* (DW), permitem implementar funcionalidades de sistemas de apoio à decisão, sendo no UPKDD delineadas tanto na modelagem inicial e multidimensional dos dados, oriundos da base de dados existente,

quanto na definição da arquitetura para implementação de consultas *Online Analytical Processing* (OLAP), ou ainda na aplicação de algoritmos de mineração de dados.

Antes da utilização do UPKDD pelos desenvolvedores, o processo foi definido pelas regras da modelagem de processo, que representam de maneira abstrata os elementos que compõem o processo. Para a representação de um processo de software, é necessário o uso de modelos de processo. A modelagem de processo de software descreve a criação de modelos do processo de desenvolvimento de software, referindo-se à definição de processos como modelos (Acuña e Ferré, 2001).

Neste contexto, a modelagem de processos de software surge como um formalismo para tratar a complexidade natural do desenvolvimento de software, oferecendo essencialmente a representação das características dos mesmos, de maneira precisa e compreensiva para os engenheiros de software. Para a modelagem do processo de software UPKDD, optou-se pela notação definida pelo metamodelo *Software Process Engineering Metamodel Specification* SPEM, segundo o OMG (2005).

O modelo conceitual de qualquer processo de software é definido pela colaboração entre entidades ativas e abstratas de Papéis no Processo, que executam operações, chamadas de Atividades, em entidades tangíveis e concretas, chamadas de Produtos de Trabalho. Papéis no Processo descrevem responsabilidades e competências de um indivíduo; Atividades ou Definições de Trabalho descrevem o que um Papel no Processo executa; e Produtos de Trabalho descrevem um pedaço de informação produzido ou usado por uma atividade.

O UPKDD tem como referência algumas diretrizes, a saber: (1) condução do processo de KDD baseado principalmente em Fayyad, Piatetsky–Shapiro e Smyth (1996) e Brachman e Anand (1996); (2) definição e implementação da arquitetura de DW segundo Kimball e Ross (2002); (3) boas práticas do *Unified Process* (UP) segundo Jacobson, Booch e Rumbaugh (1999); e (4) o Conjunto Comum de ECs (Elementos-Chave). Os Elementos-Chave representam um consenso dos autores da área de processo, arquitetura e implementação de soluções de sistemas de apoio à decisão e são divididos em artefatos, papéis e atividades. Vale ressaltar que os princípios, as fases, os *workflows* e também os elementos-chave do processo proposto, são similares aos estabelecidos no processo de software UP.

A representação da visão geral do processo proposto, por meio de diagramas de pacotes, mostra o conjunto de elementos que o compõem, que são as Disciplinas. Estas, por sua vez, representam disciplinas da engenharia de software que necessitam ser especificadas em várias perspectivas diferentes. No diagrama de pacotes, mostrado na Figura 1, o nível de abstração é grande, mas o objetivo é mostrar que todas as Disciplinas possuem a mesma importância no processo de desenvolvimento de software. Ainda que a atenção maior esteja em Requisitos, devido ao risco de projeto inerente a esta disciplina, a preocupação para o desenvolvimento de sistemas de apoio à decisão estende-se pelas outras disciplinas, refinando e validando as expectativas dos usuários quanto ao conhecimento que será visualizado pelos tomadores de decisão.

No UPKDD a disciplina de Requisitos é considerada uma tarefa essencial, porque é responsável por mapear as expectativas dos usuários quanto ao conhecimento esperado, a fim de auxiliar no processo de extração de conhecimento, servindo como filtros do processo KDD e aumentando a compreensibilidade do conhecimento extraído por meio da Lista de Inferência.

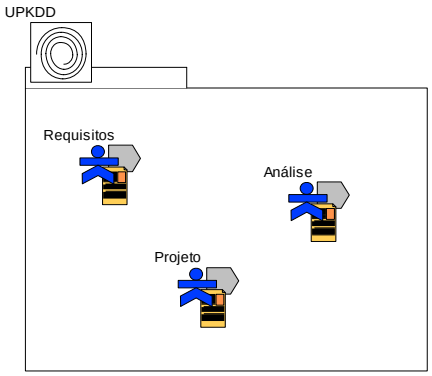

Figura 1: Diagrama de pacotes do processo UPKDD

Em suma, o UPKDD é fundamentado nos princípios da engenharia de software e divide-se em fases, que são Concepção e Elaboração. Da mesma forma, divide-se em disciplinas, que são Requisitos, Análise e Projeto. Este trabalho não considera as disciplinas de Implementação e Teste, nem as fases de Construção e Transição, devido à incerteza e dependência do ambiente de implantação da solução de apoio à decisão, e também porque não houve mudança significativa no *framework* UP nestas fases e disciplinas para este tipo de sistema.

As Disciplinas do UPKDD são compostas por Definições de Trabalho que, além de serem mecanismos de divisão semântica para as Atividades do Processo, também mostram como um elemento do modelo pode ser especificado em vários níveis de exigência, conforme a necessidade. As Definições de Trabalho explicam os relacionamentos entre os elementos do modelo.

Todos os Elementos-Chave (EC) definidos pelo UPKDD são classificados em três grupos seguindo os princípios do UP. Estes estão detalhados em Herden (2007) e no Quadro 1 são mostrados os ECs do UPKDD.

Quadro 1. Elementos-Chave do UPKDD.

Artefatos	Papéis	Atividades
Lista de Inferência	Engenheiro do Conhecimento	Analisar o Contexto de Dados
Lista de Informações de Negócio	Especialista de KDD	Analisar o Contexto de Ferramentas
Matriz de Barramento	Usuário Final (Tomador de Decisão)	Gerar Informações de Negócio
Modelo Dimensional		Projetar Arquitetura de DW
Descrição de Base de Dados Existente		Construir a Modelagem de Dados
Caracterização de Ferramentas Existentes		

Algumas vezes, as expectativas dos tomadores de decisão não são cabíveis para a implementação, mesmo em aplicações que têm como recursos as tecnologias analíticas. Portanto, compreender e *Analisar o Contexto dos Dados* existentes antes de construir uma aplicação KDD traz coesão ao desenvolvimento. Outro artefato proposto pelo UPKDD é Listar as Informações de Negócio, nele é possível identificar desde a motivação do usuário final quanto às suas expectativas de conhecimento, em nível alto de abstração, até o nível mais detalhado das oportunidades de busca na base de dados, como as tabelas e campos.

2.1 Disciplinas

Durante o desenvolvimento de um sistema de apoio à decisão, entende-se por requisitos a especificação da estrutura para tomada de decisão, agregada à especificação do desenvolvimento de aplicações tradicionais. Os detalhes desta Disciplina são vistos ainda em nível alto de abstração, em que esta é dividida em preocupações quanto às expectativas dos usuários, aos eventos dos dados existentes e às ferramentas que apoiarão a seleção, limpeza, extração e visualização das informações. A sequência de Definições de Trabalho na Disciplina de Requisitos é apresentada na Figura 2.

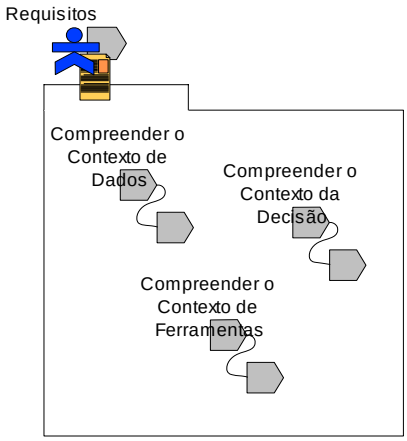

Figura 2. Diagrama de pacotes da Disciplina Requisitos do UPKDD.

A Disciplina Análise compreende as atividades que auxiliarão o desenvolvedor a desvendar a natureza subjetiva e complexa de um sistema de apoio à decisão que utilize o processo de KDD. Esta é rica em fatores não determinísticos como a situação em que se encontram os dados, instabilidade ou ilusão quanto às expectativas dos usuários finais e não formalismo das mudanças em geral. Esta Disciplina divide-se semanticamente em duas Definições de Trabalho, que são: Definir a Estrutura da Decisão e Compreender Arquitetura Dimensional.

A Disciplina Projeto refere-se à concepção do modelo de dados dimensional, envolvendo ferramentas para a construção do projeto arquitetural de DW. Compreender o Modelo Dimensional de Dados é a Definição de Trabalho dessa Disciplina.

2.2 Fases

Como o objetivo da Fase de Concepção no UP é estabelecer a viabilidade do sistema, construindo arquiteturas candidatas para implementação, o UPKDD incorpora a esta fase artefatos importantes para sistemas de apoio à decisão que utilizam o processo de KDD, que poderiam impactar o desenvolvimento.

A Figura 3 reúne os *workflows* de requisitos, análise e projeto, destacando os papéis relevantes que o desenvolvedor de aplicações KDD possui.

Neste diagrama de classes é mostrada a participação do Papel do Processo “Especialista de KDD” e também do papel de “Analista de Sistemas”, por meio do estereótipo <<*assistant*>>, que demonstra a importância do entendimento do contexto do sistema na concepção de aplicações KDD, sumarizadas na Lista de Inferências.

Este diagrama tem por propósito mostrar a visão abrangente da Fase de Concepção dos projetos e não a visão detalhista de qualquer *workflow* tratado pelo processo proposto.

Como o objetivo da fase de Elaboração no UP é estabelecer uma base arquitetônica sólida, por meio de requisitos funcionais descritos anteriormente, o UPKDD incorpora os modelos e documentos detalhados da arquitetura de DW, com o intuito de especificar tecnologias envolvidas, assim como perspectivas sobre os componentes técnicos da aplicação KDD. O fluxo desta fase pode ser encontrado em Herden (2007).

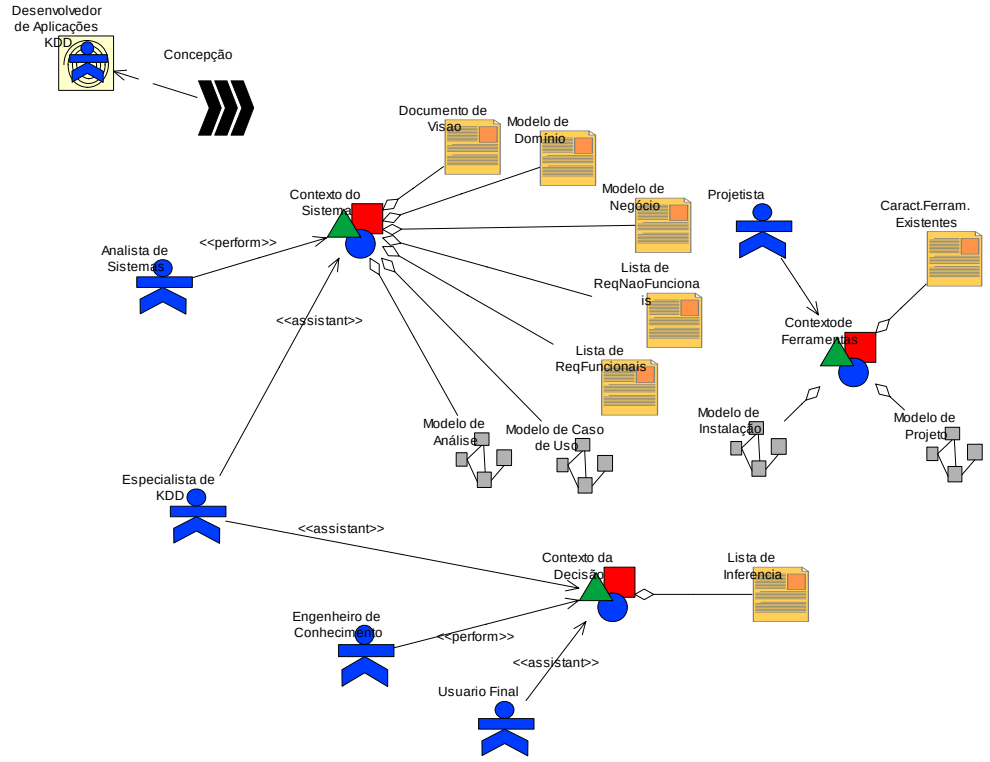

Figura 3. Diagrama de classes da Fase Concepção do UPKDD.

3. Método para Avaliação do Processo Proposto

Com o objetivo de verificar se as mudanças propostas pela utilização do UPKDD auxiliam no desenvolvimento de sistemas de apoio à decisão, foi realizada uma avaliação experimental com a instanciação do processo modelado. A avaliação experimental durou aproximadamente cinco meses, passou por quatro fases, que foram: definição, planejamento, operação e interpretação, conforme *framework* experimental de Basili, Selby e Hutchens (1985). O experimento foi em forma de estudo de caso e também 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.

Foram estabelecidas duas estratégias para o desenvolvimento de uma mesma solução KDD destinada a um Laboratório de Análises Clínicas:

(i) o grupo A, com seis desenvolvedores deveria apresentar a solução KDD apenas recorrendo à definição do UPKDD e as ferramentas prontas no mercado como OWB (*Oracle Warehouse Builder*), *Discoverer* e *Weka*; e

(ii) o grupo B, também com seis desenvolvedores deveria elaborar a especificação, análise e projeto do software, que abordasse as mesmas funcionalidades das ferramentas prontas, porém com a restrição de não utilizar as ferramentas na construção da solução KDD. O grupo B não poderia utilizar o UPKDD.

3.1. Fase Definição

Na fase de definição são preparados os conceitos da medição como: os objetivos, as questões e as métricas para o experimento. O objetivo geral foi avaliar se o UPKDD oferece Elementos-Chave (ECs) necessários para o desenvolvimento de aplicações de descoberta de conhecimento em banco de dados (KDD), do ponto de vista de seus desenvolvedores. E caracterizar quais são os ECs usados, úteis e que possuem descrição inadequada (quanto ao conteúdo descritivo).

A abordagem GQM (*Goal/Question/Metric*), proposta por Basili, Caldiera e Rombach (1994) tem a intenção de estruturar o processo experimental e serve como mecanismo que auxilia no processo de definição de métricas e hipóteses. Nesta abordagem, o experimento é direcionado por metas (*Goal*) específicas, que representam o cerne da investigação no nível conceitual. Estas por sua vez, são especificadas em questões (*Questions*), que caracterizam o caminho da avaliação (operacional). E, finalmente, as medidas (*Metrics*) servem para responder as questões (quantitativo).

O objetivo (*Goal*) da avaliação era: "Analisar o conjunto de ECs que são usados pelos desenvolvedores do UPKDD com o propósito de caracterizar com respeito à intersecção com os ECs do Conjunto Comum, do ponto de vista do desenvolvedor de aplicações KDD, no contexto de soluções de KDD."

No Quadros 3 são identificadas as hipóteses que o experimento testou e a suas associações com os objetivos de medição.

Quadro 3: Objetivos e Hipóteses utilizados na avaliação.

<table>
<tr><td rowspan="2">*Objetivo 1:* Quais são os ECs do UPKDD, que são usados.</td><td>*Questão 1*: Existem ECs do Conjunto Comum que não fazem parte do Conjunto de ECs usados do UPKDD?</td></tr>
<tr><td>*Métrica 1*: A lista de ECs do Conjunto Comum que não fazem parte do Conjunto de ECs usados do UPKDD.</td></tr>
<tr><td colspan="2">*Hipótese Nula (H0):* Os ECs usados pelos desenvolvedores do UPKDD são diferentes dos ECs considerados no Conjunto Comum como sendo fundamental para o desenvolvimento de soluções KDD.
Hipótese Alternativa (H1): Os ECs usados pelos desenvolvedores do UPKDD são similares aos ECs considerados no Conjunto Comum como sendo fundamental para o desenvolvimento de soluções KDD.</td></tr>
<tr><td rowspan="2">*Objetivo 2:* Quais são os ECs do UPKDD, que são usados - então mensurar os EC's que são considerados úteis e inúteis.</td><td>*Questão 2*: Existem ECs do Conjunto Comum e ECs usados do UPKDD que são considerados inúteis pelos desenvolvedores?</td></tr>
<tr><td>*Métrica 2*: A lista ECs do Conjunto Comum que fazem parte dos ECs usados do UPKDD e são considerados inúteis pelos desenvolvedores.</td></tr>
<tr><td colspan="2">*Hipótese Alternativa (H2):* No conjunto de ECs usados do UPKDD e que fazem parte do Conjunto Comum existem ECs que os desenvolvedores consideram úteis para o desenvolvimento de KDD.</td></tr>
<tr><td rowspan="2">*Objetivo 3:* Quais são os ECs do UPKDD que são usados e possuem descrição inadequada (quanto ao conteúdo descritivo)</td><td>*Questão 3*: Existem ECs do Conjunto Comum e ECs usados do UPKDD considerados úteis pelos desenvolvedores, cuja descrição deve ser modificada?</td></tr>
<tr><td>*Métrica 3*: A lista dos ECs do Conjunto Comum que fazem parte dos ECs usados do UPKDD e considerados úteis pelos desenvolvedores, cuja descrição deve ser modificada.</td></tr>
<tr><td colspan="2">*Hipótese Alternativa (H3):* No conjunto de ECs usados do UPKDD, que fazem parte do Conjunto Comum e considerados úteis para o desenvolvimento de soluções KDD, existem ECs cuja descrição não deve ser modificada para atingir o nível esperado pelos desenvolvedores.</td></tr>
<tr><td rowspan="2">*Objetivo 4:* Quais são os ECs que os desenvolvedores gostariam de usar, mas não usaram.</td><td>*Questão 4:* Existem ECs do Conjunto Comum que não fazem parte do conjunto de ECs usados do UPKDD, mas que os desenvolvedores gostariam de usar porque consideram úteis para o desenvolvimento de soluções KDD?</td></tr>
<tr><td>*Métrica 4:* A lista dos ECs do Conjunto Comum que não fazem parte do conjunto de ECs usados do UPKDD.</td></tr>
<tr><td colspan="2">*Hipótese Alternativa (H4):* No conjunto de ECs não usados do UPKDD existem ECs que os desenvolvedores gostariam de usar.</td></tr>
</table>

O uso da abordagem GQM também auxilia na identificação das hipóteses. Os objetivos foram avaliados de três maneiras diferentes, que são:

(i) estatística descritiva usando o cálculo da mediana e da moda nas respostas dos questionários, a fim de avaliar as variáveis de *Utilização, Utilidade e Adequação*;
(ii) teste binomial para destacar agrupamentos de ECs , a fim de gerar confiabilidade ao experimento; e,
(iii) teste estatístico não paramétrico qui-quadrado para testar as hipóteses.

3.2. Fase Planejamento

Esta fase é responsável pela elaboração do experimento, nela as hipóteses são formuladas, existe a seleção de variáveis, seleção dos participantes, preparação conceitual da instrumentação e considerações para a validade do experimento.

As hipóteses identificadas foram listadas no Quadro 3 e o procedimento de testes escolhido foi o teste não paramétrico qui-quadrado, com a margem de erro igual a 5% e grau de liberdade igual a 3,84. Já a variável independente escolhida foi a lista de ECs do Conjunto Comum, e as variáveis dependentes escolhidas foram a Utilização, Utilidade, e Adequação do EC. Durante a avaliação participaram dois grupos distintos (A e B), com seis desenvolvedores cada um, tendo como estratégia de comparação dos resultados a observação dos participantes durante a realização das suas atividades. A separação dos participantes em grupos foi de acordo com níveis semelhantes de experiência em desenvolvimento de sistemas.

O grupo A recebeu treinamento do processo UPKDD e seguiu todos os elementos-chave propostos, assim como utilizou documentos específicos, responsabilidades e atividades pré-determinadas. O grupo B não recebeu treinamento, permitindo que seus participantes buscassem soluções, documentos e definissem responsabilidades, conforme demanda.

A separação dos participantes em grupos permitiu a realização de reuniões com horários diferenciados, assim como inibiu a persuasão e preferências por metodologias de desenvolvimento de software individuais nos dois grupos. Devido à escolha por uma abordagem quantitativa, foi possível mensurar os efeitos da mudança proposta, por meio da comparação. A investigação foi direcionada para a análise das tarefas desempenhadas pelo grupo A, em comparação com qualquer sequência de tarefas realizadas pelo grupo B, a fim de verificar se o grupo B criaria artefatos semelhantes aos definidos pelo UPKDD.

O modelo de instrumentação escolhido foi um questionário, no qual a variável Utilização recebeu valores “não usado (0) ou usado (1)”, a variável Utilidade recebeu valores “não é útil (0) ou é útil (1)”, e por fim a variável Adequação recebeu valores “nível é adequado (0) ou nível não é adequado (1)”. Para a validação do experimento optou-se por observar os participantes durante a realização do estudo de caso, caracterizando o processo de avaliação como *on-line*. Os indivíduos participantes do estudo de caso foram alunos matriculados na disciplina de Banco de Dados, do Programa de Pós-graduação em Ciência da Computação da Universidade Estadual de Maringá, tendo a maioria deles experiência, com nível médio, em desenvolvimento de software. Portanto, a escolha dos participantes não foi de maneira aleatória, pois esta disciplina tem como um dos principais objetivos o ensino de tecnologias de *data warehousing*. O problema tratado foi real não simulado. Foi solicitado aos alunos, por meio de estudo de caso, a construção de um *data warehouse* juntamente com a construção de aplicações analíticas, a partir de uma base de dados relacional de um

sistema real, ou seja, não hipotético. O contexto possui caráter específico porque é focado na investigação comparativa de ECs usados pelos desenvolvedores com o Conjunto Comum de ECs.

3.3. Fase Operação

Na fase de operação os dados foram coletados e tabulados de maneira imparcial. Os questionários foram aplicados em dois momentos diferentes. No início do estudo de caso, com o intuito de traçar o perfil do grupo de participantes. As perguntas investigaram o nível e tempo de experiência em desenvolvimento e a formação do participante. E depois no final do estudo de caso para caracterização das dificuldades encontradas e do processo utilizado, assim como elaboração da análise dos casos em que houve interesse em usar outra sequência qualquer de atividades para o desenvolvimento da solução KDD.

3.4. Fase Interpretação

Esta fase é responsável por explicar os resultados do experimento, para a análise e interpretação dos resultados, optou-se inicialmente pelas medidas de tendência central do conjunto de dados de resposta, sendo calculadas a mediana e a moda de cada EC. Também foi utilizado o teste binomial para destaque de agrupamentos relevantes de ECs, e ainda foi aplicado o teste estatístico não paramétrico qui-quadrado para verificação de hipóteses.

Conforme Montgomery e Runger (2003), elaborar adequadamente a sumarização e apresentação de dados é fundamental ao bom julgamento estatístico. Neste sentido, as medidas de tendência central, como a mediana[2] e moda[3], organizam a amostra destacando os acontecimentos. Neste experimento, as medidas estatísticas como Mediana e Moda são calculadas para os valores de Utilização, Utilidade e Adequação. Os resultados são descritos na Seção 4 deste artigo.

Já as variáveis relevantes para o estudo são validadas pela aplicação do teste binominal, como: (1) quais são os ECs do UPKDD que são usados (variável Utilização); (2) quais são os ECs do UPKDD que são usados e considerados pelos desenvolvedores como úteis (variável Utilidade); (3) quais são os ECs do UPKDD que são usados e possuem descrição do seu conteúdo inadequado (variável Adequação); e (4) quais os ECs que os desenvolvedores gostariam de usar, mas não usaram (variável Utilização). O teste binomial foi utilizado no sentido amplo de análise e interpretação descritiva dos dados, destacando agrupamentos importantes de ECs.

A partir da visualização dos grupos de ECs foram testadas as hipóteses, que abrangem aspectos de similaridade e diferença dos ECs do UPKDD em relação ao Conjunto Comum de ECs. Para verificação de hipóteses optou-se pelo rigor matemático oferecido pelo teste estatístico qui-quadrado, que permite comparar frequências observadas e esperadas dada uma condição ideal estabelecida para o estudo (Montgomery e Runger, 2003). Assim, os fundamentos de aceitação ou rejeição das hipóteses estatísticas foram obtidos com grau de confiança em torno de 95%. Com a finalidade de observação do processo proposto a condição ideal é que um ECs seja usado, seja considerado útil e também que sua descrição não precise de modificação

[2] Mediana de uma amostra é uma medida de tendência central, que divide os dados em duas partes iguais, metade abaixo da mediana e metade acima. Se o número de observações for par, a mediana estará na metade da distância entre os dois valores centrais. Se o número de observações for ímpar, a mediana será o valor central.

[3] Moda da amostra é o valor da observação que ocorre com mais freqüência.

quanto ao conteúdo. Os cálculos da experimentação em engenharia de software, especificamente do processo proposto estão detalhados em Herden (2007).

4. Resultados e Discussão

Em geral, foi observado, por meio de questionários, que o grupo B mostrou a necessidade inerente do desenvolvedor de aplicações KDD por artefatos estabelecidos pelo UPKDD. Também se observou que os artefatos idealizados pelo UPKDD serviram de apoio aos dois grupos de participantes e, principalmente, o artefato "Caracterização de Ferramentas Existentes" fundamentou o software encomendado e qualificou as ferramentas escolhidas para o desenvolvimento da solução KDD.

4.1. Resultado da Estatística Descritiva (Mediana e Moda)

Cada EC foi avaliado por 12 desenvolvedores e a mediana para as variáveis (Utilização, Utilidade e Adequação) mostra que pelo menos 50% dos participantes utilizaram os quatorze ECs, considerando-os úteis e também adequados quanto ao seu conteúdo. De maneira semelhante, foi feito o cálculo da moda, sendo destacado o valor da observação que ocorre com mais frequência, observou-se que a situação que mais acontece é de grande utilização, utilidade e adequação dos ECs. Os grupos A e B declararam que independente dos ECs terem sido usados ou não, a maioria dos participantes considerou que estes são úteis para o desenvolvimento de sistemas de apoio à decisão.

4.2. Resultado do Teste Binomial

O teste binomial foi utilizado para destacar agrupamentos de ECs, conforme Quadro 4.

Quadro 4. Grupo de ECs do UPKDD que foram considerados Usados e também Úteis.

Elementos-Chave (EC) Usados e Úteis	Características
Lista de Inferência	- Mesmo os ECs tendo alto índice de utilização e utilidade, o detalhamento deve ser modificado porque foi considerado insuficiente para o entendimento. - Para todos os ECs não usados houve interesse dos participantes em usá-los. - Independente dos ECs terem sido usados ou não, a maioria dos participantes considerou que esses ECs são úteis.
Lista de Informações de Negócio	
Matriz de Barramento	
Modelo Dimensional	
Analisar o Contexto de Dados	
Projetar Arquitetura de DW	
Construir o Modelo de Dados	
Elementos-Chave (EC) Úteis e Mal Compreendidos	**Características**
Descrição de Base de Dados Existente	- Todos esses ECs exigem melhor descrição do seu conteúdo para o aumento de sua utilização, já que eles são considerados de alta utilidade. - Apesar desses ECs não terem sido usados pela maioria dos participantes, houve interesse em sua utilização.
Especialista de KDD	
Usuário Final (tomador de decisão)	
Gerar Lista de Informações de Negócio	
Elementos-Chave (EC) Úteis	**Características**
Engenheiro de Conhecimento	- Mesmo esses ECs terem sido considerados de alta utilidade, provavelmente a sua utilização poderia ter sido maior, se houvesse melhor descrição do seu conteúdo.
Analisar o Contexto de Ferramentas	
Elementos-Chave (EC) Compreendidos	**Características**
Caracterização de Ferramentas Existentes	- Esse EC foi compreendido e considerado altamente útil, mesmo não usado por alguns. - Mesmo para aqueles participantes que não usaram este EC, houve interesse em utilizá-lo.

4.3. Resultados do Teste de Hipóteses

A experimentação requer a decisão de aceitar ou rejeitar uma afirmação acerca de um parâmetro. Esta afirmação é chamada de hipótese e o procedimento de tomada de decisão sobre a hipótese é chamado teste de hipóteses. O teste de hipóteses escolhido foi o qui-quadrado, este é usado quando se deseja comparar frequências observadas com frequências esperadas (ideais). No procedimento para testar hipóteses, dentre outras atividades, destaca-se a identificação do valor qui-quadrado, e o julgamento das hipóteses, aceitando-as ou rejeitando-as. Após a aplicação do teste obteve-se o valor qui-quadrado para cada EC e o resultado está apresentado no Quadro 5.

Quadro 5. Valor qui-quadrado de cada EC do processo UPKDD.

Distribuições	Valor qui-quadrado	Elementos-Chave (ECs)
(+12): (- 0)	0,0	
(+11): (- 1)	1,04	
(+10): (- 2)	2,18	
(+ 9): (- 3)	3,43	
$X^2_{0,05;(1)}$	**3,84**	**qui-quadrado para o grau de liberdade 1. (verificação na tabela pré-estabelecida pela estatística)**
(+ 8): (- 4)	4,80	
(+ 7): (- 5)	6,32	Construir a Modelagem de Dados
(+ 6): (- 6)	8,00	Lista de Inferência, Lista de Informação de Negócio, Matriz de Barramento, Descrição de Bases de Dados Existentes Analisar o Contexto de Dados, Gerar Informação de Negócio
(+ 5): (- 7)	9,88	Modelo Dimensional, Usuário Final (tomador de decisão)
(+ 4): (- 8)	12,00	Especialista de KDD, Analisar o Contexto de Ferramentas
(+ 3): (- 9)	14,40	Caracterização das Ferramentas Existentes, Engenheiro de Conhecimento Projetar Arquitetura de DW
(+ 2): (-10)	17,14	
(+ 1): (-11)	20,31	
(+ 0): (-12)	24,00	

Seguindo o procedimento de teste, o último passo foi julgar as hipóteses. No julgamento observou-se que todos os ECs possuem valores de distribuição qui-quadrado (X^2_0) maiores que o valor qui-quadrado com grau e liberdade 1 e margem de erro de 5% ($X^2_{0,05,(1)}$). Portanto, todos os 14 ECs avaliados pelos 12 participantes possuem condições ideais em relação aos valores das variáveis: Utilização, Utilidade e Adequação. Isto prova que se determinado EC obteve como resposta *alguma* condição ideal, ele faz parte do Conjunto Comum de ECs. Logo, todos os 14 ECs usados do UPKDD fazem parte do Conjunto Comum.

Dessa maneira, a hipótese nula *(H0)* que diz: “Os ECs usados pelos desenvolvedores do UPKDD **são diferentes** dos ECs considerados no Conjunto Comum como sendo fundamental para o desenvolvimento de soluções KDD” é rejeitada.

Enquanto que a hipótese alternativa *(H1)* que diz: “Os ECs usados pelos desenvolvedores do UPKDD **são similares** aos ECs considerados no Conjunto Comum como sendo fundamental para o desenvolvimento de soluções KDD” é aceita.

Como visto anteriormente, todos os ECs usados do UPKDD fazem parte do Conjunto Comum. A partir disto, *H2* diz se os ECs foram considerados inúteis para o desenvolvimento de soluções KDD. Para a análise de *H2* é necessário entender os resultados obtidos pelo teste binomial. Nestes resultados todos os grupos de ECs apresentam a variação de 91,6% a 100% de Utilidade. Portanto, não existem ECs

considerados inúteis pelos desenvolvedores do UPKDD, devido a baixa porcentagem obtida, logo **aceita-se *H2***.

Estabelecido que todos os ECs usados do UPKDD fazem parte do Conjunto Comum e que de 90% a 100 % destes são úteis, *H3* procura ECs cuja descrição deva ser modificada, para atingir o nível esperado pelos participantes. Para analisar *H3* é necessário compreender os resultados do teste binomial. No Quadro 4 é possível visualizar que: apenas um EC (Caracterização das Ferramentas Existentes) apresenta necessidade de mudança do nível de descrição do seu conteúdo descritivo; para 10 ECs foi demonstrado que a modificação não é necessária e para 3 ECs a modificação é indiferente para esta amostra. Logo **aceita-se *H3.***

Para analisar a informação referente aos ECs não usados, mas que os desenvolvedores gostariam de usar, foi elaborada uma lista a partir dos valores das variáveis Utilização, Utilidade e Adequação = {0, X, X}. O resultado obtido foi que todos os ECs, que não foram usados, os desenvolvedores gostariam de usar. Conforme resultados obtidos na análise qualitativa dos dados, existem apenas ECs que os desenvolvedores não usaram e que gostariam de usar, logo **aceita-se *H4.***

5. Conclusões e trabalhos futuros

Este trabalho propôs um *Processo de Software para Aplicações KDD*, denominado UPKDD. A sequência de atividades do processo proposto foi modelada usando os recursos do metamodelo SPEM. Para a separação semântica de fases e *workflows* deste modelo de processo, foi escolhida a estrutura dinâmica e estática do UP. Nota-se que tanto papéis quanto artefatos e atividades, representados pelos diagramas utilizados no UPKDD, delimitam o trabalho a ser realizado, permitindo visualizar mecanismo de controle de projeto e produtividade para a equipe de desenvolvedores de aplicações KDD. Tanto para modelar o processo, usando uma notação específica como o SPEM, quanto para rever os elementos-chave do UP, foi necessário empenho considerável na configuração de um ambiente que permitisse o *design* do UPKDD e no entendimento dos detalhes do processo tradicional de desenvolvimento de software.

Por meio da avaliação do processo proposto, observou-se que o UPKDD mostrou-se adequado para aplicações deste tipo, pois os resultados quantitativos do estudo mostram alta utilização dos elementos-chave propostos comprovadamente similares ao Conjunto Comum de estudos de caso. O acompanhamento dos grupos A e B, durante o exercício de desenvolvimento de uma solução KDD para um laboratório de análises clínicas, comprovou a necessidade de um processo realmente focado nas características deste tipo de aplicação como o UPKDD.

As contribuições deste trabalho tiveram impacto nos elementos-chave do UP, para abranger a realidade de descoberta de conhecimento em bases de dados. Algumas visões foram agregadas ao processo de software tradicional, como a visão de condução do processo KDD, representado pela Lista de Inferências. Esta representa o direcionamento ou objetivo que norteia a busca por conhecimento em sistemas de apoio à decisão. Em complemento à Lista de Inferência, que trata de questões hipotéticas sugeridas pelos usuários finais, surgiu um modelo para mapear estas expectativas em situações realísticas dadas às condições das bases de dados existentes, mapeamento este denominado como Lista de Informações de Negócio.

Outra visão incorporada ao processo tradicional foi a da estrutura dos dados empresariais. Para este tipo de aplicação tem-se por base que os sistemas operacionais tratam perfeitamente situações transacionais da empresa. Portanto, durante as atividades de desenvolvimento não existe preocupação em construir o banco de dados relacional.

Esta visão nova está focada em como explorar os dados, tornando-os úteis, para isso é necessário transformá-los em uma estrutura dimensional, possibilitando prever eventos futuros usando dados históricos da empresa. Esta estrutura deve ser favorável ao uso de tecnologias analíticas que apoiam o tomador de decisões.

O próximo passo desta pesquisa é a aplicação do UPKDD em empresas de desenvolvimento de software que focam sistemas KDD, comparando os seus resultados com os obtidos no experimento conduzido.

Referências

Carmel, E. (1999) "Global Software Teams: Collaboration Across Borders and Time Zones", Prentice-Hall, EUA.

Acuña, S. T.; Ferré, X. (2001) *"Software Process Modelling"*, In: Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics: Information Systems Development (ISAS-SCI '01) – v.1, Orlando, Florida, USA.

Almeida, M. S. O. (2006) "*MGUP: RUP aplicado a jogos móveis"* 118f. Dissertação (Mestrado em Ciência da Computação) – Departamento de Informática (DIN), Universidade Estadual de Maringá (UEM), Maringá.

Álvares, P. M. R. S. (2001) *"A definição de um processo"*, In: WebPraxis: um processo personalizado para projetos de desenvolvimento para a web. Dissertação (Mestrado em Ciência da Computação) – Instituto de Ciências Exatas, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, p. 31-38.

Basili, V. R.; Caldiera, G.; Rombach, H. D. (1994) *"The goal question metric approach"*, In: Marciniak, J.J. (ed.). Encyclopedia of Software Engineering. New York: John Wiley & Sons, 1994. p.528-532.

_____; Selby, R. W. Jr.; Hutchens, D. H. (1985) "*Experimentation in software engineering*", Relatório Técnico/Científico 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, p. 37-57.

Date, C. J. (2003). *"Apoio à decisão"*, In:_____. Introdução a sistemas de banco de dados. 8. ed. Rio de Janeiro: Elsevier. p. 590-620.

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.

Freitas, A. A. (1998) *"On Objective Measures of Rule Surprisingness"*, In: Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD '98). Springer-Verlag, London, UK.

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).

_____. (2006) *"Data mining: concepts and techniques"*. 2. ed. San Francisco: Morgan Kaufmann Publishers. 770 p.

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.

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

Kimball, R.; Ross, M. (2002) *"Data warehouse toolkit: o guia completo para modelagem multidimensional"*. Rio de Janeiro: Campus. 494p.

Kruchten, P. (2003) *"Introdução ao RUP: rational unified process"*. Rio de Janeiro: Ciência Moderna. 255p.

Mitra, S.; Acharya, T. (2003) *"Data Mining: multimedia, soft computing, and bioinformatics"*. Hoboken: John Wiley and Sons, p. 14; p.401.

Montgomery, D. C.; Runger G. C. (2003) *"Inferência estatística para uma única amostra"*, In:_____. Estatística aplicada e probabilidade para engenheiros. Rio de Janeiro: LTC. p. 142-178.

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.

Rezende, S. O. et. al. (2005) *"Mineração de dados"*, In: Rezende, S. O. (Org.). Sistemas inteligentes: fundamentos e aplicações. Barueri: Manole. p.307-335.

Silberschatz, A.; Korth, H. F.; Sudarshan, S. (2006) "*Mineração e análise de dados*", In: _____. Sistemas de banco de dados. Rio de Janeiro: Elsevier. p. 485-508.

Sommerville, I. (2003) *"Engenharia de software"*. São Paulo: Addison Wesley. 592p.

Sousa, G. M. C. (2004) *"Adaptando o processo unificado para o desenvolvimento de software orientado a aspectos"*, In: Uma abordagem direcionada a casos de uso para o desenvolvimento de software orientado a aspectos*"*. Dissertação (Mestrado em Ciência da Computação) – Centro de Informática, Universidade Federal de Pernambuco (UFPE), Recife. p.69-107.