ELISA TULER DE ALBERGARIA

UM MODELO DE INTERFACE EXTENSÍVEL PARA SISTEMAS DE MINERAÇÃO DE DADOS POR REGRAS DE ASSOCIAÇÃO

Belo Horizonte
01 de julho de 2008

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

UM MODELO DE INTERFACE EXTENSÍVEL PARA SISTEMAS DE MINERAÇÃO DE DADOS POR REGRAS DE ASSOCIAÇÃO

Dissertação apresentada ao Curso de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ELISA TULER DE ALBERGARIA

Belo Horizonte
01 de julho de 2008

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Um Modelo de Interface Extensível para Sistemas de Mineração de Dados por Regras de Associação

ELISA TULER DE ALBERGARIA

Dissertação defendida e aprovada pela banca examinadora constituída por:

Profa. Raquel Oliveira Prates – Orientadora
Universidade Federal de Minas Gerais

Prof. Wagner Meira Junior – Co-orientador
Universidade Federal de Minas Gerais

Prof. Clarindo Isaís P. S. Pádua
Universidade Federal de Minas Gerais

Profa. Clarisse Sieckenius de Souza
Pontifícia Universidade Católica do Rio de Janeiro

Belo Horizonte, 01 de julho de 2008

Resumo

Atualmente, um dos grandes desafios da computação é o enorme volume de dados gerado pela facilidade de armazenamento e crescente uso de tecnologias em diversos contextos. A análise desses dados fornece apoio à tomada de decisões relacionadas a diversas áreas. Entretanto, pela grande quantidade de dados, essa análise tornou-se inviável de ser realizada sem o auxílio de técnicas computacionais. Nesse contexto, se apresenta a área de Mineração de Dados, que tem por objetivo a geração de conhecimento a partir de grandes volumes de dados. Ela abrange diversas técnicas, entre elas a de regras de associação, foco deste trabalho. Entretanto, um dos principais desafios para a ampla utilização desse tipo de sistema é a sua usabilidade, pois são vários os desafios de interação existentes. Esses sistemas normalmente são difíceis de usar, uma vez que requerem um conhecimento aprofundado de aspectos técnicos sobre o seu funcionamento.

Neste trabalho, com o objetivo de ampliar o uso de ambientes de mineração de dados, apresentamos, implementamos e avaliamos um modelo de interface extensível que permite criar novas interfaces de mais alto nível e específicas para um contexto, abstraindo o conhecimento técnico. Nossa proposta consiste em um modelo que define os componentes de um módulo de extensão a ser acoplado em sistemas de segunda geração, sistemas que envolvem diversas aplicações e abrangem diversas técnicas. Para isso ser possível, considera-se dois perfis de usuários: os especialistas e os leigos. Os usuários especialistas devem dominar tanto o domínio da aplicação quanto o sistema de mineração de dados (que requer conhecimento técnico específico). O objetivo do especialista consiste em criar um nível de abstração que permita que usuários leigos, que não possuam os conceitos técnicos envolvidos, possam usar o sistema em contextos e problemas específicos.

O modelo criado foi baseado na teoria da Engenharia semiótica, que considera que a interação consiste em um processo de comunicação entre o projetista e o usuário final. Nesse contexto, o modelo apresenta elementos em sua arquitetura que consideram esse aspecto e que permitem que os especialistas se tornem co-autores do sistema. Avaliações iniciais do modelo foram realizadas e uma implementação do mesmo foi desenvolvida, visando analisar sua viabilidade e utilidade. Os indicadores obtidos nas avaliações foram positivos, trazendo como grande benefício a possibilidade de ampliar a aplicação de técnicas mineração de dados, tanto em relação aos contextos de uso quanto ao público alvo.

Abstract

Currently, one of the main challenges of computing is the huge volume of data due to the storage facility and increasing use of technology in different contexts. The analysis of this data provides support for decisions in distinct areas. However, without efficient computational techniques it becomes unfeasible to analyze this large volume of data. Thus, data mining emerges as a promising field, since it allows for knowledge discovery from large volumes of data. Amongst the many techniques available for data mining, in this work we focus on association rules. Even though association Rules data mining systems are very popular they present users with a great challenge. These systems require users to have technical knowledge about data mining techniques in order to interact with them.

In this work we propose an extensible interface model which aims at widening the use of data mining systems. To do so, the model allows for a new abstract high level interface specific to a context to be created. This new high level interface abstracts the technical knowledge required, making it easier to interact with the system. Based on this model, an extensible module that can be added on to 2nd generation data mining systems can be developed. The model considers two distinct user profiles: the experts and final users. Expert users are those who not only have knowledge of the domain, but also of the required technical concepts to interact with the system, whereas final users have domain knowledge, but not data mining technical knowledge. Expert users interact with the extensible module and create a new high level interface specific to final users´ context with which they can interact.

The model is grounded on Semiotic Engineering theory, which perceives the interaction as designer-to-user communicative act. The model allows expert users to become co-authors of the message being transmitted by the systems, as they create new high level interfaces to final users. Preliminary evaluations of the model were executed and also a prototype was developed to provide indicators of the feasibility and utility of the model. The indicators pointed to the ability of the model to widen the use of the system to users who do not have data-mining technical knowledge at a low cost to expert users.

Agradecimentos

Em primeiro lugar, gostaria de agradecer a Deus por mais essa oportunidade. A todos os “bons fluidos” por me darem a energia necessária que precisei.

Aos meus pais, Braga e Inez, e minha irmã Elen pelo amor e apoio incondicional, tanto em relação aos estudos, mas também em todas as minhas decisões e situações vividas até hoje. Mãe, obrigada pelo conselho, estarei com ele em mente... "em tudo que fizer ou produzir tente sempre responder esta pergunta a si mesma: Em que isso pode melhorar o mundo, a humanidade?"

Agradeço ao meu marido, Leo, por todo incentivo e apoio dado, desde início desse desafio. Muitas vezes acreditou mais em mim do que eu mesma... Obrigada pela paciência, pelo carinho, pela compreensão e pelas longas conversas de incentivo.

Ao Diogo, por me ensinar o significado de amor incondicional... e por compreender a minha ausência em diversos momentos desse período. Obrigada filho, por me ensinar a cada dia algo novo.

A minha sogra, Conceição, por todo apoio, incentivo e compreensão. Obrigada pelas várias ajudas em relação ao Diogo e por estar sempre disposta a nos ajudar. Aos meus cunhados, Lu, Leandro e Francisco, pela amizade.

A minha amiga e orientadora Raquel Oliveira Prates, que se mostrou, além de ser uma excelente profissional, também uma pessoa maravilhosa. Obrigada não só pelo apoio e ensinamentos acadêmicos, mas todo carinho, paciência, atenção e amizade.

Ao Wagner Meira pelas orientações e pela amizade e carinho durante toda a trajetória que já caminhamos.

Aos professores do DCC que contribuíram para minha formação acadêmica e pessoal. Em especial, ao Clarindo pelas inúmeras oportunidades já oferecidas, incluindo o Synergia, meu atual trabalho. Ainda em relação ao Synergia, agradeço ao Robson pela oportunidade oferecida e confiança depositada.

A todos os alunos e usuários que contribuíram com o resultado deste trabalho e toda equipe envolvida no projeto Tamanduá, muito obrigada.

Aos familiares, famílias Tuler e Albergaria, pelo carinho e incentivo de todos. A todos meus amigos, pessoal do Synergia, do Speed, da UEMG, da UFMG, aos amigos criados em Lafaiete, obrigada pelo carinho de todos. As amigas, pelas conversas eletrônicas, companhia on-line em vários momentos em que estive dedicada a este trabalho.

Em especial, gostaria de agradecer ao Fernando Mourão pelo grande apoio dado durante o desenvolvimento desse trabalho, pela dedicação e amizade.

Ao meu eterno amigo Marcelo Maia, pela amizade e grande incentivo que me deu desde o início do curso de computação.

Sumário

Lista de Figuras

Capítulo 1

Introdução

Grandes instituições e empresas estão armazenando seus dados cada vez mais facilmente, gerando grandes bases de dados de natureza científica, comercial, governamental, etc. [Goldschmidt (2005)]. Este acúmulo de dados nas organizações e centros de pesquisa se tornou possível devido aos constantes avanços dos poderes computacionais.

A necessidade de transformar a "montanha" de dados armazenados em informações significativas é óbvia. Entretanto, a sua análise era demorada, dispendiosa, pouco automatizada e sujeita a erros, mal entendidos e falta de precisão [Newing (1996)].

Buscando analisar e extrair melhor conhecimento dos conjuntos de dados, surgiu uma área de pesquisa denominada KDD (*Knowledge Discovery in Databases*). "KDD é um processo de várias etapas, não trivial, interativo e iterativo, para identificação de padrões compreensíveis, válidos, novos e potencialmente úteis a partir de grandes conjuntos de dados" [Fayyad et al. (1996)].

O processo KDD envolve sistemas computacionais e suas etapas podem ser vistas na Figura 1.1. Inicialmente, o problema que se deseja resolver precisa ser analisado e compreendido, de forma a mapeá-lo em um problema de mineração de dados. Esta etapa normalmente é feita pelo próprio usuário, visto que é necessário que ele saiba como a mineração de dados pode ajudar a solucionar seu problema. A partir daí, os dados envolvidos no contexto do problema devem ser preparados para que possam ser minerados. Essa então é a segunda fase, onde é feito o pré-processamento que compreende a seleção e a preparação dos dados, sendo que alguns sistemas podem auxiliar os usuários nesta etapa. A preparação dos dados envolve tarefas como limpar a base, retirando "ruídos" e valores nulos, quando necessário, e discretizar dados, transformando números reais em intervalos de valores, por exemplo. Com os dados prontos, acontece a mineração propriamente dita (terceira etapa), quando os padrões são descobertos e explicitados. Esta fase é feita pelo sistema selecionado que utilizará uma técnica de mineração específica, escolhida pelo usuário de acordo com suas necessidades em relação ao problema existente. A quarta e última etapa é o pós-processamento e consiste na visualização dos resultados (ou modelos) e na sua interpretação, ou seja, na obtenção do conhecimento pelo usuário ao interagir com os mecanismos de visualização disponíveis no sistema [Nascimento (2005)].

A mineração de dados, apesar de ser uma das etapas do processo de KDD, é um termo comumente utilizado para referenciar todo processo. Ela surgiu há mais de uma década, como uma alternativa promissora para a análise desses grandes volumes de dados. Conjugando técnicas provenientes de diversas áreas, como estatística e banco de dados, a mineração de dados se diferencia das demais técnicas de análise pelo seu caráter exploratório. Se na estatística prevalecem os testes

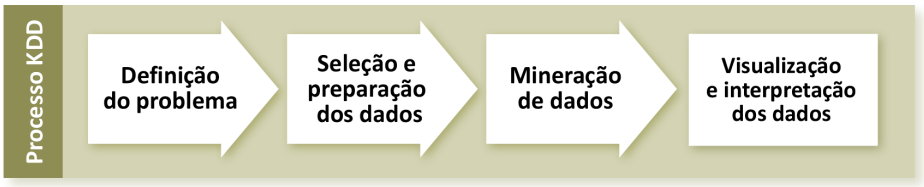

Figura 1.1: Etapas do processo KDD

de hipótese e em bancos de dados as consultas estruturadas, na mineração de dados prevalece a detecção automática de padrões. Ou seja, sem que se necessite formular previamente qualquer hipótese, toda a base de dados é analisada e uma série de padrões explicitados, fornecendo ao analista um conjunto de hipóteses potenciais que, dado o tamanho da base, só poderiam ser levantadas através da intuição.

Em termos históricos, os sistemas de mineração de dados podem ser apresentados em 4 gerações [Goldschmidt (2005)] [Piatetsky-Shapiro (1999)]. A primeira ocorreu na década de 80, em que as ferramentas focavam em uma tarefa específica como classificadores utilizando redes neurais, agrupamento (*clustering*) utilizando o algoritmo K-means [Ralambondrainy (1995)] [MacQueen (1967)] ou mesmo a visualização dos dados.

A segunda fase iniciou-se por volta de 1995, com ferramentas denominadas *suites*, dando suporte a mais de uma etapa do processo, possibilitando realizar diversas tarefas de descoberta e apresentando mais de um tipo de análise de dados. Como exemplos dessas ferramentas podemos citar: Clementine [Khabaza e Shearer (1995)], Tamanduá [Ferreira et al. (2005)] [Tamandua (2006)], WEKA [Weka (2006)] e DBMiner [Tutorial (2006)]. Entretanto, essas ferramentas normalmente requerem um grande conhecimento sobre técnicas específicas de mineração de dados por parte dos usuários para utilizá-las [Albergaria et al. (2006)].

De forma a tornar os sistemas mais "amigáveis" aos usuários, surgiu a terceira geração de sistemas no final da década de 90. Voltados para um contexto específico, os termos e conceitos utilizados tendem a ser mais próximos dos usuários, nos problemas que estão analisando. Entretanto, as ferramentas dessa fase ficam limitadas a um determinado problema e contexto. Um exemplo é o sistema voltado para detectar fraudes denominado HNC Software´s Falcon [Rainho (2001)].

A quarta geração consiste nas ferramentas de assistência ao processo de KDD, também chamadas IDA (*Intelligent Discovery Assistants*). Essas ferramentas buscam auxiliar os usuários no complexo processo de KDD, ajudando durante as tomadas de decisões entre as várias possibilidades de qual caminho seguir em uma determinada tarefa de mineração. Neste caso, os conceitos são apresentados e explicados aos usuários. Ou seja, os usuários são auxiliados no processo de aprendizagem e precisam compreender o processo para realizarem suas tarefas. Uma análise mais aprofundada dos sistemas em gerações é apresentada na seção 2.5.

Os sistemas mais amplamente utilizados são os de segunda geração, por cobrirem diversas aplicações e abrangerem diversas técnicas. Entretanto, encontramos em Albergaria et al. (2006) uma série de desafios de interação que ilustram as dificuldades dos usuários em relação a esses tipos de sistemas. Em Kriegel et al. (2007), também são levantadas algumas dificuldades existentes em relação à interação com sistemas de mineração de dados. Em geral, são conhecimentos técnicos que os usuários não possuem e muitas vezes não estão dispostos a adquirir.

Os sistemas de terceira geração abstraem os conhecimentos técnicos, mas para isso são limitados

a um tipo de problema específico. As interfaces são voltadas para os usuários, mas cada domínio demanda um novo processo de desenvolvimento, o que representa um alto custo.

Os sistemas da quarta geração ainda estão sendo pesquisados e prototipados. O objetivo destes sistemas é auxiliar o aprendizado em relação aos conceitos técnicos, sendo necessário de toda forma que os usuários aprendam o processo. Assim, eles podem facilitar o aprendizado, mas ainda requerem que o usuário esteja disposto a aprender os conceitos de mineração de dados.

Recentemente, pesquisadores têm levantando a necessidade de se criar sistemas que são fáceis de usar [Han et al. (2007)]. Porém, apesar das tentativas em abstrair os conceitos como nos sistemas de terceira geração ou apresentar os conceitos envolvidos, como no caso de sistemas de quarta geração, a usabilidade de sistemas de Mineração de Dados(MD) recentemente foi apontada em [Kriegel et al. (2007)] como um dos cinco grandes desafios da área.

Nesse sentido, o objetivo deste trabalho consiste em apresentar um modelo, aplicado e avaliado, baseado na teoria da Engenharia Semiótica [de Souza (2005)], apresentada no capítulo 3. Nossa solução consiste na proposta de um modelo de extensão a ser acoplado em sistemas de segunda geração que busca permitir a um grupo de usuários que utilizem esses sistemas sem que para isso seja necessário um entendimento (ou aprendizado) a fundo dos conceitos técnicos de mineração de dados envolvidos, sem no entanto restringir o amplo potencial de atuação dos sistemas de segunda geração. Isso em função da necessidade de que os sistemas sejam de ampla aplicação, mas que não demandem que os usuários precisem aprender os conceitos envolvidos em mineração de dados. Assim, existe a necessidade de sistemas que sejam intuitivos e aplicáveis a diversos contextos.

O modelo de extensão proposto envolve vários fatores, descritos no capítulo 4. Dentre as características, são considerados dois perfis de usuários: o **especialista**, que pode ser usuário mais experiente ou representante da equipe de design e o usuário **leigo**, que entende o contexto de aplicação, mas não as técnicas de mineração de dados.

O modelo pretende oferecer a possibilidade de sistemas de segunda geração serem extensíveis, de forma a inserir nesses tipos de sistemas a possibilidade de usuários especialistas criarem abstrações e, com isso, permitir que um maior número de usuários leigos possam utilizá-los. Isso porque sem o modelo, todos os usuários precisam ser especialistas, conhecendo os conceitos de mineração de dados e do contexto da aplicação. Com o modelo, um especialista pode criar abstrações para vários leigos, que só precisam entender do problema a ser analisado. Ou seja, o modelo proposto visa possibilitar que usuários especialistas definam perguntas interessantes, permitindo que pessoas que não conheçam os conceitos envolvidos em mineração de dados possam obter informações úteis para elas nos ambientes em que atuam.

Um exemplo de aplicação seria o dono de um determinado supermercado que deseja saber quais produtos são vendidos de forma conjunta nos sábados a noite. Como ele pode utilizar um sistema de mineração para responder à sua pergunta? Nesse caso, ele não conhece as técnicas de mineração de dados e não está disposto a estudá-las. A idéia então consiste em criar uma camada de abstração por usuários especialistas (ou mesmo pela equipe de *design*) através de mecanismos de extensão em sistemas de segunda geração. Essa camada criada consiste em uma interface de fácil interação para um conjunto de usuários finais de um determinado domínio, nesse caso a gerência do supermercado. Dessa forma, o gerente conseguiria executar as perguntas criadas pelos especialistas e obteria respostas sem que seja necessário conhecer os conceitos de mineração.

Então, poderia ser criada pelos especialistas uma pergunta da seguinte forma: "Quais os produtos mais vendidos no(a) <DIA DA SEMANA> juntamente com o produto <PRODUTO>?" (<DIA DA SEMANA> consiste na lista de dias possíveis e <PRODUTOS> lista os produtos exis-

tentes no supermercado). O gerente então iria escolher sábado e um determinado produto, como cerveja, e solicitar a resposta, que apresentaria a listagem dos produtos que responde a consulta realizada. Nesse caso, o gerente não precisou conhecer nenhum conceito envolvido no contexto da mineração, mas obteve a resposta que desejava.

Em nosso trabalho, o ambiente de aplicação será o sistema de mineração de segunda geração denominado Tamanduá [Tamandua (2006)] e em relação às técnicas, estamos focados neste trabalho na técnica de Regras de Associação, sendo ela bastante popular e de grande aplicação [Hipp et al. (2000)]. A seguir apresentamos como este trabalho está dividido.

1.1 Organização da dissertação

Este trabalho está organizado em mais cinco capítulos, além desta introdução. O capítulo 2 apresenta os conceitos de mineração de dados, aprofundando na técnica de mineração de regras de associação, foco deste trabalho. Apresentamos também os desafios identificados para um uso mais amplo dos sistemas de mineração de dados e soluções existentes para alguns destes desafios.

A fundamentação teórica do trabalho é apresentada no capítulo 3, em especial a teoria da Engenharia Semiótica [de Souza (2005)], juntamente com uma introdução a sistemas extensíveis. O modelo aqui proposto está descrito no capítulo 4, onde são apresentados seus objetivos, arquitetura e características.

A instanciação do modelo foi feita desenvolvendo um protótipo que está apresentado no capítulo 5, além da descrição do sistema de segunda geração utilizado para aplicar o modelo, o Tamanduá [Tamandua (2006)]. Nesse capítulo também são descritas avaliações realizadas, inclusive com a participação de usuários reais.

Para finalizar, as conclusões são apresentadas e discutidas no capítulo 6, além de contribuições e trabalhos futuros.

Capítulo 2

Mineração de dados por regras de associação

Este capítulo visa apresentar conceitos em mineração de dados, detalhando a técnica de mineração de regras de associação, que é o contexto onde o modelo desenvolvido é aplicado.

Mineração de dados surgiu da necessidade de extrair conhecimento e padrões de grandes bases de dados. Isso porque a análise de grandes quantidades de dados tornou-se inviável sem o auxílio de ferramentas computacionais [Goldschmidt (2005)]. Conforme apresentado no capítulo 1, a mineração é uma etapa do processo KDD, porém diversos autores referem-se à mineração de dados e ao processo KDD de forma indistinta. É na etapa de mineração que se realiza a busca efetiva por conhecimentos úteis e implícitos.

Assim, mineração de dados refere-se a uma forma automática e inteligente de analisar, interpretar e relacionar grandes quantidades de dados, tomando as informações obtidas como suporte para decisões nos negócios.

Mineração de dados apresenta diversas técnicas, que podem ser classificadas como preditivas ou descritivas. A mineração preditiva constrói modelos para a previsão das tendências e das propriedades de dados desconhecidos. Ela prevê dados não disponíveis a partir de dados disponíveis, podendo indicar diretamente uma descoberta (auxiliar uma decisão) ou servir como passo intermediário para uma descoberta mais complexa. Alguns tipos de inferência que podem ser citados como preditivas são classificação [Mitchell (1999)] e regressão [Weiss e Indurkhya (1998)].

A mineração descritiva descreve conceitos ou conjuntos de dados relevantes de forma concisa, discriminante e informativa. Representa a área de investigação nos dados que busca fatos relevantes, não-triviais e desconhecidos dos usuários, sem que existam hipóteses previamente elaboradas. Alguns exemplos são sumarização [Jiawei Han (2001)], clusterização [Berkhin (2002)] e as regras de associação [Agrawal et al. (1993)].

2.1 Regras de associação

Nesse trabalho, estamos focados na técnica de Regras de Associação, uma das técnicas mais populares, tendo uma grande variedade de aplicação [Hipp et al. (2000)]. Essa técnica tem a funcionalidade objetiva de encontrar correlações interessantes entre os itens de uma base de dados. A mineração de regras de associação foi introduzida por Agrawal et al. em [Agrawal et al. (1993)]. A técnica consiste em encontrar conjuntos de itens que ocorram simultaneamente e de forma

freqüente em um banco de dados. Assim, muitos algoritmos relacionados à tarefa de regras de associação baseiam-se na seguinte propriedade [Goldschmidt (2005)]: um conjunto somente pode ser freqüente se todos os seus subconjuntos forem freqüentes.

A aplicação mais conhecida de regras de associação consiste em auxiliar na compreensão dos hábitos de compra dos clientes de um supermercado, que ficou conhecida como análise do carrinho de compras. A idéia era descobrir como as vendas de alguns produtos influenciavam nas vendas de outros, para que se pudesse planejar melhor as promoções, organizar de forma mais conveniente a disposição das prateleiras e avaliar o impacto que a descontinuidade nas vendas de um produto poderia provocar nas vendas de outros. Por exemplo, através dessa técnica é possível descobrir quais produtos são vendidos de forma conjunta. Assim, o gerente de um supermercado pode descobrir, por exemplo, que arroz e óleo são mais vendidos, de forma conjunta, aos sábados pela manhã. A mesma aplicação pode ser estendida à sites de comércio eletrônico, por exemplo, onde permite descobrir se existe uma grande afinidade na preferência de seus compradores [Cortes (2002)].

A generalidade da mineração de regras de associação permitiu, no entanto, que ela fosse utilizada para as mais diversas aplicações. Exemplos de aplicações reais incluem: análise de crédito no setor financeiro, detecção de fraudes na área de seguros, database marketing (generalização da análise do carrinho de compras), detecção de intrusos na área de segurança de redes, leilões eletrônicos, etc. Em última instância, a mineração de regras de associação é aplicável sempre que se deseja encontrar algum tipo de correlação dentro de uma base de dados.

Algoritmos de mineração de regras de associação geram um conjunto de regras que devem ser interpretadas pelos usuários. Uma **regra de associação** representa uma relação entre dois ou mais itens de uma base de dados. Considere, por exemplo, a regra apresentada a seguir:

[PÃO], [MANTEIGA] => [LEITE] (30.00, 60.00)

O conjunto dos itens do lado esquerdo da regra (pão e manteiga) é chamado de **antecedente** e o conjunto dos itens do lado direito da regra (leite) é chamado de **conseqüente**. Essa regra mostra a relação que existe entre a compra de pão, manteiga e leite em uma padaria hipotética e deve ser lida da seguinte forma: trinta por cento das compras realizadas pelos clientes da padaria incluem pão, leite e manteiga; e das compras que incluem pão e manteiga, sessenta por cento também incluem leite. Um exemplo de um conjunto de vendas ilustrando um contexto onde essa regra pode ter sido gerada pode ser visualizado na tabela 2.1.

Número da Compra	Pão	Manteiga	Leite
1	sim	sim	sim
2	não	sim	não
3	não	sim	sim
4	sim	sim	sim
5	sim	não	sim
6	sim	sim	não
7	sim	sim	não
8	sim	sim	sim
9	não	não	não
10	não	não	não

Tabela 2.1: Cadastro de vendas de uma padaria

O primeiro valor apresentado na regra (30.00) corresponde ao **suporte** da mesma. O suporte[1] representa a freqüência de ocorrência do evento, formado pela união entre o antecedente e o conseqüente da regra e dá uma medida da sua significância estatística. No nosso exemplo, observamos que em 3 das 10 transações ocorreram as compras de pão, manteiga e leite simultaneamente. Sendo assim, temos que o suporte da regra é de 30%.

O segundo valor (60.00) que aparece entre os parênteses corresponde a confiança da regra. A **confiança** representa a freqüência relativa (ou probabilidade condicional) entre a ocorrência do evento no conseqüente e a ocorrência do evento no antecedente. Podemos dizer que a confiança dá uma medida do poder de previsão da regra: se já soubermos que uma determinada compra inclui pão e manteiga, e arriscamos dizer que ela também incluirá leite, qual será a nossa chance de acerto? Pela regra acima, a nossa chance de acerto será de 60%. Os termos confiança, freqüência relativa e probabilidade condicional podem ser usados de forma intercambiável. Ela é calculada da seguinte forma:

$conf(A \rightarrow B) = \mathrm{P(B \mid A)} = \frac{P(AeB)}{P(A)} = \frac{suporte(A \rightarrow B)}{suporte(A)}$

onde P(B | A) é a probabilidade de B ocorrer, visto que A ocorreu, que é calculada como a probabilidade de A e B, dividida pela probabilidade de A.

Utilizando o exemplo, temos:

- Suporte de A (pão e manteiga): 50% (aparecem em 5 das 10 transações)
- Suporte de A $\rightarrow$ B (pão, manteiga e leite juntos): 30% (aparecem em 3 das 10 transações)
- Confiança = 30/50 = 60%

Além das medidas de suporte e confiança, existem outras medidas de interesse que auxiliam na análise das regras de associação. A seguir são apresentadas as definições de *leverage*, *lift* e *conviction*.

O ***leverage*** é uma medida de interesse que relaciona o suporte esperado com o que é realmente obtido. Por exemplo, existindo dois dados, a probabilidade de sair o número 6 em um dado é 1/6, já a probabilidade de sair o número 6 nos dois dados é dada por $1/6 * 1/6 = 1/36$. Ou seja, dados os eventos A e B, temos que a probabilidade de ocorrer os eventos A e B juntos é: P(A).P(B).

Assim, no cálculo do *leverage*, primeiro calcula-se os suportes de A e B separadamente. Posteriormente, esses valores são multiplicados gerando o valor esperado. Calcula-se também o suporte de A e B juntos (os itens ocorrendo simultaneamente), encontrando o valor obtido. O leverage é a diferença entre os valores encontrados:

$\mathrm{leverage(A \rightarrow B)} = (P(A \text{ e } B) - (P(A)P(B)))$

leverage = suporte obtido − suporte esperado

O ***lift*** é uma medida de interesse que relaciona a confiança esperada com a obtida, sendo semelhante ao *leverage*. É uma das medidas mais utilizadas para avaliar dependências. Dada uma regra de associação A $\rightarrow$ B, o *lift* indica o quanto mais freqüente torna-se B quando A ocorre.

[1] A importância em relação ao valor da suporte pode variar de acordo com o contexto. Por exemplo, regras que apresentam um suporte abaixo de um determinado valor podem ser consideradas pouco relevantes em análise de carrinhos de compras, já que se busca alto grau de relacionamento entre produtos. Já em detecção de fraudes, a exceção pode ser o dado procurado e, neste caso, regras com valor de suporte baixo serão relevantes.

Ela pode ser explicada através do exemplo a seguir. Dadas as transações apresentadas na tabela 2.2, vamos analisar a regra regra PÃO => MANTEIGA.

Número da Compra	Pão	Manteiga
1	*sim*	sim
2	*sim*	sim
3	não	sim
4	não	não
5	*sim*	não
6	não	não
7	*sim*	sim
8	não	não
9	não	sim
10	*sim*	não

Tabela 2.2: Exemplificação do Lift

Considerando todas as compras realizadas, temos que em 50% das transações o item manteiga foi comprado. Quando consideramos a regra PÃO => MANTEIGA, reduzimos nosso domínio apenas às transações onde houve a compra de pão. Devemos então verificar em quantas delas houve o consumo de manteiga. Em 5 dessas transações houve o consumo de PÃO e em 3 delas também foi consumido o item MANTEIGA. Sendo assim, temos uma confiança de $3/5 = 60\%$. Vimos assim, que a confiança obtida com a regra foi maior que o suporte inicial esperado para o pão, o que pode indicar que o consumo de pão está relacionado ao de manteiga. A fórmula para calcular o lift é:

$$lift(A \rightarrow B) = lift(B \rightarrow A) = \frac{P(AeB)}{P(A)P(B)} = \frac{conf(A \rightarrow B)}{suporte(B)} = \frac{conf(B \rightarrow A)}{suporte(A)}$$

onde $conf(A \rightarrow B) = \frac{P(AeB)}{P(A)}$

Ou seja, considerando nosso exemplo, temos o seguinte cálculo:

$$\text{lift} = \frac{conf(pao \rightarrow manteiga)}{suporte(manteiga)} = \frac{(60)}{(50)} = 1,2$$

Quanto maior o *lift*, maior é a possibilidade de que A e B juntos em uma transação não seja um fato aleatório, e sim que tenha sido causado por alguma relação.

Calculando em termos de porcentagem, temos a seguinte expressão:

$$(lift - 1)100 = (1,2 - 1)100 = 20\%$$

Ou seja, a regra apresenta uma confiança 20% acima da esperada.

O ***conviction*** (convicção) é uma medida de interesse que relaciona a regra complementar a que está sendo analisada, onde a regra contendo a negação do conseqüente pode ser muito mais expressiva. Ela quantifica o impacto da regra quando comparada com a sua regra complementar (o conjunto de regras onde o conseqüente é invertido).

Dada uma regra A $\rightarrow$ B, *conviction* é a freqüência com que A ocorre sem B, dividida pela freqüência com que as duas ocorrem juntas.

Primeiramente, calcula-se o *lift* da regra complementar (negação da regra que está sendo analisada). Posteriormente, seu valor é invertido: $\frac{1}{lift}$.

Para analisar os valores obtidos, temos as seguintes regras:

- quando *conviction* é igual a 1, significa que a regra e o seu complemento tem igual valor,
- quanto maior o valor de *conviction*, mais forte é a própria regra,
- se o valor do *conviction* for menor que 1, deve-se analisar as regras complementares.

A fórmula de cálculo de *conviction* é:
$\text{conviction}(\text{A} \rightarrow \text{B}) = \frac{P(A)P(notB)}{P(AandnotB)} = \frac{(1-supp(B))}{(1-conf(A \rightarrow B))}$

Cada medida de interesse deve ser analisada de forma complementar à análise dos resultados obtidos em um processo de mineração de regras de associação. Normalmente, as medidas mais utilizadas são suporte e confiança, pois o entendimento das mesmas é mais simples, sendo assimilada com mais facilidade pelos usuários [Albergaria et al. (2006)].

2.2 Tarefa de mineração de regras de associação

A aplicação da técnica de regras de associação é ampla e abrange diversos contextos. Porém, independente do contexto, em um sistema de mineração de segunda geração são necessários alguns passos para a criação de tarefas de mineração de dados. A seguir serão ilustrados os passos a serem seguidos utilizando o sistema de segunda geração denominado Tamanduá [Tamandua (2006)].

O primeiro passo a ser realizado pelo usuário consiste em criar uma tarefa de mineração. As primeiras informações solicitadas são nome e descrição para a tarefa, conforme ilustra a Figura 2.1. Posteriormente, o usuário precisa escolher a base a ser minerada juntamente com os atributos da mesma (Figura 2.2). É necessária também a escolha do algoritmo a ser utilizado, além dos valores dos parâmetros que serão utilizados, que no caso de regras de associação são suporte e confiança (Figura 2.3).

A tarefa então deve ser executada e os resultados são apresentados ao usuário. A tela dos resultados do sistema Tamanduá é apresentada na Figura 2.4 em que mostra o conjunto das regras obtidas na mineração. A tela apresenta as seguintes informações:

1. Filtros que podem ser utilizados para escolher os atributos presentes nas regras;
2. Possibilidade de mudança das medidas de interesse para visualização gráfica das regras;
3. Matriz de medidas de interesse, onde cada ponto é uma regra (ou um conjunto de regras com os mesmos valores nas medidas de interesse);
4. Detalhe de uma regra. Ao clicar em um dos pontos do gráfico (regra ou conjunto de regras) são apresentadas informações detalhadas na lateral.

Como trata-se de um processo iterativo, ao visualizar os resultados obtidos, o usuário pode sentir necessidade de mudar atributos, parâmetros ou filtros, tendo que executar novamente a tarefa. Para realizar essa interação, o usuário deve conhecer bem os conceitos envolvidos e o impacto de cada mudança que pode realizar.

A seguir são apresentados problemas que surgem na interação dos usuários com sistemas de segunda geração de forma geral.

Figura 2.1: Tela de criação de uma tarefa de mineração do sistema de segunda geração Tamanduá (Dados da Tarefa)

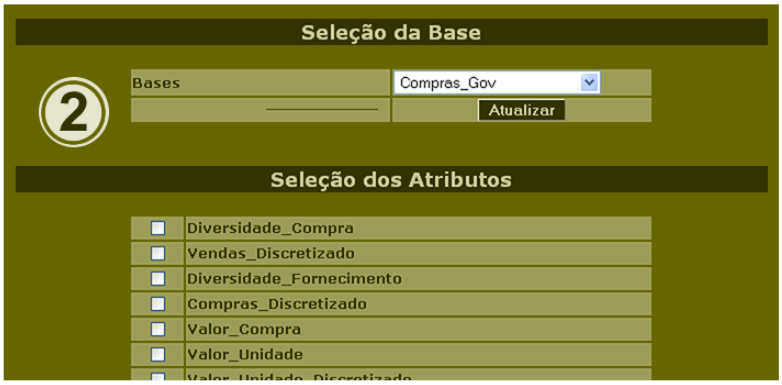

Figura 2.2: Tela de criação de uma tarefa de mineração do sistema de segunda geração Tamanduá (Seleção Base/Atributos)

2.3 Desafios no uso de aplicações de segunda geração

Recentemente, foi apresentado em [Kriegel et al. (2007)] que um dos desafios em mineração de dados consiste em aumentar a usabilidade de sistemas de Mineração de Dados (MD). As dificuldades experimentadas pelos usuários se distribuem ao longo do processo de mineração, desde a definição de parâmetros para mineração até sua visualização. Isso envolve configurar uma série de parâmetros, em um processo iterativo que envolve ajustar os resultados obtidos, selecionar e interpretar regras resultantes [Albergaria et al. (2006), Hofmann et al. (2000), Kriegel et al. (2007), Mei et al. (2006)]. O impacto dos problemas no uso do sistema é grave tanto para o usuário (que pode ser levado a interpretar erroneamente o resultado, não obtendo o conhecimento desejado), quanto para os responsáveis pelo sistema (o usuário pode desistir de utilizar o sistema).

Analisando a dificuldade dos usuários na interação dos sistemas disponíveis atualmente, encontramos em [Gonçalves (2001)] um estudo da aplicação de algumas ferramentas de mineração de dados no contexto de uma rede de supermercados. Esses sistemas podem ser considerados de segunda geração e os usuários foram gerentes dos supermercados, não especialistas em mineração de dados. Como resultado da avaliação, chegou-se à conclusão de que as ferramentas analisadas

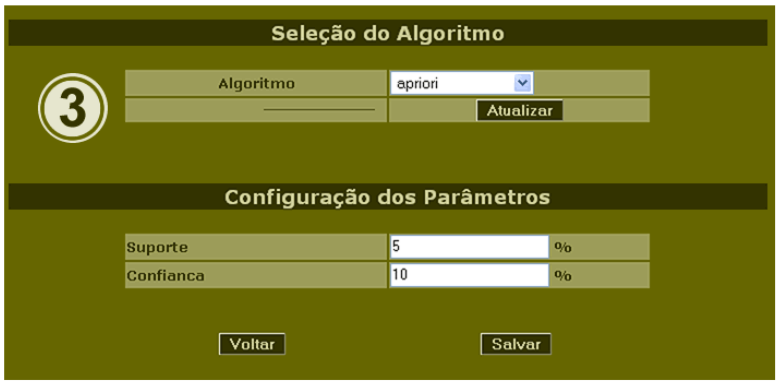

Figura 2.3: Tela de criação de uma tarefa de mineração do sistema de segunda geração Tamanduá (Seleção Algoritmo/Parâmetros)

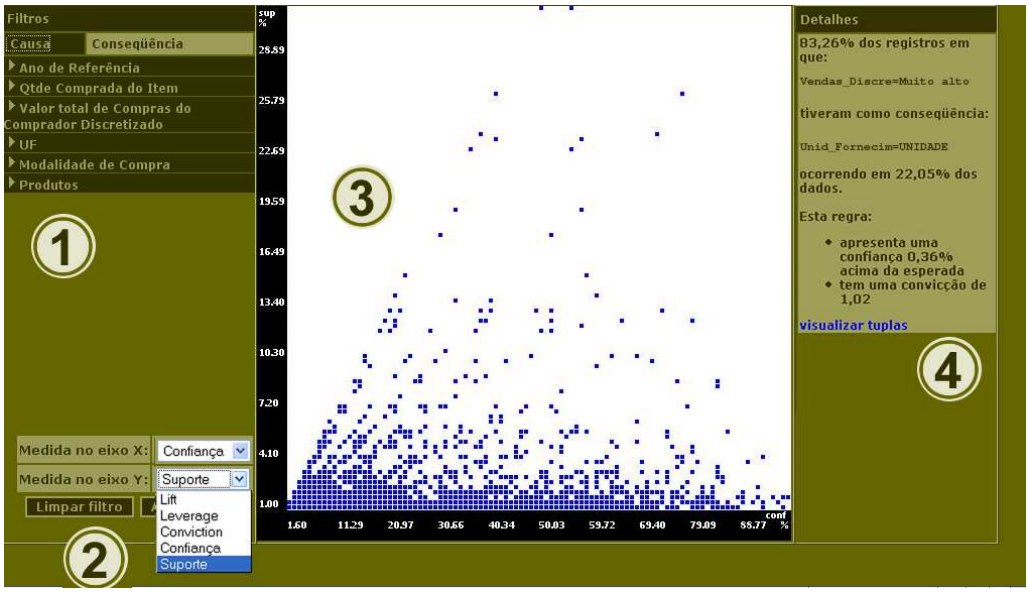

Figura 2.4: Tela de visualização dos dados do sistema de segunda geração Tamanduá

não auxiliaram os tomadores de decisão da empresa. Isto ocorreu pelo fato de não conseguirem utilizá-las de maneira eficaz, não trazendo assim nenhum benefício ao serem usadas. A dificuldade enfrentada pelos usuários ocorreu, em grande parte, em função da linguagem e interface do sistema que não faziam parte do domínio dos usuários.

Da forma semelhante, em [Albergaria et al. (2006)] são levantadas algumas dificuldades de interação dos usuários. A seguir estão apresentados alguns desafios de interação com sistemas de segunda geração de mineração de dados. Em geral, estes desafios podem ser relacionados aos termos técnicos empregados e aos conceitos envolvidos durante todo o uso dos sistemas que não fazem parte do domínio dos usuários.

2.3.1 Definição dos parâmetros de entrada da mineração

Um primeiro desafio compreende a definição dos parâmetros de entrada da mineração. Os algoritmos de mineração de regras de associação geralmente exigem que o usuário defina alguns

parâmetros iniciais para que eles possam ser executados. Os dois parâmetros mais tradicionais desses algoritmos são suporte e confiança mínimos. O usuário deve fornecer o valor mínimo de suporte que uma regra deve apresentar para que ela seja gerada, o mesmo valendo para a confiança. Estes conceitos não fazem parte (normalmente) do domínio do usuário, e, além disso, os valores mais adequados para esses parâmetros dependem da base de dados que vai ser minerada e do tipo de conhecimento desejado pelo usuário. Desta forma, a definição destes parâmetros não é intuitiva e depende da experiência do usuário tanto com a base de dados, quanto com o sistema de mineração.

2.3.2 Seleção dos atributos

Um aspecto relacionado à geração das regras que merece ser mencionado refere-se à escolha dos atributos a serem minerados na base de dados. O problema aqui se refere à escolha dos atributos pelos usuários, visto que em muitas bases de dados há diversos atributos que são redundantes ou parcialmente redundantes. Por exemplo, numa base de compras os atributos "código do produto" e "nome do produto" em geral são redundantes, já que cada código corresponde a um único produto (e.g. o código "123" corresponde ao produto "Mouse XYZ") . Já os atributos "nome do produto" e "categoria do produto" são parcialmente redundantes, já que cada produto é de uma única categoria (e.g. o produto "Mouse XYZ" pertence à categoria "Periféricos"). Quando o usuário seleciona atributos redundantes ou parcialmente redundantes, o sistema pode gerar regras óbvias, como as seguintes:

- [Código=123] → [Nome=Mouse XYZ] (100.00, 1.00)
- [Nome=Mouse XYZ] → [Categoria=Periféricos] (100.00, 1.00)

É óbvio que 100% dos produtos de código "123" são "Mouse XYZ", assim como é óbvio que 100% dos "Mouse XYZ" sejam "Periféricos". Como o sistema não tem como saber que os atributos são redundantes, essas regras irão aparecer em destaque, já que possuem uma confiança alta e um *lift* também alto, sendo o *lift* dado pela razão entre a confiança da regra e a confiança que seria esperada. O *lift* da primeira regra, por exemplo, tem valor 100, indicando que a confiança da regra é 100 vezes maior que a freqüência do conseqüente. Ou seja, o fato de sabermos que o código do produto em uma determinada compra é igual a "123" aumenta em 100 vezes a chance do nome do produto na mesma compra ser "Mouse XYZ", o que é óbvio.

2.3.3 Análise das regras de associação resultantes da mineração

O resultado de um sistema de mineração de dados por regras de associação é um conjunto dessas regras. No entanto, o conceito de regras de associação não faz parte do domínio do usuário e deve ser aprendido por ele para que possa fazer uso do sistema. O usuário deve entender que cada regra de associação representa uma possível correlação entre itens de uma base de dados. Possível porque o fato de existir uma regra de associação entre dois ou mais itens não significa necessariamente que eles estejam correlacionados. Vamos considerar a regra abaixo:

[Pão], [Manteiga] → [Leite] (80.00, 50.00)

Esta regra indica uma possível correlação entre a compra de pão e manteiga e leite. Como vimos, ela nos diz que os itens pão, manteiga e leite são comprados juntos com uma freqüência de 50%, e que 80% das compras que incluem pão e manteiga também incluem leite. Esta última porcentagem corresponde também à chance de acerto de uma previsão da compra de leite dado que já ocorreu a compra de pão e manteiga. Se o usuário não compreender corretamente os conceitos envolvidos em uma regra de associação, ele corre o risco utilizá-las de forma equivocada ou não conseguir atingir o objetivo que tinha ao utilizar o sistema.

2.3.4 Seleção do subconjunto de regras

Além de permitir ao usuário visualizar as regras geradas e suas características, o sistema de mineração de regras de associação deve também permitir ao usuário selecionar um subconjunto de regras que seja mais interessante para ele. Para isso, o usuário deve definir quais itens o interessam, em que lado da regra ele quer que um determinado item esteja presente ou qual o número de itens uma regra deve ter no antecedente ou no conseqüente para ser considerada interessante. Assim, o usuário deve entender não apenas a estrutura da regra (e.g. que a regra é formada por um antecedente e um conseqüente), mas também o que significa um item estar presente de um lado ou do outro.

2.3.5 Seleção das métricas de interesse

Valores de suporte e confiança altos não necessariamente indicam uma correlação entre os itens. Para avaliar essa correlação, são necessárias outras métricas de interesse. Para ser capaz de utilizá-las, o usuário deve antes aprendê-las, uma vez que elas também não fazem parte do seu domínio.

Na literatura, são encontradas dezenas dessas métricas, algumas mais adequadas a determinadas situações que outras. Para entendermos a utilização destas regras, vejamos o *lift*, já apresentado anteriormente. O *lift* dá uma medida do quanto a confiança de uma regra é surpreendente em relação ao que era esperado. Uma confiança de 80%, por exemplo, na regra ([Pão], [Manteiga] → [Leite] (80.00, 50.00)]) indica que 80% das compras que incluíram pão e manteiga também incluíram leite. Embora essa confiança pareça alta, não podemos afirmar isso com certeza sem olharmos a freqüência da compra de leite na base de dados. Se 80% de todas as compras efetuadas na padaria incluíram leite, então a confiança de 80% já era esperada, e a regra não teria trazido nenhuma informação surpreendente. Por outro lado, se apenas 40% de todas as compras efetuadas na padaria incluíram leite, então a confiança de 80% é o dobro da esperada, indicando que a compra de pão e manteiga influencia positivamente na compra de leite, o que é uma informação interessante. O *lift*, conforme já apresentado, é dado pela razão entre a confiança da regra e a confiança que seria esperada. Se a confiança esperada era de 80% e a confiança da regra foi de 80%, o lift é 1. Da mesma forma, se a confiança esperada era de 40% e a confiança da regra foi de 80%, o *lift* é 2. Quanto mais o lift divergir do valor 1, maior será a intensidade da correlação expressa pela regra e mais surpreendente ela será. Esse exemplo ilustra bem a necessidade de se analisar mais de uma medida de interesse antes de se tomar qualquer decisão ou tirar conclusões dos resultados apresentados.

***

Depois de apresentados alguns desafios, a próxima seção ilustra uma tarefa de mineração de dados (baseada nos passos apresentados na subseção 2.2), apresentando alguns pontos de dificuldades de interação.

2.4 Cenário de uso

A seção 2.2 apresentou os passos para se realizar uma tarefa de mineração de regra de associação e na seção 2.3 foram ilustrados desafios de interação em sistemas de segunda geração. Nessa seção será apresentado um cenário de uso ilustrando uma real tarefa de mineração sendo executada.

Cenários [Carroll (2000)] foram definidos como plausíveis e detalhadas narrativas textuais que descrevem uma situação específica. Eles têm sido usados em diferentes fases de concepção do software e sua principal contribuição é permitir uma visão mais ampla da utilização do sistema. Embora não seja real, é uma situação plausível, baseada em experiências reais. A seguir, será apresentado um cenário que ilustra uma aplicação, ilustrando uma tarefa de mineração de regras de associação sendo executada. O cenário apresentado é baseado no contexto de auditoria de compras governamentais e o sistema de segunda geração utilizado foi o Tamanduá [Tamandua (2006)].

2.4.1 Tarefa de auditoria

O setor de auditoria do governo resolveu verificar se existiam indicativos de fraudes em compras realizadas pelos órgãos públicos. Pedro, funcionário do setor de auditoria, achou que seria interessante realizar esse trabalho e resolveu utilizar técnicas de mineração de dados, apesar de não conhecer profundamente os conceitos envolvidos.

A primeira tarefa realizada por Pedro foi identificar os fenômenos fraudulentos que gostaria de analisar. Resolveu então focalizar a busca em três deles, listados abaixo:

- **Favorecimento:** seleção de fornecedores por meios não previstos em lei. (organizações públicas podem agir somente no limite do que é previsto em lei, enquanto organizações privadas podem fazer tudo o que não é proibido em lei)
- **Formação de Cartel:** tabelamento de preços de um tipo de produto por parte dos fornecedores.
- **Super-faturamento de compras:** Preços médios pagos para determinados produtos muito acima do preço de mercado.

Buscando identificar se houve indícios de fraude, Pedro determinou algumas premissas relacionadas aos fenômenos que ele selecionou:

- **Favorecimento:** Um fornecedor não é capaz de ganhar todas as licitações de um mesmo tipo de produto durante um ano inteiro sem ser favorecido.
- **Formação de Cartel:** Não é possível que todos os fornecedores de um mesmo tipo de produto o vendam com o mesmo preço, em um mesmo período, sem formar cartel.
- **Super-faturamento:** Um produto não pode ser vendido a um preço muito alto, repetidas vezes, sem haver a ocorrência de super-faturamento de compras.

Pedro então resolveu utilizar o sistema Tamanduá para realizar sua tarefa, achando apropriada a aplicação da técnica de mineração de regras de associação. Ele deveria utilizar o sistema para mapear o problema que ele tinha (baseado nas premissas que elaborou) em tarefas de mineração de dados. A partir da primeira premissa de que um "fornecedor não é capaz de ganhar todas as licitações de um mesmo tipo de produto durante um ano inteiro sem ser favorecido", Pedro considerou que se um fornecedor ganhar mais que 70%, por exemplo, das compras de um produto Y, existem indícios que esse fornecedor poderia ter sido favorecido. Pedro então resolveu criar uma tarefa de mineração, utilizando os seguintes dados abaixo:

- **Nome:** Tarefa de auditoria - Fornecedor e **Descrição:** Tarefa que busca analisar se há favorecimento para algum fornecedor específico (tela da figura 2.1)
- **Base de dados:** Base de compras, que contém 27.834 registros (tela da figura 2.2)
- **Atributos:** Foram escolhidos os atributos: (tela da figura 2.2)
  - produto
  - ano
  - órgão
  - valor efetuado na venda
  - fornecedor (nome, código, endereço)

Em relação a escolha dos parâmetros, Pedro se sentiu confuso ao fornecer valores. Isso porque não são conceitos familiares a ele e Pedro não sabia ao certo o impacto que cada valor poderia ter. Pedro já tinha executado algumas tarefas no Tamanduá, mas utilizando bases diferentes para contextos distintos, o que não pode ser considerado como uma experiência, pois cada tarefa é diferente. Por esse motivo, ele atribui alguns valores que considerou pertinentes, mas sendo essa escolha um desafio para ele.

- **Algoritmo:** foi escolhido o *Eclat*, relacionado às regras de associação (Figura 2.3)
- A escolha dos valores dos parâmetros foram:
  - **Suporte:** Pedro atribuiu o valor de 0.27, que consiste em 75 ocorrências na base, valor considerado relevante por ele.
  - **Confiança:** o valor mínimo determinado foi 70%, direcionado de acordo com a premissa de que um fornecedor é favorecido se obtém grande parte das vendas e essa porcentagem já é um indício, segundo a visão de Pedro.

Pedro então salvou a tarefa e executou a mesma. Como resultado, Pedro obteve um conjunto de regras de associação, como o ilustrado na tela da figura 2.4. Pedro encontrou algumas regras que aparentemente eram interessantes e estavam em destaque como:

[Código_fornecedor = 0156] → [Nome_fornecedor = ETA Ltda] (100.00, 1.00)

Pedro então se sentiu frustrado ao verificar que se tratava de regras óbvias, visto que o código do fornecedor "ETA Ltda" é "0156". Depois de alguma análise, Pedro descobriu que havia escolhido

atributos redundantes, como nome e código do fornecedor. Pedro então teve que editar a tarefa, modificando os atributos escolhidos e executando novamente a tarefa.

Pedro então achou que foram geradas poucas regras, mas não sabia o que poderia ser feito para visualizar mais. Depois de uma análise, Pedro descobriu que o valor fornecido para a confiança estava alto e, por isso, precisava diminuí-lo para serem geradas mais regras. Pedro redefiniu o valor da confiança e executou novamente a tarefa.

Ao analisar os dados, Pedro gostaria de visualizar somente as regras onde aparecia o mês de Janeiro e, para isso, observou que deveria utilizar os filtros. Ele não sabia, entretanto, onde selecionar o atributo desejado, onde seria mais interessante e qual o motivo dos valores aparecerem no antecedente ou conseqüente. Após fazer sua escolha (sem muita certeza que de estava correta), Pedro ainda teve dúvidas ao interpretar as regras obtidas. Ele não conseguia aplicar bem os conceitos envolvidos nas outras medidas de interesse e, por isso, sentia dificuldade de utilizá-las. Ele considerava sempre que valores altos de suporte e confiança representavam regras interessantes, mas não conseguia afirmar isso com a certeza necessária. Depois de muitas dificuldades e dúvidas, Pedro conseguiu obter as informações que desejava, mas sem a certeza de que havia encontrado tudo que poderia da melhor forma possível.

Os desafios vividos por Pedro são enfrentados de uma maneira geral por usuários que não dominam os conceitos envolvidos e isso pode acabar limitando o público e área de atuação das técnicas de mineração de dados [Albergaria et al. (2006)].

Após alguns desafios de interação serem levantados e um exemplo de interação ser apresentado, a próxima seção consiste em analisar soluções que foram propostas em diversos trabalhos que buscam minimizar os problemas existentes.

2.5 Propostas existentes

Primeiramente, um estudo foi feito de forma a verificar que os desafios identificados não eram específicos de um determinado sistema, mas sim gerais de sistemas de segunda geração de mineração de regras de associação. Na tentativa de identificar soluções existentes, foi feita uma pesquisa na literatura e em sistemas de mineração de regras de associação de objetivo geral, que não focam em nenhum domínio específico. Em relação ao desafio de definição de parâmetros de entrada, por exemplo, todos os sistemas analisados [Dama (2006), Analysis (2006), Weka (2006), Tamandua (2006) XLMiner (2006), Tutorial (2006)] apresentam na sua linguagem de interface os mesmos conceitos técnicos que formam os desafios para os usuários. Para ilustrar, as Figuras 2.5, 2.6 e 2.7 mostram telas do DBMiner [Tutorial (2006)], Dama [Dama (2006)] e XLMiner [XLMiner (2006)], respectivamente, na qual os usuários entram com os dados necessários para a mineração de dados.

Poucos trabalhos foram encontrados sobre a necessidade de entendimento por parte dos usuários dos conceitos de mineração de dados. Dentre estes, destacamos Thearling e colegas [Thearling et al. (2002)] que chamam a atenção para a importância do usuário entender e confiar em sistemas de mineração de dados, mas não apresenta quais aspectos são necessários para isso.

Em relação à visualização dos resultados, existem alguns esforços em melhorar a usabilidade de sistemas de mineração de regras de associação. Uma estratégia consiste em ajudar os usuários a explorar a grande quantidade de regras apresentadas como resultado, auxiliando-os no processo de identificação de regras interessantes. Nesse contexto, duas abordagens são geralmente utilizadas. A primeira consiste em construir ferramentas e diferentes formas de visualizar as regras, possibilitando que os usuários possam ter uma visão geral dos resultados mais facilmente. Em

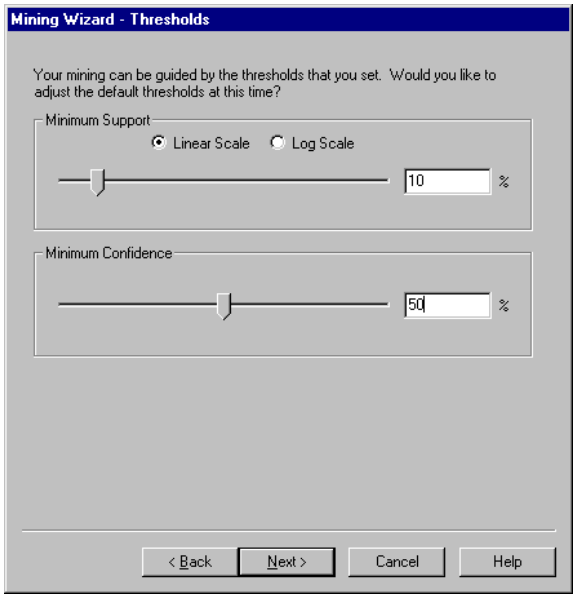

Figura 2.5: DBMiner: Parâmetros de entrada (suporte e confiança mínimos)

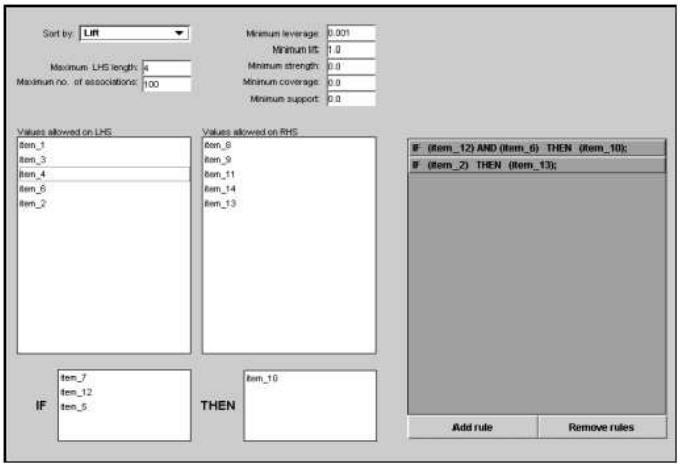

Figura 2.6: DAMA Prototype: Parâmetros de entrada (suporte e confiança mínimos, dentre outros)

[Zaki e Phoophakdee (2003)] e [Rainsford e Roddick (2000)], por exemplo, são usadas técnicas de grafos para apresentar um conjunto de regras. Nesse tipo de visualização, os nodos dos grafos representam os itens ou conjunto de itens e as arestas as regras, onde o nodo de origem é o antecedente e o de destino o conseqüente, como pode ser visualizado na Figura 2.8. Em diversas aplicações como DBMiner [Han et al. (1996)] e IBM Intelligent Miner, as regras são apresentadas em formas gráficas, onde um eixo é o antecedente e o outro o conseqüente, como mostra a Figura 2.9. O problema é que essa forma de apresentação não é escalável para muitos atributos. De forma a minimizar esse problema, foi proposto em [Wong et al. (1999)] uma visualização 3D, plotando regras e atributos e apresentando os valores de suporte e confiança de forma conjunta, que podem ser visualizados na Figura 2.10. Porém, essa visualização tornou-se complexa em termos de identificação dos atributos em relação à regra (identificar qual atributo está no antecedente e qual está no conseqüente).

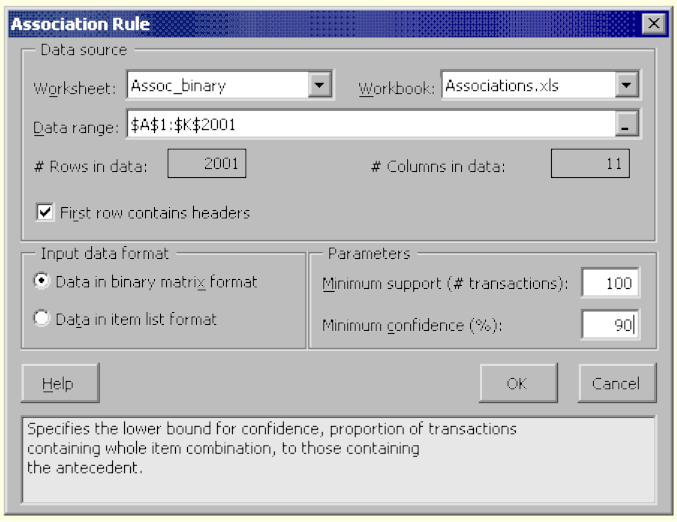

Figura 2.7: XLMiner: Parâmetros de entrada (suporte e confiança mínimos)

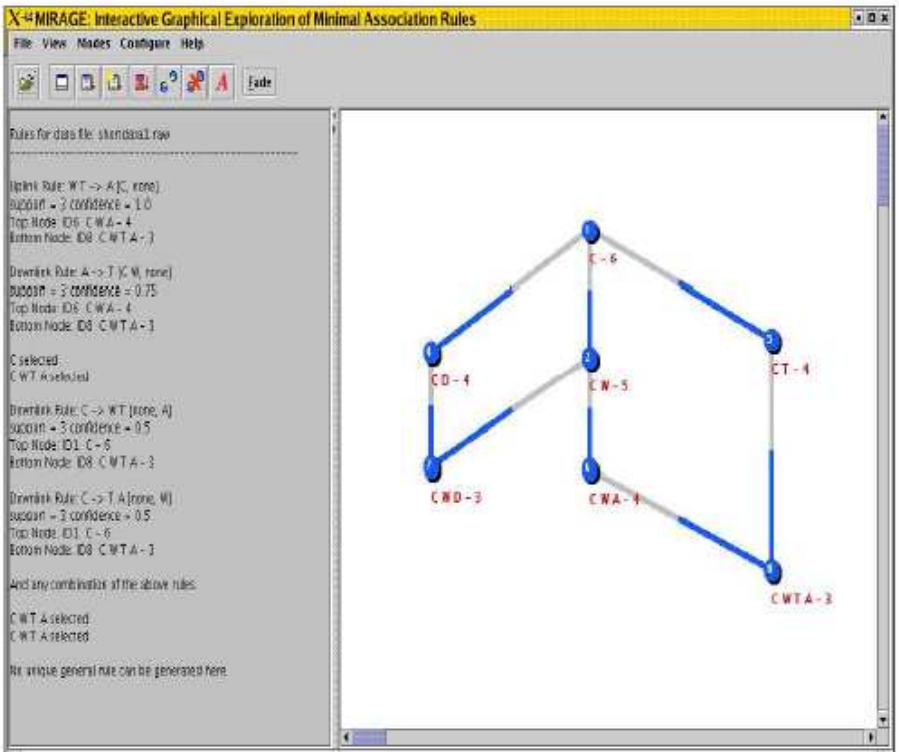

Figura 2.8: Framework Mirage, visualização proposta por Zaki [Zaki e Phoophakdee (2003)]

A segunda abordagem consiste em diminuir a quantidade de regras a serem apresentadas, onde uma das possibilidades é utilizando taxonomias. Em [Srikant e Agrawal (1997)], por exemplo, todas as regras possíveis (com e sem taxonomias) são apresentadas e num segundo momento buscam retirar as regras que não são interessantes, de acordo com uma determinada medida que deve ser escolhida. Em [Domingues e Rezende (2005)] também é utilizada a generalização de regras de associação utilizando taxonomias, onde é proposto um algoritmo denominado GART (*Generalization of Association Rules using Taxonomies*). Porém, as regras são generalizadas na etapa de

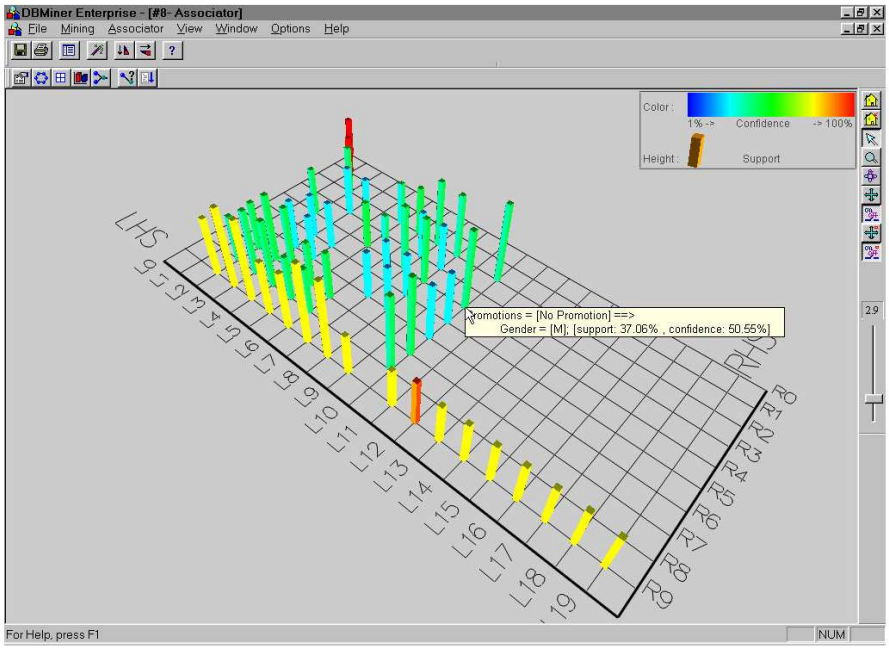

Figura 2.9: Visualização do DBMiner: rule body(LHS) x rule head(RHS)

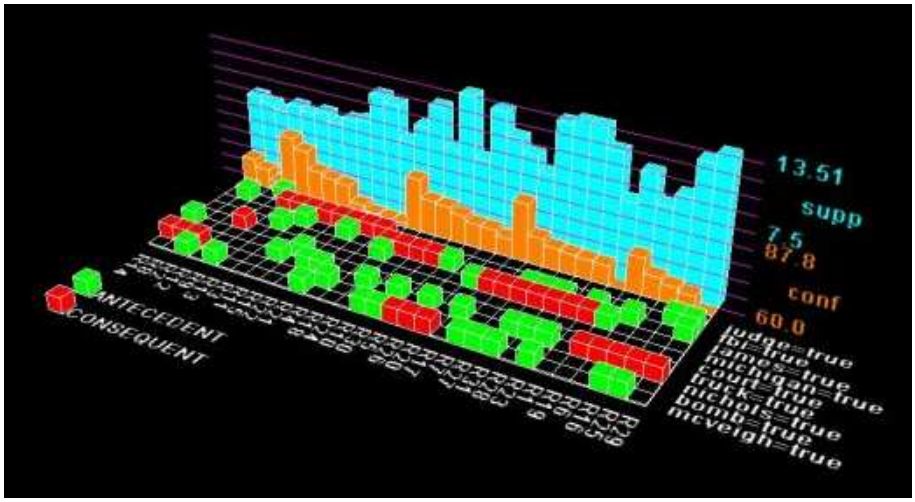

Figura 2.10: Visualização 3D proposta por Wong [Wong et al. (1999)]

pós-processamento, não aumentando o volume de regras geradas como ocorre no trabalho citado anteriormente.

Outro objetivo de recorrente interesse consiste em auxiliar os usuários a analisarem as regras geradas. Em [Hofmann et al. (2000)], a técnica de *Mosaic Plots* é utilizada para melhorar o entendimento das regras (*Mosaic plots* são metáforas visuais para as tabelas de contingência). A regra é apresentada em forma gráfica, onde a área é o suporte e a confiança é apresentada pelo total preenchido. No entanto, esse trabalho oferece apoio aos usuários que já conhecem os conceitos envolvidos no contexto das regras de associação, não auxiliando no entendimento em si. Ele visa oferecer uma ferramenta para que usuários que já possuam o conhecimento técnico necessário possam analisar a relevância de uma determinada regra dentro do conjunto em que ela está inserida.

Uma forma de ajuda para o entendimento das regras é apresentada em [Mei et al. (2006)]. A idéia do artigo foi inspirada na linguagem natural, onde a semântica da palavra pode ser inferida do contexto, onde as palavras que compartilham contextos tendem a ser similares. Eles apresentam no trabalho uma forma de gerar automaticamente informações semânticas de um determinado padrão, denominadas anotações semânticas. Tais anotações consistem em um conjunto de fortes indicadores contextuais, um conjunto de transações representativas e um conjunto de padrões semanticamente similares. O método pode ser aplicado a qualquer técnica de mineração de padrões freqüentes como um passo para facilitar a interpretação de padrões encontrados.

Todas essas propostas são avanços na melhoria da usabilidade em sistemas de mineração de regras de associação. No entanto, eles continuam a exigir que os usuários de mineração de dados aprendam conceitos técnicos, a fim de interagir com o sistema. Sistemas de terceira e quarta geração propõe estratégias diferentes para melhorar a usabilidade dos sistemas de KDD.

Sistemas de terceira geração são voltados para um contexto específico, com conceitos próprios dos usuários. Um exemplo é o sistema o ADS (*Advanced-Detection System*), apresentado na Figura 2.11 que detecta fraudes no comportamento do Nasdaq Stock Market, de acordo com regulamento NASD [Senator et al. (2002)]. Outro exemplo é o HNC Software Falcon para detecção de fraude de cartão de crédito [Rainho (2001)]. Embora tente "esconder" os conceitos de mineração envolvidos no processo, as ferramentas de terceira geração ficam restritas a um determinado contexto e problema. Assim, as interfaces de sistemas de terceira geração são orientadas para os usuários, mas eles são voltados para contexto e tarefas específicos, sendo necessário um novo desenvolvimento para cada domínio.

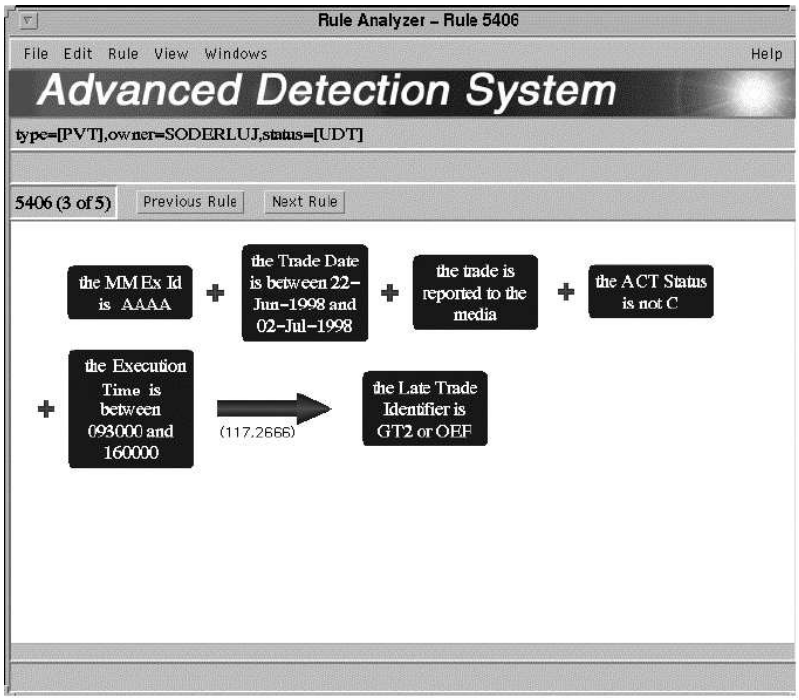

Figura 2.11: Tela do sistema ADS - Representação de uma regra

Os sistemas de quarta geração buscam auxiliar os usuários no complexo processo de KDD e são chamadas de ferramentas de assistência ao processo KDD. Elas fornecem aos usuários uma enumeração sistemática dos processos válidos de mineração, onde os algoritmos possuem pré-condições e

efeitos compatíveis em uma determinada seqüência. Além disso, permite criar um "ranking" dos processos classificados pela velocidade e exatidão, facilitando a escolha de qual processo deve ser executado. Essas ferramentas ajudam nas tomadas de decisões entre as várias possibilidades de qual caminho seguir em uma determinada tarefa de mineração [Goldschmidt et al. (2002)]. Em [Goldschmidt (2003)] é proposta uma ferramenta de assistência, uma máquina de assistência inteligente à orientação do processo de KDD (também chamada IKDD - *Intelligent Assistance in KDD*), proposta para ser uma ferramenta didática voltada para a formação de profissionais. Nesse sistema, os usuários são guiados a entenderem o processo, aprendendo gradualmente os conceitos envolvidos. Na visão contextual clássica do processo KDD, de um lado está o homem e do outro um conjunto de recursos utilizados na execução das etapas do processo KDD [Goldschmidt (2003)]. Esse conjunto refere-se, de uma maneira geral, a um repositório de algoritmos KDD, integrados ou não. A máquina IKDD entra como um componente auxiliar (Figura 2.12, apresentada em [Goldschmidt (2003)]). A máquina não é responsável por executar o processo e sim sugere alternativas de ações.

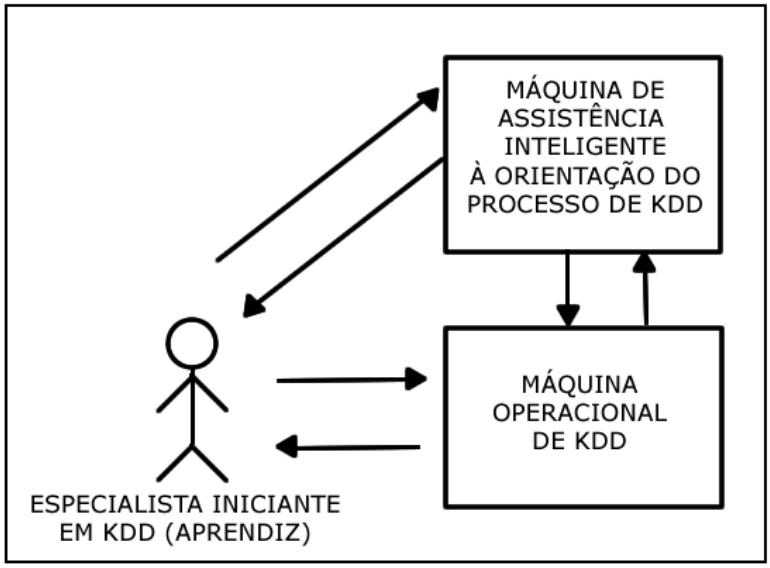

Figura 2.12: Visão contextual da máquina IKDD segundo Goldschmidt [Goldschmidt (2003)]

Para que as ferramentas de assistência (sistemas de quarta geração) possam auxiliar os usuários no processo KDD, devem existir algumas definições sobre o mesmo. Por exemplo, é necessário que a ferramenta saiba quais os pré-processamentos para cada técnica e qual o impacto da execução de cada uma, por exemplo. Isso porque, ao indicar ao usuário qual decisão melhor a ser tomada, a indicação deve ser baseada em uma teoria sólida e correta. Alguns trabalhos estão voltados para essa linha teórica. Por exemplo, em [Bernstein e Provost (2001)][Goldschmidt et al. (2002)] os autores buscaram descrever os algoritmos de KDD e suas características através de ontologias, mostrando suas pré-condições e efeitos. Em [Morik (2000)] foi analisada a influência de ações de pré-processamento nos desempenhos das tarefas de mineração e em [Brazdil et al. (2003)][Soares et al. (2001)] analisou-se a aderência de algoritmos de mineração a um conjunto de dados.

Voltando às ferramentas de segunda geração, como vimos, elas demandam dos usuários um grande conhecimento da teoria envolvida em mineração de dados. Embora os trabalhos apresentados permitam a melhoria de diferentes aspectos relacionados ao uso de sistemas de segunda geração, de forma geral eles continuam a demandar esse conhecimento. As ferramentas de terceira geração

buscam construir soluções voltadas para um determinado contexto, buscando abstrair os conceitos envolvidos, mas são construídos para um problema específico e, se o problema mudar ou evoluir, o sistema pode deixar de ser útil. Os sistemas de quarta geração buscam auxiliar aos usuários no processo, mas apresentando os conceitos envolvidos aos mesmos. O auxílio é em relação às decisões possíveis, visto que no processo KDD são inúmeras as possibilidades de interação. Nesse caso, os usuários precisam aprender os conceitos, o que pode representar um alto custo para alguns.

O que apresentamos nesse trabalho consiste em uma abordagem diferente das que foram apresentadas. Consiste em um modelo que pode ser acoplado a sistemas de segunda geração, buscando criar uma camada de abstração. Diferentemente dos de terceira geração, essa solução não é fixa para um determinado contexto e pode ser aplicada em diferentes domínios. Além disso, não se busca apresentar os conceitos envolvidos como as ferramentas de quarta geração, mas a idéia é que um grupo de usuários (denominados especialistas) possam criar um nível de abstração para que o sistema possa ser usado por usuários finais sem que esses precisem aprender conceitos técnicos de mineração de dados. Assim, essa solução é capaz de permitir uma interface em um nível maior de abstração, não limitando o contexto de aplicação.

Depois dos conceitos envolvidos no contexto de mineração de dados serem apresentados, a fundamentação teórica do modelo proposto será descrita no próximo capítulo.

Capítulo 3

Fundamentação teórica

No processo de interação usuário-sistema, a interface é o combinado de software e hardware necessário para viabilizar e facilitar os processos de comunicação entre o usuário e a aplicação [Preece et al. (1994)]. Segundo Moran [Moran (1981)], a interface de usuário deve ser entendida como sendo a parte de um sistema computacional com a qual uma pessoa entra em contato de forma física, perceptiva e conceitual.

O termo Interação Humano-Computador (IHC) foi adotado na década de 1980 para descrever um novo campo de estudo. O termo não é apenas para abranger interfaces, mas todos os aspectos relacionados a interação entre pessoas e sistemas computacionais [Preece et al. (1994)]. Trata-se de uma matéria multidisciplinar que relaciona ciência da computação, design, ergonomia, psicologia, sociologia, semiótica, lingüística e áreas afins.

Um ponto importante a ser compreendido em IHC está relacionado à qualidade de um determinado sistema em relação à interação. Isso porque acrescentar funcionalidades não significa melhorar a interação e também não pode ser desculpa para um design pobre [Preece et al. (1994)]. Um bom exemplo é o dado por Norman [Norman (1988)] com relação aos carros. Ele afirma que "interagir" com carros, que normalmente possuem cerca de 100 comandos ou mais (dentre funcionalidades de rádio, ventilação, janelas, direção, luzes, etc.) muitas vezes não é tão difícil como uma tarefa de programar um horário de gravação em um vídeo. Um fato relacionado consiste no feedback dado pelos comandos do carro serem mais imediatos e óbvios. Além disso, os símbolos utilizados em carros seguem determinados padrões e não se diferenciam tanto de um carro para outro. Assim, as pessoas que já dirigiram um carro, sabem o que esperar em qualquer outro.

Os objetivos de IHC podem ser resumidos em "desenvolver ou melhorar a segurança, utilidade, eficácia, eficiência e usabilidade de sistemas computacionais" [Barlow et al. (1989)]. Sistemas aqui não está se referindo a software ou hardware especificamente, mas todo o contexto de uso. Utilidade refere-se às funcionalidades do sistema, o que ele faz. Eficácia relaciona-se com a precisão, completeza com que os usuários atingem objetivos específicos, acessando a informação correta ou gerando os resultados esperados. Já a eficiência está relacionada com a precisão, completeza com que os usuários atingem seus objetivos em relação à quantidade de recursos gastos. Usabilidade envolve o sistema ser fácil de aprender e fácil de usar.

Por sua característica multidisciplinar, várias foram as abordagens elaboradas para analisar a forma de interação. Uma abordagem bastante difundida, por exemplo, refere-se a engenharia cognitiva [Norman (1986)]. Ela é baseada na psicologia cognitiva e possui como objetivo entender o sistema humano de processamento de informação. Consiste na elaboração de modelos cogniti-

vos mentais que permitem aos projetistas entenderem o processo cognitivo humano e possam ser utilizados durante a interação. Norman considera que o designer precisa entender o processo de interação e propõe a teoria das ações (ilustrada na Figura 3.1) para ajudá-lo.

A teoria das ações envolve dois golfos, o de execução e o de avaliação. O de execução consiste na definição da meta, onde o objetivo do usuário deve ser traduzido em comandos de interface. Já o de avaliação consiste na análise dos resultados obtidos, em que respostas do sistema devem ser traduzidas em uma avaliação sobre o quanto se atingiu do objetivo inicial. As distâncias semântica e articulatória são utilizadas como métricas para se avaliar a qualidade da linguagem de interface. A distância semântica representa a distância entre a intenção do usuário e o conteúdo dos signos presentes na linguagem de interface, onde signo é aquilo que representa alguma coisa para alguém [Peirce (1958)]. Já a distância articulatória representa a distância entre o conteúdo dos signos e sua expressão na interface.

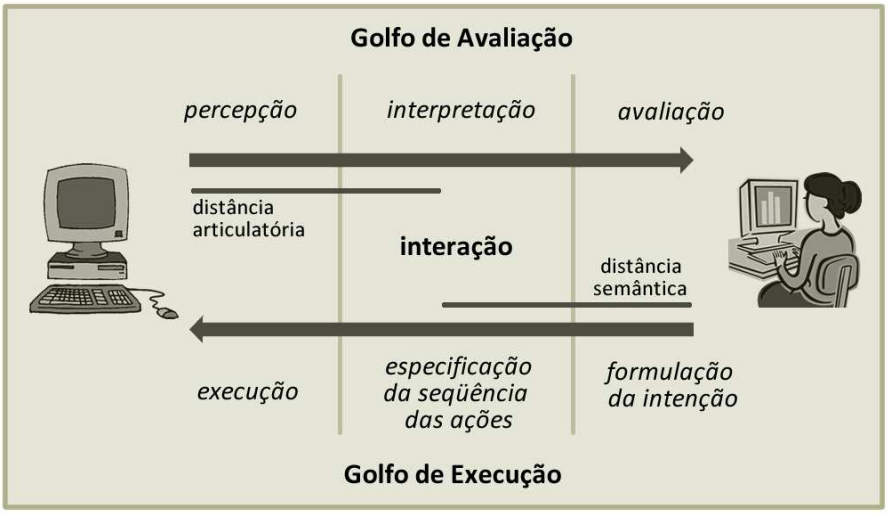

Figura 3.1: Teoria das ações - processo de interação dos usuários

A Engenharia cognitiva foca na análise da relação interface-usuário e não na relação designer-sistema. Além disso, e sendo esse o aspecto principal, a Engenharia cognitiva considera o usuário e suas intenções como únicos. Adota-se a premissa de que os usuários possuem o mesmo conjunto de intenções e interpretam os signos apresentados pela interface da mesma forma [de Souza (2005)]. Assim, o caráter evolutivo e contingente das interpretações e usos dos usuários não são considerados.

Esse aspecto das variações na interação é abordado pela Engenharia Semiótica, teoria adotada neste trabalho como base teórica. Ela pode ser considerada, dentre outras, como uma teoria pós-cognitiva [Bim et al. (2007)]. Ela trata o envolvimento entre designer e sistema como um processo de comunicação em que os projetistas devem transmitir suas mensagens, no lugar de apenas os usuários interpretarem o que foi previamente projetado.

A Engenharia Semiótica é uma das poucas tentativas de juntar a semiótica e IHC de maneira concisa e consistente, para suportar a organização e a descoberta do conhecimento novo, o estabelecimento de métodos úteis de pesquisa para análise e síntese.

3.1 Engenharia Semiótica

A Engenharia Semiótica é uma teoria que caracteriza a interação humano-computador como um caso particular de comunicação humana mediada por sistemas computacionais [de Souza (2005)]. Trata-se do designer(projetista) se comunicando com o usuário, mediado pelo sistema, onde a interface é uma mensagem para o usuário representando a maneira como o designer projetou, para que e por que ela foi construída.

Em uma teoria, uma ontologia é utilizada para descrever conceitos e relacionamentos entre os mesmos, além de categorizá-los. A teoria da Engenharia Semiótica possui como ontologia quatro categorias: o processo de comunicação, o processo de significação, os interlocutores envolvidos e o espaço do design. O processo de significação envolve os conceitos de signos e semiose, enquanto o de comunicação a intenção, conteúdo e expressão. Os interlocutores envolvem os projetistas, os sistemas e os usuários. Já o espaço de design envolve os termos emissor, receptor, mensagem, contexto, códigos, canal e mensagem.

Assim, a Engenharia Semiótica envolve o estudo dos signos, o processo de significação e o de comunicação voltados para o contexto de IHC. O processo de significação é aquele pelo qual uma determinada cultura associa sistematicamente um conjunto de expressões a um conjunto de conteúdos, que envolve pela produção e interpretação dos signos. Já o processo de comunicação é aquele pelo qual o grupo de uma cultura explora os sistemas de significação disponibilizados para interagir com outros indivíduos ou grupos.

Como já citado, um signo, segundo Peirce [Peirce (1958)], é aquilo que representa alguma coisa para alguém. Peirce apresenta a estrutura do signo como um conjunto de três constituintes: *representamen* (representação), objeto (referente) e significado (interpretante), apresentados na Figura 3.2(A). O significado é sempre o mediador entre a representação e o que é referenciado [de Souza (2005) p.41]. Por exemplo, tomemos um objeto que é utilizado para cortar materiais de pouca espessura e que não requeiram grande força de corte, a tesoura. Uma tesoura é um objeto que pode ser representado tanto pela palavra "tesoura" quanto pela imagem. Assim, o "objeto cortante", cuja representação pode ser pela imagem ou palavra "tesoura" pode ter como um dos significados "uma tesoura de criança" (Figura 3.2(B)).

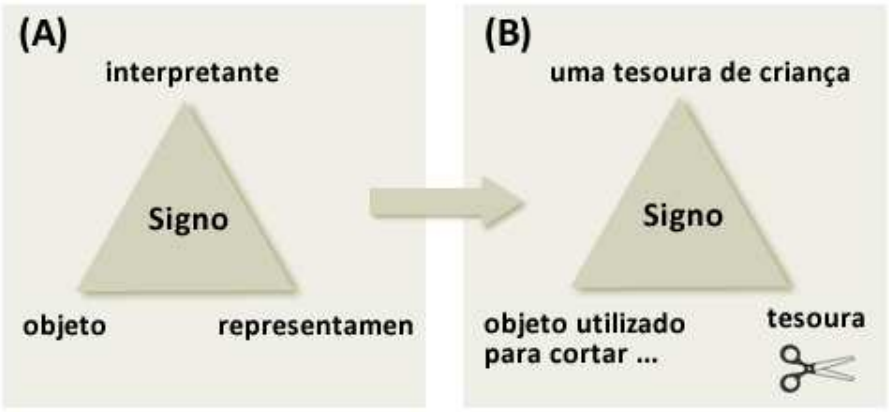

Figura 3.2: Estrutura do signo, segundo Peirce

Uma pessoa, ao perceber um signo e tentar interpretá-lo, gera uma idéia relacionada, um significado (interpretante). Ao ouvir a palavra "tesoura", alguém pode pensar, por exemplo, em

uma tesoura pequena de pontas arredondadas utilizadas por crianças. Esse significado gerado na mente do ouvinte é seu interpretante. Um interpretante pode dar origem a outras idéias, iniciando o processo de interpretação, denominado semiose (Peirce, 1931-1958).

Teoricamente, trata-se de um processo infinito, denominado semiose ilimitada [Eco (1976)]. Na prática, o processo de semiose termina quando a pessoa desiste (é interrompida ou inicializa uma outra atividade, por exemplo) ou quando ela encontra um significado satisfatório. Voltando ao exemplo da tesoura, imagine uma situação onde, em um grupo, uma pessoa pede uma tesoura emprestada. Uma mulher presente, mãe de um menino, procura em sua bolsa e encontra o objeto, retirando e emprestando ao colega que solicitou. Ao realizar esse ato, a mãe interpreta inicialmente que a tesoura pertence ao seu filho. Logo em seguida, ela se lembra que "as aulas de seu filho irão começar na próxima semana" e ela ainda "não comprou os materiais escolares". Ela então se lembra que precisa comprar o material e, para isso, precisa pegar a "lista que se encontra na internet". Assim, o processo de semiose pode se estender por esse caminho ou outros relacionados. Por exemplo, se a mãe tivesse preocupada com o dinheiro a ser gasto, poderia com isso se lembrar de seus problemas financeiros, lembraria de outras contas a serem pagas e assim por diante. Em outro caso, se não existissem os problemas, a mãe poderia programar um passeio junto com o filho para comprar o material, planejando também outras compras e passeios. Assim, a cadeia de semiose varia de acordo com as circunstâncias que as pessoas se encontram, além de influência da cultura e hábitos.

A perspectiva da semiose na comunicação é utilizada na teoria da Engenharia Semiótica ao visualizar o processo de interação, onde projetistas se comunicam com os usuários através das interfaces [de Souza (2005)]. Isso porque os signos apresentados pelos designers (na interface são representados por comandos, imagens, ajudas, mensagens) devem ser interpretados de forma compatível com o que foi inicialmente projetado. Variações em relação aos usuários (como contexto, cultura e hábitos) são desafios do designer no processo de significação, onde a consistência entre os interpretantes do designer e os dos usuários é a situação ideal. Porém, não há como garantir essa consistência, visto que não há como prever os interpretantes que podem ser gerados pelos usuários. Assim, a Engenharia semiótica é uma teoria explicativa, onde busca estimular o designer a se preocupar com essa correspondência.

A codificação e decodificação dos signos utilizados para se comunicar constituem o processo de significação. O processo de comunicação se dá quando signos são codificados de forma a serem transmitidos, através de um canal, aos destinatários (humanos ou não). O modelo de comunicação proposto por Jakobson [Jakobson (1960)] apresenta 6 elementos em sua definição: emissor, mensagem, receptor, canal, contexto, código. O **emissor** transmite uma **mensagem** para o **receptor** através de um **canal**. A mensagem é expressa em um **código** e se refere a um **contexto**. Na comunicação, emissores e receptores alternam os papéis de interlocutores.

A Engenharia Semiótica, considerando a interação humano-computador como um processo de comunicação, tem o computador como o canal por onde a mensagem é transmitida ao usuário, canal da comunicação usuário-sistema. Como o designer não pode estar presente fisicamente na interface, ele é representado por seu preposto [Prates et al. (2000)].

O projetista deve então criar e definir um artefato intelectual para transmitir sua mensagem. Artefatos intelectuais são objetos não naturais criados por humanos. Alguns são concretos como garfos e facas, utilizados para auxiliar a pessoa a se alimentar e alguns são abstratos como segurança. Alguns são de finalidade física como cadeiras, outras somente mentais como tabelas da verdade.

Sistemas são artefatos intelectuais lingüísticos e possuem as seguintes características [de Souza

(2005)]:

- Possui compreensão ou interpretação específicas de uma situação do problema;
- Possui um conjunto específico de soluções para perceber a situação do problema;
- A codificação da situação do problema e sua solução são fundamentalmente lingüísticas;
- A finalidade do artefato pode ser completamente atingida por seus usuários se estes puderem formular dentro do sistema lingüístico no qual o artefato está codificado.

As interfaces são consideradas como artefatos intelectuais que possuem "produtores" e "consumidores", usando um mesmo conjunto de regras na linguagem. O processo de comunicação ocorre através de sistemas computacionais, sendo esses artefatos de metacomunicação, pois as mensagens são comunicadas através de si mesmas. Nessa perspectiva, a interface é vista como uma mensagem única e indireta, enviada de projetistas a usuários. E ela é unidirecional visto que o usuário não consegue se comunicar com o designer durante a interação. A metamensagem possui como objetivo comunicar ao usuário o seguinte conteúdo (Figura 3.3):

"Eis minha interpretação de quem você é, o que aprendi que você tem ou quer fazer, preferencialmente de que formas e por quê. Eis, portanto, o sistema que concebi para você, o qual você pode ou deve usar assim, a fim de realizar uma série de objetivos associados com esta minha visão".

Figura 3.3: Metamensagem - Engenharia Semiótica

O processo da metacomunicação envolve projetistas e usuários dentro do contexto de IHC e possui os seguintes passos [de Souza (2005)]:

- Os projetistas realizam uma análise de contexto, verificando tarefas, ambiente e perfil dos usuários;
- Os projetistas expressam suas visões através dos sistemas computacionais;
- Os usuários interagem com os sistemas, analisando as mensagens que são transmitidas;
- Após a interação, os usuários são capazes de analisar o artefato produzido.

Em Design Centrado no Usuário [Norman (1986)], os projetistas tentam identificar mais precisamente quanto possível, o que os usuários desejam. Estudos dos usuários e análise de tarefas ajudam os projetistas a desenvolverem o modelo como uma imagem do sistema, onde os usuários devem interagir para atingirem seus objetivos. A imagem do sistema é a chave final; se ela apresentar relações fáceis e intuitivas, os usuários poderão facilmente usar o sistema e relembrar suas funcionalidades após um tempo de uso. O sucesso do Design Centrado no Usuário depende do estudo realizado com os usuários para descobrir seus comportamentos e reações.

Em Engenharia Semiótica, o projetista busca entender o que os usuários querem e precisam e tenta comunicar a sua visão à eles. Através do preposto, o projetista se comunica com os usuários, de forma que o usuário ache a visão do projetista fácil e útil. O designer tem o desafio de guiar o comportamento dos usuários, buscando comunicar a eles aquilo que é necessário para se fazer melhor uso do sistema. A comparação entre as duas perspectivas, Design Centrado no Usuário e Engenharia Semiótica, pode ser visualizada na Figura 3.4.

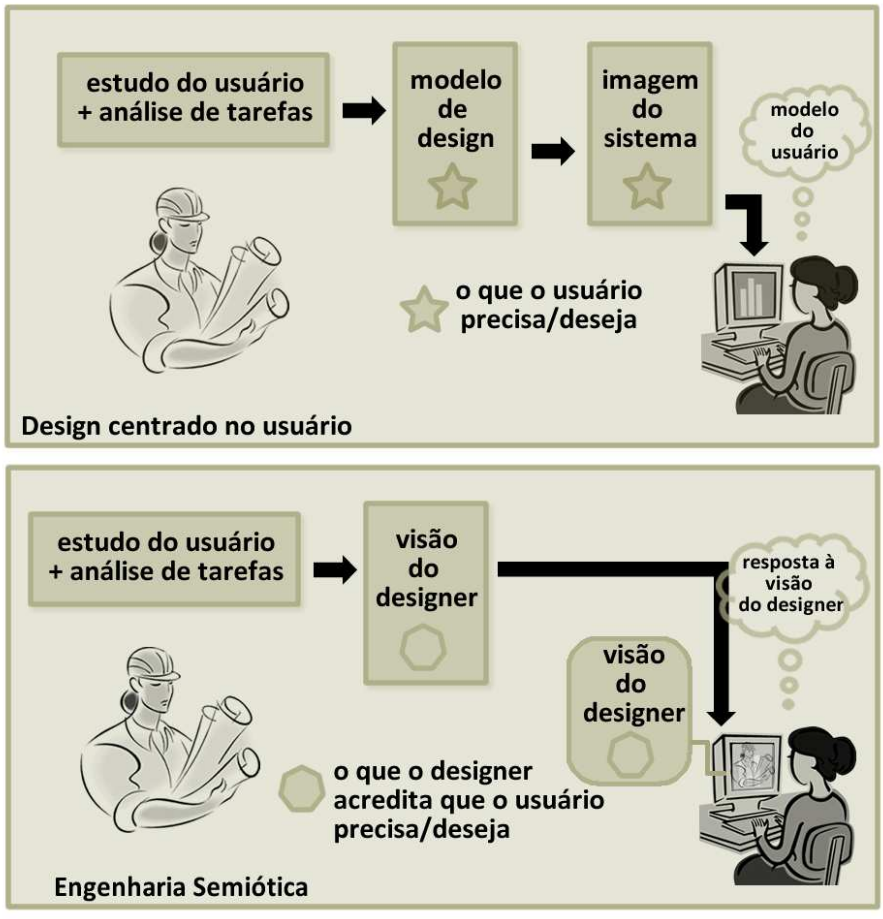

Figura 3.4: Design Centrado no Usuário x Engenharia Semiótica[2] [de Souza (2005), pag.8]

Quando é o emissor no processo de comunicação designer-usuário, o projetista deve escolher cuidadosamente os signos a serem utilizados, considerando as dimensões de conteúdo, de expressão e de intenção de comunicação. Quando o usuário for o emissor na comunicação usuário-sistema, a interface deve permitir que ele consiga se expressar com os signos desejados.

[2]A imagem foi adaptada pela autora deste trabalho.

Mesmo com todo cuidado na escolha dos signos, problemas na comunicação podem ocorrer. Assim, devem existir formas de recuperação, prevenção e tratamento desses mal-entendidos. Ferramentas que podem auxiliar o designer na análise dos conhecimentos relacionados com o design são chamadas ferramentas epistêmicas. Elas devem auxiliar o designer na produção e análise constante de novos conhecimentos que estão diretamente relacionados com questões de design [de Souza (2005) p.105]. Ou seja, elas devem ajudar o designer a aumentar o seu conhecimento sobre os problemas, permitindo hipotetizar, avaliar e restringir soluções candidatas. Assim, ferramentas epistêmicas são utilizadas para geração de conhecimento cujo objetivo é ajudar designers a identificar e colocar em perspectiva os elementos de situações de design únicas. Porém, uma ferramenta epistêmica é usada para aumentar o entendimento de uma pessoa sobre o problema que ela está resolvendo; não para lhe dar diretamente uma resposta ou focar diretamente a sua solução, como é o caso de diretrizes e regras, por exemplo [de Souza (2005)].

3.2 Desenvolvimento por usuários finais

Um passo do processo da metacomunicação que os projetistas encontram muita dificuldade consiste na análise de contexto, pela variedade de usuários e situações e a dificuldade de levantamento e entendimento das necessidades dos usuários. Além disso, as necessidades e contextos dos usuários mudam com o tempo [Fischer (2007)]. Isso porque os usuários precisam de adaptações, mudanças e até mesmo novas funcionalidades ou comportamentos diferentes do que foi especificado do sistema.

Existem trabalhos como [Myers (1992)][Suchman (1987)] que apontam os motivos da dificuldade de um software oferecer soluções para cada usuário em particular em problemas específicos. Estas pesquisas buscam criar mecanismos que façam com que os usuários sejam capazes de modificar os sistemas, adaptando para as novas necessidades. Uma das soluções propostas é o desenvolvimento pelo usuário final (EUD - *End User Development*). Isto é, permitir ao usuário final que possa adequar e adaptar o sistema para as utilizações que surgirem com o tempo.

Segundo [Lieberman et al. (2006)], EUD pode ser definido como um conjunto de métodos, técnicas e ferramentas que permitem que os usuários dos sistemas, atuem como desenvolvedores não profissionais de *software*, e em algum ponto possam criar, modificar ou estender o sistema.

As soluções em EUD variam de forma a oferecer aos usuários desde oportunidades de customizar os sistemas até mecanismos de reprogramação de componentes [Fischer et al. (2004), Fischer (2007), de Souza (2005), de Souza e Barbosa (2006)].

Em [Morch (1997)] são apresentadas diferentes formas de se alterar e adaptar um *software*, classificadas como:

- **Customização:** a partir de um conjunto de configurações pré-definidas, é possível modificar a aparência dos objetos (cor, fonte, períodos de atualização) ou mudar valores de seus atributos.
- **Integração:** sem acessar o código do sistema diretamente, é possível conectar componentes e acrescentar funcionalidades, indo além da customização.
- **Extensão:** possibilita acrescentar novas funcionalidades em pontos definidos pelo designer, sendo possível a adição de código. Dependendo do tipo de extensão, pode ser executada por usuários finais, desenvolvedores ou pela própria aplicação.

Ao levantar a possibilidade de o usuário final desenvolver, uma questão importante de ser apresentada refere-se ao tipo de linguagem de programação a ser utilizada. São inúmeras as formas existentes, onde cada uma apresenta vantagens e desvantagens. Em [Nardi (1993)] é apresentado um gráfico que ilustra desde a tarefa de dar valores a parâmetros até a programação tradicional, relacionando a possibilidade de construção interativa e a expressividade (Figura 3.5). É possível perceber que linguagens tradicionais, por exemplo, apresentam alta expressividade, mas baixa construção interativa. O ideal é encontrar a linguagem adequada para o problema a ser aplicado. Programação visual, por exemplo, apresenta uma boa relação entre os dois parâmetros, mas, de toda forma, deve ser analisada sua utilização, pois depende do perfil dos usuários e os resultados que se deseja obter.

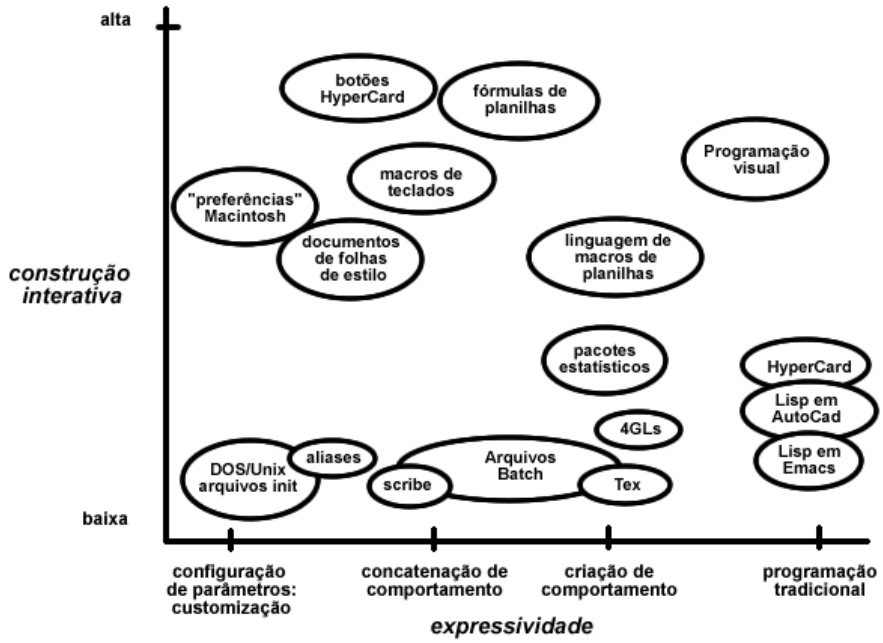

Figura 3.5: Linguagens de Programação - por Nardi[3] [Nardi (1993)]

Assim como Nardi, Fischer apresenta uma análise em relação às linguagens de programação [Fischer et al. (2004)], porém utilizando outros fatores. Em sua análise, ele relaciona o custo de aprendizagem de uma determinada linguagem com o escopo em que pode ser aplicada. Aqui se pretende analisar qual a linguagem escolher, de acordo com o objetivo que se tem. Assim, por exemplo, linguagens tradicionais como Java e C++ possuem um alto custo de aprendizagem para usuários finais, mas elas podem ser aplicadas em um contexto mais amplo, desenvolvendo diferentes tipos de sistemas e extensões. Já linguagens específicas de um determinado domínio, por exemplo, possuem baixo custo de aprendizado, mas só podem ser aplicadas onde foram desenvolvidas. A Figura 3.6 mostra a relação entre custo de aprendizagem versus escopo apresentada por Fischer.

Como uma solução EUD, Fischer apresenta o conceito de meta-design [Fischer (2007)]. Meta-design caracteriza objetivos, técnicas e processos para criar novas mídias e contextos que permitem que "os donos do problema" (os usuários) atuem como projetistas [Fischer (2007)]. O objetivo principal do meta-design é criar um contexto sócio-tecnológico que dê "poder" aos usuários para

[3]A imagem foi traduzida pela autora deste trabalho.
[4]A imagem foi traduzida pela autora deste trabalho.

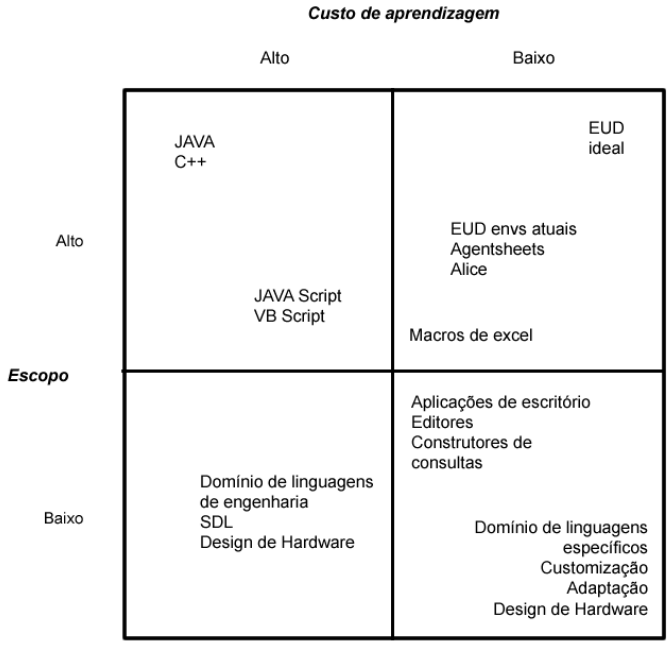

Figura 3.6: Linguagens de Programação - por Fischer[4] [Fischer et al. (2004)]

estarem continuamente desenvolvendo no lugar de fazerem uso restrito de sistemas existentes.

Uma condição necessária, mas não suficiente, para meta-design é o sistema apresentar recursos avançados que permitam aos usuários criarem complexas extensões e customizações. Meta-design provê oportunidades, ferramentas e estruturas sociais para estender o sistema de acordo com as necessidades. O que meta-design propõe então é que usuários finais possam ser co-criadores dos sistemas. Dentro da teoria da Engenharia Semiótica, o usuário seria um co-autor da mensagem a ser transmitida através do sistema. A próxima seção apresenta a visão da Engenharia Semiótica em relação a EUD.

3.2.1 Visão da Engenharia Semiótica

A visão da Engenharia Semiótica em relação às extensões envolve os pilares considerados na comunicação: o sistema de significação e o de comunicação em si. A comunicação humana não é limitada (semiose ilimitada, descrita na seção 3.1). Além disso, existem inúmeras formas humanas de se expressar, como humor, ironia, criação de novos termos e significados. Expressão, conteúdo e intenção são três dimensões fundamentais no processo de comunicação humana, pois exploramos e associamos expressões existentes em nossa cultura para ativar certos efeitos em nossos ouvintes [de Souza e Barbosa (2006)].

Além do processo da semiose ser ilimitado, cada signo apresentado pode ter diferentes significados para cada usuário. Outro processo interessante de interpretação é chamado abdução. Trata-se do contrário da dedução, onde na dedução aplicamos regras conhecidas em fatos conhecidos para tirarmos conclusões e na abdução observamos fatos e tiramos conclusões hipotéticas. Por exemplo, se não conseguir renomear um arquivo, o usuário pode pensar que é porque ele está aberto, porque ele conseguiu renomear um que estava fechado.

Com os computadores, não há intenção, semiose ilimitada ou interpretações diferentes. Cada elemento presente na interface possui ações e traduções pré-definidas e que não podem ser modificadas de acordo com contextos, reagindo a fatores externos, por exemplo. O sucesso da comunicação humano-computador depende da interpretação que os usuários fazem dos signos presentes e os efeitos que eles representam.

Verifica-se assim que há uma grande dificuldade na comunicação humano-computador. Os computadores não conseguem realmente processar a intenção humana e, por esse motivo, permitir que usuários sejam também designers pode ser uma boa alternativa. Utiliza-se assim técnicas de EUD, como a extensão, ampliando os significados em sistemas computacionais.

Assim, os sistemas devem suportar atividades dos usuários como projetistas. Deve ser possível inspecionar e elaborar modificações e extensões através de técnicas de EUD. Seguindo a teoria da Engenharia Semiótica, projetos de sistemas de computação que envolvem EUD devem [de Souza (2005)]:

- sintetizar o sistema de significação para suportar interação humano-computador;
- comunicar a visão do projeto através de um sistema específico de significação;
- comunicar as regras e princípios;
- comunicar se e como os princípios e regras podem ser mudados;
- comunicar como mudanças de significado podem efetivamente ser usadas na interação com a aplicação.

Assim, em sistemas EUD a metamensagem a ser transmita aos usuários deve englobar outros aspectos além dos já apresentados na metamensagem original. Isso porque os usuários se tornam co-criadores, ou seja, tornam-se co-autores da mensagem a ser transmitida através do sistema. A metamensagem então fica da seguinte forma [de Souza (2005)]:

"Eis minha interpretação de quem você é, o que aprendi que você tem ou quer fazer, preferencialmente de que formas e por quê. Eis, portanto, o sistema que concebi para você, o qual você pode ou deve usar assim, a fim de realizar uma série de objetivos associados com esta minha visão". ***Mas sei que você pode querer modificar minha visão, objetivando realizar coisas (de alguma forma) que eu não havia pensado. Posso lidar com as mudanças que você pode querer fazer, desde que você me fale o que quer nesse código específico."***

Porém, os usuários só podem se beneficiar das possibilidades de intervir no projeto de sistemas de computação se [de Souza e Barbosa (2006)]:

- entenderem o sistema de significados;
- formularem uma hipótese satisfatória sobre como os significados são codificados no sistema;
- dominarem o uso para comunicar intenções ao sistema e conseguir uma variedade das finalidades com elas;
- formularem hipóteses satisfatórias de como novos significados (ou os que mudaram) podem ser codificados;

- codificarem alguns significados no sistema e incorporá-los com as possíveis variedades de interação com a aplicação.

Uma análise semiótica em EUD é apresentada em de Souza e Barbosa (2006) mostrando o envolvimento de processos de significação em sistemas computacionais e o processo de comunicação humano. Os sistemas computacionais apresentam uma perspectiva de manipulação de símbolos que pode ser léxica, sintática ou semântica. Essas dimensões podem ser caracterizadas como:

- **léxicos:** usada para formar o vocabulário, para criar as sentenças da linguagem. São elementos da interface como botões, comandos, campos.
- **sintáticos:** são combinações dos elementos léxicos para alcançar os objetivos desejados.
- **semânticos:** permite criar novas significações dentro do sistema.

De forma a comparar o processamento de significação do computador e a perspectiva da comunicação humana, em [de Souza e Barbosa (2006)] é apresentado o esquema da figura 3.7.

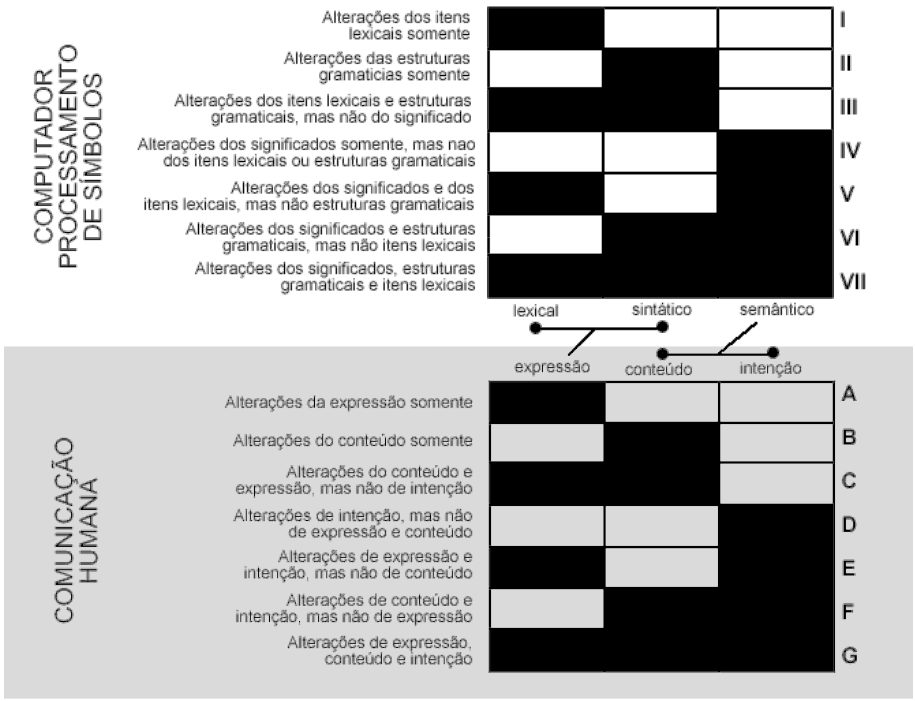

Figura 3.7: Dimensão semiótica de manipulações das linguagens[5] [de Souza e Barbosa (2006)]

A primeira tabela se refere ao processo de significação presente nos computadores e a segunda está relacionada à perspectiva da comunicação humana. Como se vê, não há uma correspondência direta entre as dimensões humanas e as da computação, sendo que o léxico e o sintático estão relacionados à expressão e o conteúdo e intenção humanos à dimensão semântica.

O processo de significação presente nos computadores pode ser dividido em dois grandes grupos, os que envolvem e os que não envolvem mudanças de significado. Essa pode ser considerada a diferença entre extensões e customizações [de Souza e Barbosa (2006)]. As customizações não

[5]A imagem foi traduzida pela autora deste trabalho.

apresentam mudanças nos significados, ou seja, na semântica da aplicação, sendo representados pelos tipos I, II e II da primeira tabela. Mudanças só em elementos léxicos e/ou sintáticos não representam mudanças em aspectos semânticos. Por exemplo, ao criar um ícone de uma "lixeira" e substituir pelo comando "excluir", trata-se de uma variação léxica, visto que nenhuma funcionalidade foi criada, mas apenas um novo elemento na interface foi inserido. A variação sintática pode ser exemplificada pela criação de macros, casos em que um comando executa um "pacote" de tarefas. Por exemplo, a abertura, mudança de tamanho e fechamento de um arquivo de uma imagem poderiam ser executados colocando apenas no comando "redimensionar imagem".

Os tipos IV, V, VI e VII apresentados na primeira tabela representam mudanças nos elementos semânticos, o que altera significados do sistema. Sendo assim, são consideradas extensões, onde novos significados até então não existentes são criados. Voltando ao exemplo da imagem, mudar a cor ou outra característica da imagem poderiam ser novas funcionalidades criadas, significados que não existiam.

Dentro do contexto do processo de significação, dois conceitos importantes que devem ser citados, relacionados à manipulaçao de sistemas significativos, são a identidade da aplicação e o conceito de impermeabilidade [de Souza e Barbosa (2006)]. Impermeabilidade remete a signos cujos significados são sempre preservados, ou seja, não podem ser alterados pelo usuário. Assim, a identidade de um sistema computacional pode ser definida como o conjunto mínimo de signos impermeáveis que a representam.

Em relação ao processo humano, os dois grandes grupos podem ser divididos a partir da variação da intenção, o que representa diferenciação entre customizações e extensões. Os tipos A, B e C representam os grupos que mantêm a intenção. Um exemplo simples aqui, que remete ao tipo A, pode ser ilustrado como a criação de novos *labels* em comandos, por exemplo, em que não há variação de intenção ou conteúdo. Já o outro grupo de alterações (D, E, F e G) envolvem a variação da intenção. Um exemplo desse grupo (tipo D) é quando o usuário associa novas intenções a componentes existentes, como no caso em que ele usa tabelas em HTML para ajustar textos em vários navegadores, sendo que seu uso inicial era apenas tabular dados.

Essa taxonomia criada tem uma grande importância no contexto de EUD, visto que apresenta uma caracterização teórica dos tipos de extensões possíveis e a diferenciação entre customizações e extensões. Mostra também a distância entre o processo de significação humana e a computacional e a relação existente entre eles. Outra contribuição da análise semiótica de extensões é que permite uma melhor análise entre os custos e benefícios das diferentes técnicas de EUD existentes.

***

Vista a teoria da Engenharia Semiótica e sua visão dentro do contexto de EUD, no próximo capítulo será apresentado o modelo proposto neste trabalho, suas características, arquitetura e aplicações.

Capítulo 4

Modelo proposto - EDeM

Como apresentamos no capítulo 2, sistemas de segunda geração exigem que os usuários conheçam conceitos técnicos envolvidos no contexto de mineração de dados para interagirem com os mesmos [Albergaria et al. (2006), Kriegel et al. (2007)]. Embora alguns usuários fiquem interessados em aprender esses conceitos necessários, outros percebem isso apenas como um alto custo. Assim, o modelo aqui apresentado, EDeM, *End User Development Conceptual Model*, propõe a arquitetura de um módulo a ser acoplado a sistemas de segunda geração, cujo objetivo é permitir que usuários criem extensões, possibilitando novas interações que não exigem conhecimento técnico [Albergaria et al. (2008a)]. Para isso, o módulo oferece aos usuários mecanismos para definir tarefas de mineração específicas relacionadas a um problema e contexto com uma interação direta que permite executá-las. Vale ressaltar aqui que, teoricamente, o modelo pode ser acoplado a qualquer sistema de segunda geração existente de mineração por regras de associação.

A idéia da solução proposta surgiu da estratégia que nos foi apresentava em uma entrevista com um usuário de sistema de segunda geração que tinha como contexto tarefas de auditoria de compras governamentais. Ele comentou que atuava como um "minerador" para sua equipe de trabalho. Ele utilizava o sistema para identificar padrões interessantes e, em seguida, apresentava os indicadores aos outros auditores para identificar o que deveria ser auditado.

Baseado nessa estratégia, o modelo considera dois perfis de usuários possíveis: o especialista e o leigo. O usuário especialista não apenas conhece o domínio de aplicação, mas também os conceitos técnicos necessários para interagir com sistemas de segunda geração. O usuário leigo (também chamado de usuário final) pode ser considerado como um perito no domínio de aplicação, mas não está disposto a "arcar" com os custos da aprendizagem de todos os conceitos técnicos a fim de se beneficiar do uso do sistema de mineração. No modelo, os usuários especialistas atuam como co-criadores de um nível de abstração de um problema e domínio específicos a ser utilizado pelo usuário leigo.

O modelo é proposto no âmbito de IHC, tendo como fundamentação teórica a Engenharia Semiótica, apresentada no capítulo 3, que considera a interação como um processo comunicativo. O modelo permite que usuários especialistas criem suas "mensagens" aos usuários leigos, possibilitando tanto uma comunicação indireta através da nova camada de interação, quanto uma comunicação direta sobre esta através da base de conhecimento, um componente do modelo que será descrito a seguir. Nossa solução permite que usuários especialistas se tornem co-criadores do sistema, ao criarem a camada de abstração. A arquitetura e funcionamento do modelo são descritos na próxima seção.

4.1 Arquitetura do modelo

Sem a aplicação do modelo, devido aos desafios existentes na interação de sistemas de segunda geração de mineração de regras de associação, todos os usuários desse tipo de sistema devem possuir o perfil de especialistas. Utilizando o modelo proposto, é possível que um grupo de usuários possua o perfil de leigos, onde eles não necessitem dos conhecimentos técnicos envolvidos. A figura 4.1 ilustra a seqüência de interação dos usuários leigos e especialistas utilizando o modelo. O primeiro passo consiste na necessidade de o usuário leigo obter algumas informações para seu negócio, dentro do seu contexto de trabalho, em que podem ser utilizadas técnicas de mineração de dados. Assim, ele solicita ao especialista que crie uma forma direta de interação, que consiste em abstrações para que ele possa interagir com o sistema de mineração de dados(1), isso para que ele não tenha que aprender os conceitos técnicos necessários. O especialista cria então abstrações de entrada de dados, de forma que o leigo possa executá-las de forma periódica e variável(2). O leigo então interage com o sistema, através das abstrações criadas, sem ter que conhecer os conceitos envolvidos no processo de mineração de dados, obtendo as informações que deseja(3).

Figura 4.1: Interação dos perfis dos usuários utilizando o modelo

O segundo passo apresentado na figura 4.1 requer todo um mecanismo de extensão em sistemas de segunda geração que visa garantir que esse mapeamento possa ser realizado. O modelo proposto neste trabalho envolve esse mapeamento e é constituído de três componentes: o **Gerador**, a **Linguagem Abstrata de Interface com o Usuário** (LAIU) e a **Base de Conhecimento**. O gerador é utilizado pelos usuários especialistas para a criação da LAIU; a base de conhecimento contém a concepção do usuário especialista em relação a personalização feita e a LAIU é a interface resultante com a qual o usuário leigo irá interagir. De forma resumida podemos ilustrar como os usuários interagem com os componentes. Os usuários especialistas interagem com o gerador para criar a linguagem de abstração (LAIU) com a qual os leigos irão interagir. Durante o processo de criação, ele entra com explicações sobre as decisões tomadas que são armazenadas na base de conhecimento e que são apresentadas em parte aos usuários leigos, de forma a facilitar o entendimento dos mesmos em relação às abstrações.

A estrutura do modelo pode ser visualizada na figura 4.2. A seguir, os componentes do modelo

são descritos de forma detalhada, explicando também como ocorre a relação entre eles.

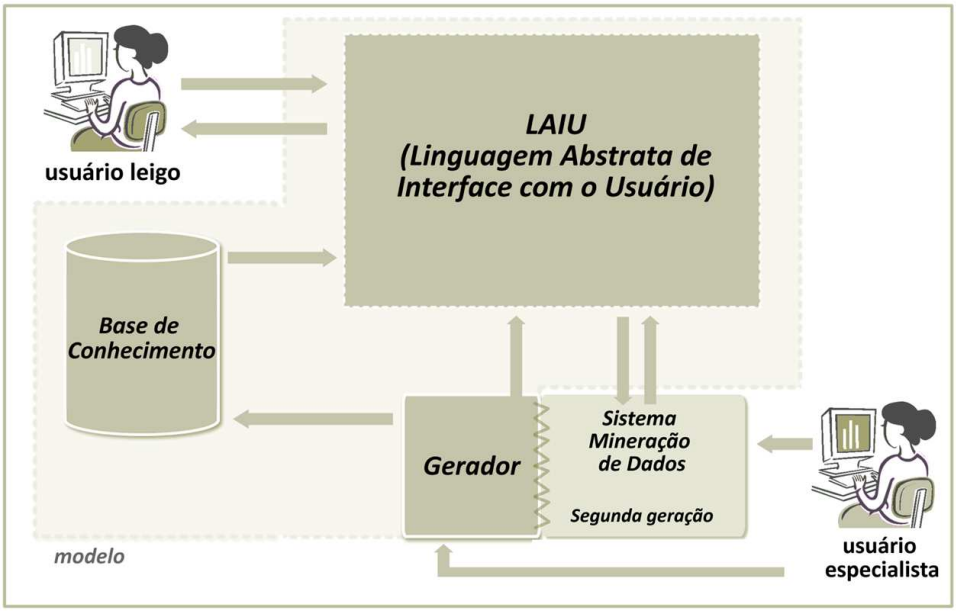

Figura 4.2: Modelo proposto

4.1.1 Linguagem abstrata de interface com o usuário (LAIU)

A LAIU (Linguagem Abstrata de Interface com o Usuário) é criada pelo especialista para o leigo. A LAIU é composta de duas partes distintas, o nível de abstração, sendo a parte léxica e sintática da linguagem e o interpretador, sendo a parte semântica da linguagem. O nível de abstração conta com a entrada e a saída de informações, que consiste nos elementos lexicais da linguagem.

O usuário leigo utiliza-se do nível de abstração para fazer uma consulta ao sistema. A consulta consiste em um elemento sintático, uma combinação de elementos lexicais da linguagem (i.e. uma sentença) que tem um significado (ou funcionalidade) associado. Uma consulta é uma abstração de uma tarefa de mineração de regra de associação na qual o especialista define quais parâmetros devem ser considerados no processo de mineração, quais são fixos e quais serão definidos pelos usuários finais durante a interação. Assim como o especialista especifica a consulta, ele também define como os resultados serão apresentados. Nesse caso, serão apresentados na parte da saída do nível de abstração, como elementos lexicais.

Voltando ao exemplo da auditoria (descrita na seção 2.4.1), a consulta "Durante o ano de <DEFINE_ANO>, algum fornecedor ganhou mais que <PORCENTAGEM> das ofertas de compra do produto <PRODUTO>?" é uma forma de interação dos usuários leigos. O especialista, ao criar essa consulta, já definiu os parâmetros relacionados aos conceitos técnicos de mineração. Assim, algoritmo, base, atributos e valores mínimos de parâmetros como suporte e confiança já foram selecionados. A correspondência entre os atributos e parâmetros também já foi feita, como definir que a <PORCENTAGEM> da consulta refere-se ao parâmetro confiança.

Um usuário final, interagindo com a camada de abstração, poderia realizar a seguinte consulta ao sistema: "Durante o ano de **2007**, algum fornecedor ganhou mais que **40%** das ofertas de compra do produto **toner**?". Esse seria um exemplo de entrada. Porém, os usuários especialistas devem

definir também como os resultados serão apresentados aos usuários leigos na camada de abstração. Isso porque a camada de abstração não estaria consistente se fosse mostrada aos usuários leigos uma regra de associação como resposta à consulta, como a apresentada abaixo.

[2006][ETA Inc.] => [toner] (12.05, 55.19)

Assim, o especialista deve, como foi feito para a entrada, definir a forma de abstração para a saída. Um exemplo seria apresentar a resposta como sendo: "ETA Inc. foi o fornecedor, em 55,19% das ofertas de compra produto toner." Dessa forma, permite-se ao usuário realizar uma consulta e obter a resposta sem conceitos técnicos envolvidos. Essa abstração deve ser feita pelo especialista. Porém, não há como o modelo garantir que haverá correspondência entre as formas de abstração da entrada e saída. O modelo tenta auxiliar os usuários na tarefa de abstração e assim apresenta elementos como o dicionário (a ser apresentado na próxima seção), que visa auxiliar os especialistas.

Assim como em qualquer linguagem, a LAIU apresenta elementos léxicos, sintáticos e semânticos. Os elementos léxicos compõem o texto da consulta, assim como do resultado, bem como os valores dos parâmetros. A sintaxe consiste nas possíveis combinações dos elementos léxicos, ou seja, as possíveis combinações que o usuário especialista pode gerar para criar uma consulta. Finalmente, a semântica refere-se ao comportamento da consulta quando essa é executada como uma tarefa de mineração pelo sistema de segunda geração. A camada de abstração contém todos os elementos da interface com os quais os usuários finais irão interagir.

Dentre os elementos que compõem a LAIU, está o interpretador. Ele funciona como uma forma de comunicação entre a camada de abstração, a aplicação de segunda geração e o gerador. Ele é responsável por receber uma consulta específica do usuário final e transformá-la em uma tarefa de mineração, onde essa transformação é feita em função das regras definidas pelo especialista no gerador. De forma análoga, o interpretador recebe o resultado gerado pelo sistema de segunda geração, que juntamente com as especificações feitas pelo especialista, cria o resultado final a ser apresentado na camada de abstração para o usuário final. Além disso, ele recebe da base de conhecimento as explicações dadas pelos especialistas para serem apresentadas aos usuários finais.

Na perspectiva da Engenharia Semiótica, a LAIU é uma metacomunicação principalmente do especialista para o usuário final. Isso porque ele comunica as decisões que foram tomadas ao serem criadas as abstrações, facilitando o entendimento das mesmas. Os especialistas são os principais autores das mensagens enviadas aos usuários finais, mas não são os únicos. O módulo de extensão criado define que tipos de consultas o usuário especialista pode gerar. Assim, os designers deste módulo também participam, isto é, são co-autores da mensagem que é enviada ao usuário leigo. Além disso, os designers dos sistemas de segunda geração especificaram inicialmente quais as técnicas e os dados da mineração poderiam ser utilizados para se definir o significado das as consultas. Assim, a LAIU é uma metamensagem composta pelo especialista, os projetistas dos sistemas de segunda geração e dos designers do módulo de extensão.

4.1.2 Gerador

Podemos definir o gerador como sendo a interface disponível para o usuário especialista que permite criar as abstrações (passo 2 da figura 4.1). O usuário especialista interage com o gerador visando criar a camada de abstração da LAIU para o usuário final poder interagir. O gerador

contém a especificação onde o usuário especialista define quais parâmetros devem ser considerados para a criação da camada de abstração. As regras de geração devem ser definidas tanto para os elementos relativos à entrada, quanto para a saída de dados e informações. Em outras palavras, o gerador permite ao usuário especialista criar uma nova interface para o usuário final, especificando os elementos com os quais será possível interagir, bem como o seu comportamento. Denominamos de consultas os elementos de entrada disponibilizados na camada de abstração para usuários leigos. Nesse caso, os especialistas definem os parâmetros fixos e os variáveis que serão definidos pelos leigos em tempo de interação (*template* de consulta). Em relação às visualizações, o especialista define como as regras de associação serão apresentadas aos leigos, quais tipos de visualizações serão possíveis e como os conceitos serão traduzidos.

Para ilustrar como o gerador trabalha, vamos voltar ao cenário de auditoria (seção 2.4.1) e à consulta citada na seção 4.1.1: "Durante o ano de <DEFINE_ANO>, algum fornecedor ganhou mais que <PORCENTAGEM> das ofertas de compra do produto <PRODUTO>?" Nesta consulta, o usuário final tem que atribuir valores a ANO, PRODUTO e PORCENTAGEM, conforme vimos anteriormente, que são conceitos bem conhecidos em seu domínio. Ao criar esta consulta, o usuário especialista definiu a relação desses valores aos parâmetros (por exemplo, <DEFINE_PORCENTAGEM> consiste no valor atribuído à confiança e <DEFINE_PRODUTO> é o conjunto das instâncias do atributo produto), bem como definiu o valor necessário para os outros parâmetros que não são mencionadas na consulta, como o suporte, a escolha de outros atributos e do algoritmo a ser utilizado. Todas essas definições são armazenadas no gerador.

4.1.3 Base de conhecimento

Como citado, o modelo proposto é baseado em Engenharia Semiótica, que considera a interação como um ato de comunicação. A base de conhecimento é utilizada nesse contexto para possibilitar ao projetista, no caso o especialista atuando como co-autor, documentar suas decisões. Assim, o especialista pode registrar suas intenções e decisões em relação às abstrações criadas, potencializando a qualidade da comunicação entre ele e o usuário leigo.

A base de conhecimento possui assim o objetivo de permitir que usuários especialistas registrem suas justificativas e explicações para as abstrações que criam. A base de conhecimento tem dois subcomponentes: as explicações e o dicionário. As explicações armazenam todos os esclarecimentos registrados pelos especialistas sobre suas decisões. As explicações são classificadas em dois níveis: as que são colocadas à disposição dos usuários finais (leigos), e as direcionadas aos outros usuários especialistas. As que se destinam aos leigos possuem explicações mais gerais, como o objetivo da consulta e as hipóteses que foram assumidas. Já as explicações para os especialistas são explicações técnicas que envolvem descrições sobre a seleção dos parâmetros, técnicas e até mesmo algoritmos.

Utilizando novamente o cenário da auditoria ilustrado em 2.4.1, na consulta "Durante o ano de <DEFINE_ANO>, algum fornecedor ganhou mais que <PORCENTAGEM> das ofertas de compra do produto <PRODUTO>?" o especialista poderia associar a seguinte explicação: *Essa consulta permite explorar os dados relativos às compras governamentais e identificar se, para um determinado ano, houve algum fornecedor favorecido. A hipótese é de que nenhum fornecedor pode ganhar todas as licitações. Em nossa experiência, já teremos um indício de fraude se o fornecedor ganhar acima de 40%. No entanto, é possível examinar candidatos para valores mais altos...*

Essa explicação permitiria ao usuário final compreender o significado da consulta, bem como o valor mínimo atribuído ao parâmetro que ela possui. Na perspectiva da Engenharia Semiótica,

essa explicação é muito importante para que o processo de comunicação designer-usuário tenha sucesso, que nesse caso é do usuário especialista para usuário final.

Além das explicações para usuários leigos, a base de conhecimento pode armazenar decisões técnicas tomadas pelo usuário especialista na criação de uma consulta. Um exemplo de explicação mais técnica poderia apresentar o motivo da escolha de alguns parâmetros como sendo flexíveis e outros fixos e até mesmo o motivo dos valores escolhidos como fixos poderia ser apresentado. Assim, para o exemplo, uma explicação técnica poderia ser: *O valor de suporte para essa consulta (que gera uma tarefa de mineração) foi de 0.27. Esse valor foi definido depois de várias tarefas serem executadas e esse pode ser considerado um valor relevante porque...* Essa explicação apresenta termos que são conceitos próprios de mineração e relacionados às decisões da modelagem do problema em tarefa de mineração de dados. Não é esperado que o usuário final entenda esse tipo de explicação, mas ela é importante para documentar a escolha de decisões técnicas.

O outro subcomponente da base de conhecimento é o dicionário. O dicionário contém os elementos que aparecem na LAIU e que têm a semântica correspondente no sistema de segunda geração. Por exemplo, a PORCENTAGEM que aparece na consulta acima teria uma entrada no dicionário onde seriam definidas suas dimensões léxicas e semânticas, além de uma explicação. O léxico é definido pelo especialista para compor a LAIU. A semântica é o que o termo representa no contexto de mineração de dados, ou seja, para a aplicação de segunda geração. Além disso, pode-se documentar uma explicação sobre o termo. Uma entrada para PORCENTAGEM poderia ser como a apresentada a seguir:

- **Léxico:** <PORCENTAGEM>% dos lances
- **Semântica:** Confiança
- **Explicação:** Como a confiança representa a freqüência relativa (ou probabilidade condicional) entre a ocorrência do evento no conseqüente e a ocorrência do evento no antecedente de uma regra, então no contexto da auditoria ela será usada...

O dicionário tem duas funções principais no modelo. A primeira é fazer com que o especialista pense sobre os elementos de sistemas de segunda geração que ele acredita que devem fazer parte da LAIU, bem como a forma como eles devem ser representados. A outra função é apoiar o especialista na manutenção da consistência entre os elementos que são mostrados na entrada e na saída da LAIU. As definições das abstrações de entrada e saída podem ser feitas em momentos diferentes e o dicionário pode ajudar a manter uma única representação na LAIU.

Embora a criação e execução de consultas não dependam da base de conhecimento, ela tem um papel fundamental na forma como as pessoas as utilizam e interpretam. Na perspectiva da Engenharia Semiótica, ela é um componente essencial, na medida que exige que os especialistas adicionem elementos metalingüísticos, que explicam outros elementos ou aspectos da LAIU. Esses elementos permitem que usuários especialistas, quando estiverem desempenhando papel de projetistas, enviem mensagens diretas sobre suas intenções ou decisões aos usuários leigos.

***

A Figura 4.2 apresenta os componentes do modelo e na Figura 4.3 as relações entre eles são detalhadas. A interação com o modelo inicia com o usuário especialista criando extensões, interagindo com o gerador. O objetivo a partir daí é criar o nível de abstração para o usuário leigo,

que compõe a LAIU. Ao criar uma extensão, são repassados ao interpretador as especificações da LAIU e as definições e explicações à base de conhecimento. O interpretador recebe da base de conhecimento as explicações inseridas, que serão apresentadas aos leigos junto com os resultados. O interpretador possui o papel de realizar a comunicação entre o gerador, o sistema de mineração de segunda geração e o nível de abstração da LAIU. A interação entre o interpretador (parte semântica da LAIU) e o nível de abstração (partes léxicas e sintáticas) se inicia com a interação com as consultas, onde os leigos definem valores para os parâmetros disponíveis. Os resultados também são "traduzidos" de acordo com as regras definidas pelos especialistas, onde os leigos podem modificar as visualizações utilizando os filtros disponíveis.

O modelo como um todo pode ser considerado uma ferramenta epistêmica para o especialista. Isso porque o faz refletir sobre as necessidades do usuário, como representá-las como tarefas de mineração de dados e como apresentá-las aos usuários leigos. A base de conhecimento é um componente essencial nesse contexto, uma vez que registra os entendimentos e decisões do projetista (nesse caso o especialista) tanto em relação à abstração para o usuário, quanto ao seu significado no sistema de mineração.

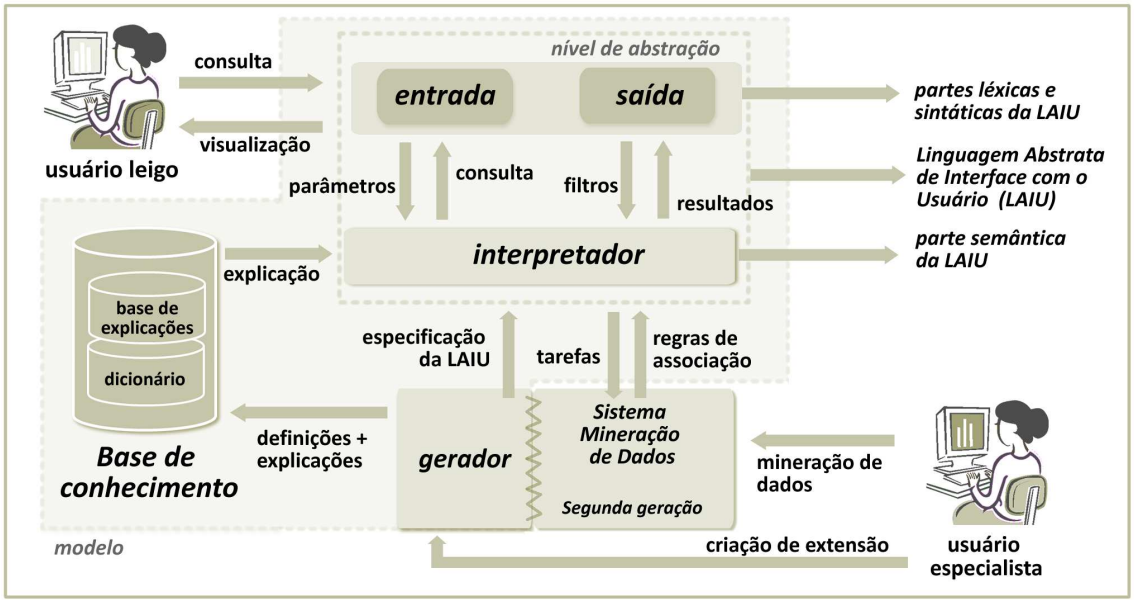

Figura 4.3: Modelo proposto

É importante ressaltar que o modelo pode ser acoplado a qualquer sistema de segunda geração, que pode continuar sendo utilizado pelo usuário especialista para realizar tarefas de mineração de dados diretamente, caso seja desejado. Acoplado ao modelo, ele recebe tarefas de mineração e retorna como resultado um conjunto de regras de associação, passos apresentados na seção 2.2. Como afirmamos anteriormente, teoricamente o modelo pode ser acoplado a qualquer sistema de mineração de segunda geração. Isso porque o interpretador possui o papel de gerar e apresentar as consultas, sendo assim deve existir uma linguagem para importar dados e um formato de comunicação entre interpretador e sistema.

A seguir é apresentada uma caracterização e análise das extensões geradas pelo modelo.

4.2 Análise das Extensões Geradas

Como já descrito, o nosso modelo permite que sistemas de segunda geração sejam extensíveis, de forma a inserir nesses tipos de sistemas a possibilidade de se criar abstrações e, com isso, permitir que um maior número de usuários possam utilizá-lo. Com isso, criamos novas possibilidades de acessar as informações.

Como apresentado na seção 3.2.1, existem diferentes formas de caracterizar e analisar extensões. Fischer apresenta uma relação custo x benefício das linguagens existentes em função do custo de aprendizagem e escopo de sua aplicação [Fischer et al. (2004)]. Considerando estes critérios temos que o custo de aprendizagem para o especialista é bem baixo, pois não requer novos conhecimentos, apenas o que já tinha para criar uma tarefa na interface do sistema de segunda geração. Em relação ao escopo, podemos também observar que é baixo, visto que só permite a criação de consultas no sistema a que o módulo foi acoplado. Assim, a possibilidade de criar a extensão descrita aqui pode ser considerada como uma linguagem de domínios específicos, possuindo baixo custo de aprendizagem, porém baixo escopo de aplicação também.

A análise apresentada por [de Souza e Barbosa (2006)] descreve uma classificação baseada em análise dos processos de significação de sistemas computacionais e de seres humanos. No processo de significação de sistemas computacionais, são consideradas dimensões de manipulações de símbolos os itens lexicais, sintáticos e semânticos. As extensões são consideradas as modificações que envolvem mudanças nos significados (semântica). Nos processos de significação de seres humanos são consideradas dimensões de expressão, conteúdo e intenção.

Analisando as extensões oferecidas pelo EDeM de acordo com essa perspectiva, temos que para os sistemas de mineração de dados, elas preservam o significado, visto que só podem ser manipulados itens lexicais e sintáticos Pode-se criar novos itens lexicais (termos usados nas consultas) e sintáticos (que são as próprias consultas). Entretanto, elementos semânticos não, visto que os usuários não podem alterar as funcionalidades já existentes.

Em termos da perspectiva de comunicação dos usuários especialistas, as extensões possuem as intenções preservadas. Isso porque eles não podem criar nada que já não podiam antes, ou seja não têm novas intenções associadas ao sistema. No entanto, permite introduzir sinônimos no sistema de significação, visto que as consultas podem ser vistas como macros, que agilizam o trabalho dos especialistas. Macros são caracterizadas quando um comando aciona uma serie de eventos, que são repetidos freqüentemente. Nesse sentido, há variação de expressão, existindo um novo elemento de interação.

Em relação aos usuários leigos, eles não estão diretamente envolvidos na criação das extensões, mas as consultas são destinadas a eles. Para os leigos, são as extensões que proporcionam novas intenções, conteúdo e expressão, uma vez que antes eles não possuíam acesso ao sistema. Nesse caso, para os leigos as consultas são novos elementos, onde eventos são encapsulados em forma de macros, visando abstrair conceitos envolvidos. Para os leigos, o benefício é alto, visto que permite interagir com sistemas e técnicas que até então ele não utilizava.

Em relação ao custo de criação das extensões, para os especialistas, o custo de aprendizado e execução são baixos, pois ele já possui o conhecimento técnico necessário. O maior custo é relacionado à geração das explicações, mas que além de ser um grande benefício para os leigos serve como documentação sobre as decisões tomadas.

A seguir são apresentadas as etapas de avaliação iniciais realizadas para o modelo.

4.3 Avaliação

A solução proposta pelo modelo foi baseada em duas hipóteses. A primeira foi que os usuários especialistas poderiam criar abstrações (consultas) úteis para os usuários finais, que não conhecem os conceitos técnicos de mineração. A segunda foi que o modelo pode ser acoplado (implementado) a um sistema de mineração de segunda geração. Essa seção apresenta os esforços para analisar a primeira hipótese, sendo que a segunda é discutida no capítulo 5.

Buscando avaliar a utilidade real do modelo, uma análise foi feita de forma a verificar se os usuários especialistas conseguiriam realmente criar um nível de abstração em um domínio relevante para os usuários finais. Uma avaliação inicial nesse sentido foi feita utilizando cenários [Carroll (2000)], sendo realizada em duas etapas. A primeira consistiu em se partir de uma tarefa de mineração já criada em um sistema de segunda geração e seus resultados e então gerar uma abstração (sem os conceitos técnicos envolvidos), onde um exemplo é apresentado na seção 4.3.1. O objetivo dessa etapa era verificar se seria possível criar uma abstração a partir dos dados da mineração. O segundo passo foi verificar se outros usuários especialistas, que não participaram na criação do modelo, eram capazes de criar propostas de abstrações usando cenários para diferentes domínios, onde os resultados estão apresentados na seção 4.3.2.

4.3.1 Abstração de uma tarefa de mineração

Como citado, o primeiro passo da avaliação foi tomar um problema modelado como uma tarefa de sistemas de mineração de dados e criar uma proposta de abstração com o qual o usuário final pudesse interagir e obter os resultados desejados. O problema escolhido foi de um real cliente do grupo de pesquisa em mineração de dados, o departamento de auditoria do Estado de Minas Gerais. O foco da tarefa foi em relação ao favorecimento de fornecedores e a mineração foi feita usando o sistema de segunda geração, o Tamanduá (mesmo problema descrito na seção 2.4.1).

Tendo a modelagem original da tarefa, o objetivo era criar um cenário descrevendo uma abstração que seguisse a proposta do modelo. Assim, foi criada a consulta:

"Durante o ano de <DEFINE_ANO>, algum fornecedor ganhou mais que <PORCENTAGEM> das ofertas de compra do produto <PRODUTO>?"

Através dessa abstração, não é necessário que o usuário leigo tenha os conhecimentos técnicos relacionados a mineração. Para criar essa consulta, o especialista selecionou algoritmos, base, atributos e valores de parâmetros, por exemplo, deixando para o leigo uma forma de interação pronta e relacionada ao seu contexto de trabalho.

Uma vez que os parâmetros da modelagem tenham sido definidos, deve-se também configurar a forma de saída. Neste caso, precisava-se gerar uma regra geradora que permitisse que uma regra de associação fosse traduzida em uma abstração. Como primeiro passo, optou-se por traduzir a regra de associação em um texto. Assim, foi definida a regra geradora da forma: A; E ===> B;C;D; onde A, E, B, C, D são atributos que formam a regra. A quantidade de atributos na regra não é limitada e pode ser representada pelo *template* textual apresentado abaixo:

- Em CONFIANÇA% das vezes que A OCORRE e E OCORRE, B OCORRE e C OCORRE e D OCORRE...
  Este padrão ocorreu SUPORTE*TAMANHO_BASE vezes na base de dados BASE_DADOS

onde:

- CONFIANÇA: é a confiança da regra;
- OCORRE: é o significado inerente a cada atributo. Este significado é dado pelo usuário especialista ao construir o modelo. Por exemplo, sendo os atributos: produto, valor e fornecedor, o significada de OCORRE dado para cada atributo foi:
  - produto: é comprado
  - fornecedor: fornece
  - valor: é o valor do produto
- SUPORTE: é o suporte da regra
- TAMANHO_BASE: é o número de registros da base de dados
- SUPORTE*TAMANHO_BASE: é o número de registros da base em que a regra se verifica
- BASE_DADOS: nome da base de dados

Essas definições são as explicações dadas para os termos no dicionário. Assim, uma interação do usuário leigo envolve execuções de consultas e visualização de resultados, como apresentado a seguir:

- "Durante o ano de **2007**, algum fornecedor ganhou mais que **40 %** das ofertas de compra do produto **cartucho**?"
- Em **49.20%** das vezes que **cartucho** é comprado, **Happy Printer Inc.** fornece.

  Este padrão ocorreu **437** vezes na base de dados **Compras_Governo.**

A abstração criada mostrou como o modelo poderia apoiar a criação de consultas que seriam interessantes para os usuários finais. Assim, a primeira hipótese foi analisada de forma a verificar que é possível criar abstrações úteis para os usuários finais, sem envolver os conhecimentos técnicos relacionados.

O próximo passo foi verificar se outros usuários especialistas seriam capazes de criar consultas relevantes para diferentes domínios. Essa análise está apresentada na próxima seção (4.3.2).

4.3.2 Cenários de aplicação

Essa fase da avaliação foi realizada como parte de um projeto de classe para o curso de Mineração de Dados, no nível da graduação (com alguns alunos também da pós). Normalmente, o trabalho final da disciplina requeria que os alunos desenvolvessem um projeto de mineração de regras de associação modelando um problema real e apresentando os resultados reais. No projeto desenvolvido no segundo semestre de 2007, os estudantes também tiveram que criar um nível de abstração a ser apresentado a usuários finais. Em outras palavras, tiveram que definir questões que usuários finais estariam interessados em solicitar e identificaram qual seria o domínio de respostas específicas que poderia ser gerado do conjunto de regras de associação. Os cenários criados foram analisados de forma a verificar se eles abstraíam de fato os conceitos técnicos e se realmente poderiam ser criados a partir dos componentes definidos no modelo.

De forma geral, os projetos foram desenvolvidos em grupos de 2 estudantes. Dos 14 grupos que iniciaram o projeto, 12 terminaram e, desses, 8 foram avaliados como tendo alcançado os objetivos do projeto, ou seja, tendo conseguido fazer a modelagem desejada a partir dos cenários definidos. Os motivos pelos quais alguns trabalhos foram considerados como não tendo atingido os objetivos do projeto envolvem várias questões, como a produção de trabalhos incompletos ou cuja modelagem da tarefa de mineração não foi considerada satisfatória[6].

Foram sugeridos alguns temas para o projeto em sala de aula, mas os alunos também poderiam encontrar outros de seu interesse. O projeto exigiu que os alunos interagissem com um usuário final, buscando definir e analisar um contexto real de aplicação, onde alguns projetos foram disponibilizados, assim como usuários. Eles também tiveram acesso às bases de dados necessárias para modelagem do problema. Os projetos considerados foram feitos em 3 diferentes domínios: Temperatura e consumo de energia elétrica em diferentes edificações (1 grupo), Criminalidade em uma cidade (1 grupo) e qualidade de questões do vestibular de uma universidade (6 grupos).

Todos os 8 grupos foram capazes de criar bons níveis de abstração (entrada e saída de dados). Por sucesso, consideramos que uma abstração poderia ser aplicada em um determinado problema, onde não houvesse necessidade de entender os conceitos técnicos pelos usuários finais. Um grupo superou o que foi solicitado e efetivamente implementou uma consulta em um sistema de mineração, o Weka [Weka (2006)]. No projeto foi ainda solicitado aos alunos que explicassem suas consultas, assim como a modelagem feita para o problema. Um resultado interessante foi que, embora não tenha sido apresentado aos usuários o dicionário, a maioria dos trabalhos apresentou um mapeamento entre as abstrações das consultas e os elementos da interface do sistema de mineração.

De forma a ilustrar as abstrações propostas, foram criados cenários de uso[7]. A seguir são apresentados exemplos de cenários que foram obtidos nos trabalhos gerados.

Vestibular

O primeiro cenário de aplicação foi em relação a qualidade das questões de vestibular de uma determinada instituição. A base de dados disponibilizada para o trabalho contém informações dos anos de 1995 a 2005. Ela apresenta dados sócio-econômicos, assim como notas nas provas dos vestibulares e em diversas disciplinas na graduação dos candidatos aprovados.

O usuário real do contexto, considerado leigo nas técnicas de mineração de dados, desejava analisar a validade das questões do vestibular serem preditores de desempenho do aluno na graduação. O usuário leigo apresentou seu problema, contexto e uma "classificação" apresentada na figura 4.4 a ser aplicada nas questões do vestibular. Em relação à classificação, um aluno pode errar uma questão no vestibular e ter fracasso em uma determinada disciplina, ou grupo delas, na graduação. Pode errar a questão, mas ter sucesso posteriormente na graduação; acertar a questão e ter sucesso ou acertar e ter fracasso. Assim, as boas questões são as consideradas, segundo o leigo, as que conseguem pré-determinar o desempenho do aluno nas aulas (erro:fracasso ou acerto:sucesso). Um aluno que acerta uma questão (ou obtém sucesso em uma prova específica no geral) de física, por exemplo, e depois vai bem na matéria de física, pode demonstrar que a questão foi uma boa forma

[6] A avaliação do modelo foi conduzida após o término da disciplina. Para isso, foi solicitada então a autorização dos alunos para a utilização de seus trabalhos para este fim. O termo de consentimento utilizado encontra-se no Apêndice C.1.

[7] Cenários foram definidos em [Carroll (2000)] como narrativas textuais plausíveis e detalhadas que descrevem uma situação específica.

de seleção. Já o aluno que vai muito bem numa prova de matemática e depois tem fracasso em várias disciplinas relacionadas, como cálculos, pode demonstrar que a prova não está sendo uma boa seleção dos alunos.

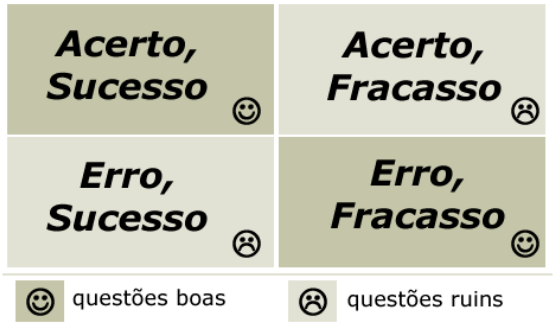

Figura 4.4: Classificação das questões de vestibular, segundo a visão do usuário leigo

Diante dessa demanda, os alunos buscaram apresentar premissas para criarem algumas consultas relacionadas e que poderiam ser úteis para o usuário final. Muitas vezes foi necessário um pré-processamento da base, de forma a organizar e limpar os dados a serem minerados. Por exemplo, o atributo nota foi discretizado em alguns casos, sendo apresentado em conceitos. Assim, foram consideradas, por exemplo, faixas de valores para um determinado conceito: até 50 (F); de 50 a 60 (E) e assim por diante. Além disso, foram criados novos atributos ou critérios como: alunos que tiraram acima de 70% foram bem e que tiraram abaixo foram mal, onde o atributo criado FOI_BEM era preenchido com S(sim) ou N(não), de acordo com as notas tiradas na disciplina.

Assim, foram criadas consultas relacionadas às necessidades do leigo, como a exemplificada abaixo:

- "Quais as melhores questões do vestibular de <ANO> da matéria <MATERIA_ VESTIBULAR>?"

De forma a analisar o desempenho dos alunos no vestibular e nas disciplinas, foram considerados pares de dados matéria_do_vestibular x matéria_da_graduação. Essa relação foi considerada coerente e baseada nos dados existentes. Exemplos de relações feitas são apresentadas a seguir:

- Física - Fundamentos de Mecânica
- Matemática - Cálculo Diferencial e Integral I
- Química - Química Geral

A explicação apresentada para a consulta deve tentar auxiliar os usuários a entenderem o que foi modelado. Um exemplo de explicação dada foi: *"Esta tarefa tentará relacionar as questões de uma das provas do vestibular de um determinado ano com o desempenho no primeiro período letivo dos alunos aprovados. O usuário deve selecionar um ano específico e uma das matérias do vestibular para realizar a consulta. O resultado será dado relacionando as questões da prova escolhida com o desempenho das matérias cursadas pelo aluno."* Além dessa explicação, a premissa que foi adotada em relação a considerar uma questão boa deve ser descrita (como encontrar as melhores?) Uma explicação sobre a consideração feita foi: *"As relações entre as matérias do vestibular e da graduação foram adotadas da seguinte forma[...] Essa relação foi assim feita por serem consideradas matérias*

(do vestibular e da graduação) relacionadas e a relação entre elas ser coerente. Uma questão foi considerada boa quando o conceito obtido no vestibular foi igual ou diferente de um conceito em relação a matéria da graduação. Os conceitos utilizados foram[...] Assim, se um aluno tirou B em uma questão de física, a questão será considerada boa se o aluno tirar A, B ou C na matéria de Fundamentos de Mecânica[...]" Todas as explicações e descrições que foram desenvolvidas são consideradas no modelo como parte da base de conhecimento.

Como resultado, foram geradas regras onde os antecedentes eram nome da prova, número da questão e conceito da questão e como conseqüente o conceito da disciplina. As formas de apresentação sugeridas nos trabalhos para os resultados foram diversas. Tabelas, gráficos, imagens e textos foram formas que apareceram. Para essa questão em especial, foi apresentada uma tabela com o número da questão, conceito da questão, conceito da disciplina e percentual de ocorrência (confiança). Templates textuais também foram sugeridos, como o exemplo abaixo:

- Dos <TAMANHO_BASE> alunos do departamento <NOME_DEPT> que fizeram o vestibular de <ANO> e cursaram <DISCIPLINA>, <TAMANHO_BASE>*<SUPORTE> (<SUPORTE>%) tiveram <RESULTADO_DISCIPLINA> nessa disciplina e tiveram <RESULTADO_QUESTAO> na prova de <MATERIA_QUESTAO>.
- Dos <TAMANHO_BASE> alunos do departamento <NOME_DEPT> que fizeram o vestibular de <ANO> e cursaram <DISCIPLINA>, <TAMANHO_BASE>*<SUPORTE> (<SUPORTE>%) tiveram <RESULTADO_DISCIPLINA> nessa disciplina e tiveram <ACERTO_QUESTAO> na questão <NUMERO_QUESTAO> de <MATERIA_QUESTAO>.

Os atributos e parâmetros são apresentados entre as marcações "<>", o que representa que serão preenchidos pelos valores existentes. Exemplos de resultados obtidos, onde as regras já foram "traduzidas" segundo o *template* definido são apresentados a seguir:

- Dos **1000** alunos do **departamento de física** que fizeram o vestibular de **2005** e cursaram **fundamentos de mecânica**, **700** (**70%**) tiveram **sucesso** nessa disciplina e tiveram **sucesso** na prova de **física**.
- Dos **1000** alunos do **departamento de física** que fizeram o vestibular de **2005** e cursaram **fundamentos de mecânica**, **700** (**70%**) tiveram **sucesso** nessa disciplina e tiveram **acerto** na questão **3** de **física**.

Temperatura e consumo de energia elétrica

O segundo cenário aqui descrito, apresentado nos trabalhos, foi em relação ao controle de consumo de energia elétrica em algumas edificações. Esse monitoramento é importante para analisar o gasto com energia, levantando possíveis formas de melhor aproveitamento e economia.

A base de dados consiste nas medições feitas em indústrias e edificações, que identifica o consumo individual (equipamentos) e geral de energia. Dentre as consultas elaboradas no trabalho para os usuários reais selecionados, uma buscava analisar o comportamento em relação ao consumo de energia de um determinado equipamento quando submetido a uma temperatura específica. A seguir, a consulta sugerida está apresentada.

- Qual comportamento dos equipamentos, em relação ao consumo de energia, no <MES>, de acordo com a faixa de <TEMPERATURA> do período analisado?

Como já citado, as explicações associadas às consultas podem ser técnicas ou direcionadas para os usuários finais. Em relação às informações técnicas, uma explicação foi dada da seguinte forma: *"Para gerar essa consulta, o valor escolhido para confiança foi de 0.7, isso porque acima desse valor é que foram encontradas regras interessantes para a questão. O valor para suporte foi de 0.01 porque representa 2% da base, sendo esse valor considerado relevante."*

Como forma de apresentação para usuário final, foi proposto uma visualização tabular, onde é apresentada a intensidade da informação em formas de barras. A freqüência é outro valor apresentado, que consiste na confiança da regra. A freqüência de ocorrência é apresentada graficamente pela de quantidade "estrelas", onde quanto mais estrelas, maior o valor associado à confiança.

***

De forma geral, ambas as etapas de avaliação foram positivas ao demonstrar que um nível de abstração poderia ser criado de forma eficiente. Foram criadas consultas relevantes e úteis a serem aplicadas nos contextos reais dos usuários leigos, tanto as abstrações criadas a partir de tarefas, quanto as propostas por usuários especialistas. No terceiro cenário desenvolvido, relacionado à criminalidade, as abstrações também foram criadas de forma satisfatória. Esse cenário será apresentado na seção 5.3, onde foi utilizado para avaliações com os usuários reais.

Após essa avaliação preliminar do modelo, o próximo passo foi verificar que um módulo baseado no modelo proposto poderia de fato ser implementado e acoplado a um sistema de mineração já existente, o Tamanduá [Tamandua (2006)]. Esse passo do trabalho consiste no protótipo apresentado no próximo capítulo.

Capítulo 5

Protótipo

No capítulo 4 vimos o modelo e uma avaliação preliminar de como ele possibilita a criação de abstrações por usuários especialistas para usuários leigos. Continuando a análise do modelo, este capítulo refere-se à implementação do modelo em um sistema de mineração de regras de associação de segunda geração. Assim, foi criada uma instância do modelo de forma a avaliar seu funcionamento e estrutura. Para isso, um protótipo foi desenvolvido e foram feitas avaliações com a participação de usuários especialistas e leigos.

A seguir serão apresentados o protótipo e o sistema de segunda geração utilizado como base, o Tamanduá, além da arquitetura do mesmo e o custo de acoplamento do protótipo, ou seja, as adaptações necessárias no sistema para que o protótipo pudesse funcionar. Além disso, serão apresentadas também as avaliações realizadas com o protótipo e a análise dos resultados obtidos.

5.1 Tamanduá

O sistema Tamanduá foi desenvolvido no departamento de Ciência da Computação da UFMG com o objetivo de permitir a pesquisa básica e aplicada relacionada à mineração de dados distribuída [Tamandua (2006)] [Ferreira et al. (2005)]. Podemos considerar o Tamanduá um sistema de mineração de objetivo geral, no sentido que ele procura oferecer aos usuários a oportunidade de encontrar padrões interessantes em uma base de dados, sem focar em nenhum domínio específico. O Tamanduá já vem sendo utilizado por algumas instituições públicas brasileiras, com aplicações em diferentes contextos, dentre os quais: segurança pública, saúde e compras governamentais. Ele tem apoiado a gestão governamental em tarefas de auditoria e tem sido utilizado também como ferramenta de análise por cientistas sociais.

O Tamanduá é uma plataforma que visa proporcionar serviços de mineração de dados de forma escalável e eficiente, possuindo algumas características como:

- A interoperabilidade é baseada na utilização e extensão de padrões abertos e internacionalmente reconhecidos para a construção de serviços web, mineração e armazém de dados;
- A escalabilidade refere-se à sua arquitetura modularizada, o que permite fácil replicação e adaptação desses componentes para os variados cenários de uso da plataforma, assim como a utilização de uma plataforma computacional paralela baseada em máquinas de baixo custo;
- O paradigma da computação utilizado é o orientado a serviços, o que permite que cada servidor seja instanciado mais de uma vez, tornando a solução flexível e escalável.

Cada tarefa executada no Tamanduá pode demandar uma grande carga de dados e processamento. Buscando apoiar essa demanda, o Tamanduá foi concebido como um conjunto de componentes distribuídos que oferecem os seus serviços através de interfaces bem definidas, de modo que possam ser usados para satisfazer as necessidades da aplicação.

A versão atual do Tamanduá dá apoio às seguintes fases do processo de mineração:

- **Seleção dos dados:** escolha pelo usuário de quais atributos da base serão utilizados;
- **Engenharia dos dados:** formatação e ajuste dos dados para possibilitar a execução das técnicas de mineração;
- **Determinação de padrões:** execução do algoritmo em si, de acordo com a técnica escolhida;
- **Análise dos padrões:** análise dos resultados obtidos utilizando técnicas de visualização de padrões.

A arquitetura do Tamanduá pode ser vista na figura 5.1. E a seguir iremos descrever cada componente.

- **Servidor de Aplicação:** O servidor de Aplicação é responsável por garantir o controle de acesso aos dados e serviços oferecidos.
- **Servidor de Mineração:** é responsável por disparar as requisições de mineração realizadas pelos usuários, controlando os nodos de execução definidos no sistema. Ele é responsável pela comunicação entre os outros servidores garantindo a correta execução da lógica de negócio proposta pelo Tamanduá.
- **Servidor de Dados:** O Servidor de dados (SD) é responsável pela interface para acessar todos os conjuntos de dados e metadados associados. O SD pode obter descrições das bases de dados e os seus metadados, realizar a transferência de bases de um SD para outro e executar consultas SQL sobre os dados.
- **Servidor de Processamento:** O Servidor de processamento (SP) executa os algoritmos, processa um conjunto de dados e produz novos conjuntos como resultado. Para lidar com grandes bases de dados e os custos computacionais associados, o SP é distribuído através de um cluster.
- **Servidor de Visualização:** O Servidor de visualização recebe um conjunto de dados como entrada (normalmente o resultado de uma mineração anterior) e opera em analisar os dados e produzir uma representação visual.

A figura 5.2 apresenta o ciclo de vida de descoberta de conhecimento utilizando o Tamanduá [do Tamandua (2005)]. Uma tarefa de descoberta de conhecimento é executada através dos seguintes passos:

1. O **servidor de aplicação** contém as interfaces, serviços web disponibilizados através de páginas web; além disso, ele realiza o controle necessário de acesso aos dados. Assim, ele é responsável pela entrada dos dados dos usuários.

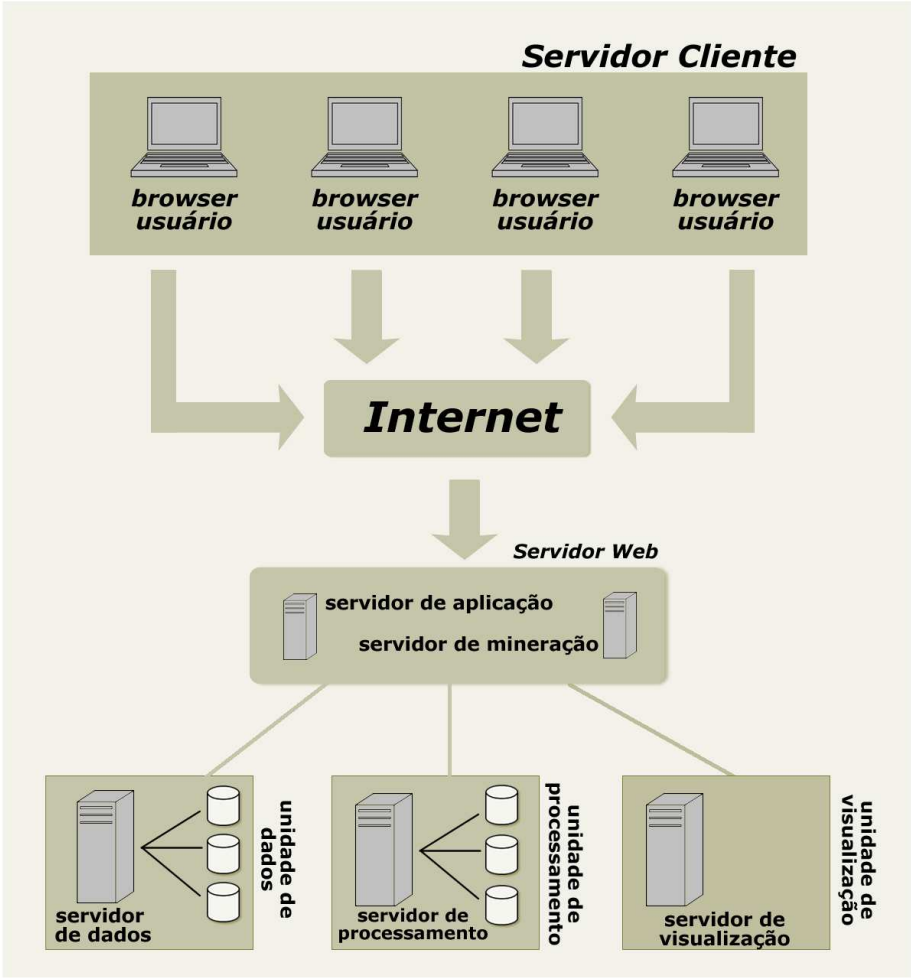

Figura 5.1: Estrutura do Tamanduá

2. O **servidor de mineração** recebe os dados de entrada e é responsável por acionar os outros servidores; ele controla todo o processo.

3. Os dados a serem minerados ficam armazenados no **servidor de dados** que envia partições ao servidor de mineração.

4. O **servidor de processamento** é responsável pela mineração em si, executando os algoritmos de mineração.

5. Por último, os dados são tratados pelo **servidor de visualização** que trata os resultados a serem apresentados.

A próxima seção apresenta o protótipo desenvolvido, suas características e definições.

5.2 O protótipo

Como já citado, a princípio, o modelo proposto pode ser acoplado a qualquer sistema de mineração por regras de associação. Neste trabalho, o modelo foi implementado desenvolvendo um

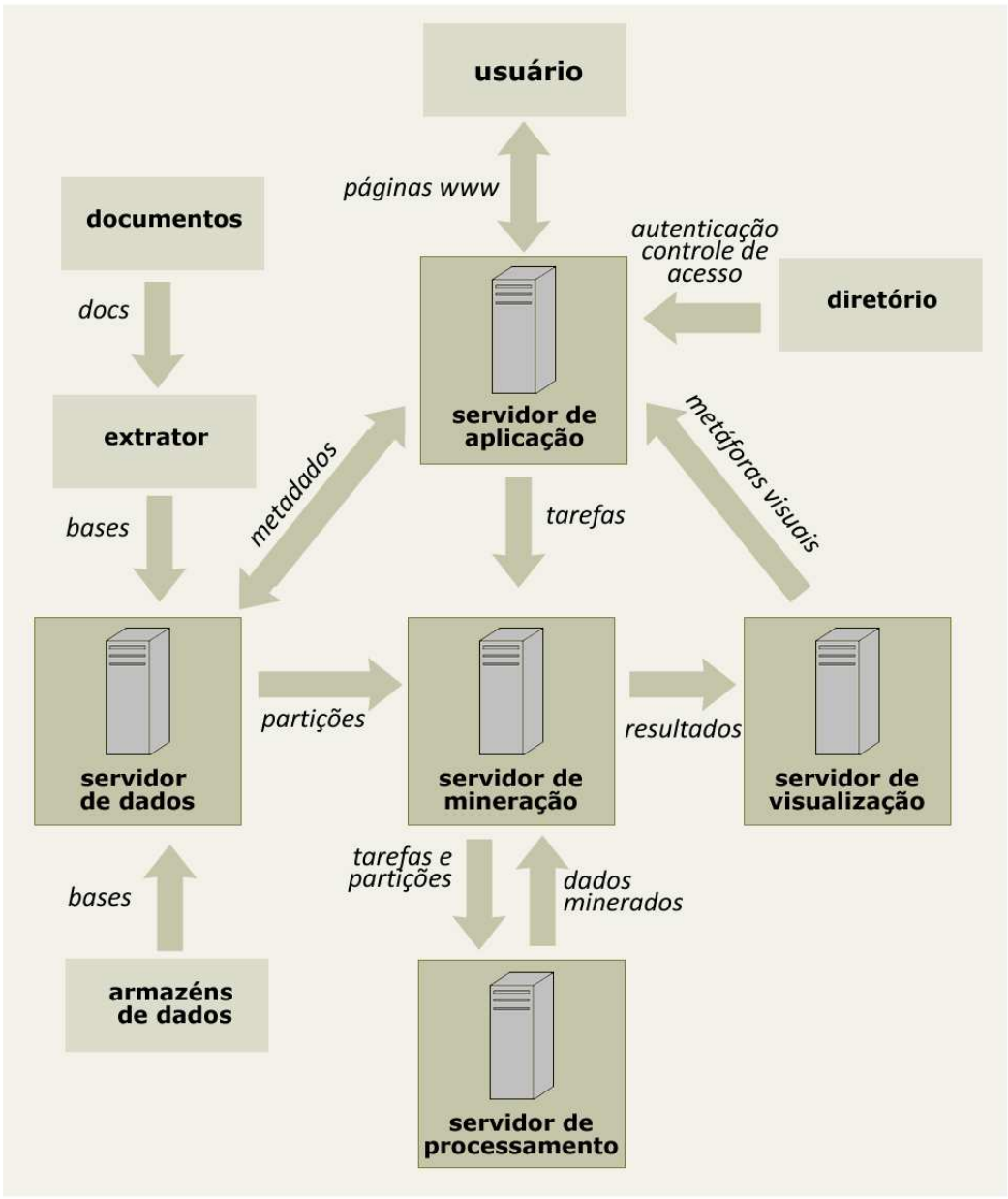

Figura 5.2: Ciclo de vida de descoberta de conhecimento utilizando o Tamanduá

protótipo utilizando o sistema de segunda geração Tamanduá. O protótipo foi desenvolvido na plataforma Java, utilizando Java Server Faces como suporte a interfaces criadas com conceito Ajax [Kuranov et al. (2001)]. Uma primeira versão do protótipo foi desenvolvida com o objetivo de demonstrar que é possível implmentar um módulo extensível baseado no modelo EDeM.

A seguir são apresentadas as modificações que foram necessárias na arquitetura do sistema Tamanduá para a implementação do modelo, além da descrição de conceitos e modelagens que foram criadas [Mourão et al. (2008)]. O objetivo de descrever essas alterações consiste em mostrar o custo para acrescentar ao sistema escolhido o módulo extensível.

5.2.1 Adequação do tamanduá

As mudanças necessárias na arquitetura do Tamanduá para a implementação do modelo estão apresentadas na figura 5.3. Foram necessárias três novas unidades: unidade de usuários, servidor de extensões e um novo servidor de visualização, para geração dos resultados abstraídos. A unidade dos usuários (*home*) armazena os arquivos pessoais de cada usuário, como tarefas e visualizações. Já a unidade de extensão possibilita ao usuário especialista a interação com o sistema para criação das extensões. Ela contém os componentes do modelo responsáveis por gerar a camada de abstração dos usuários leigos: o gerador, o interpretador e a base de conhecimento. Para armazenamento dos dados e comunicação entre os componentes (interpretador, gerador, etc.), foi necessária a criação de uma linguagem denominada *Pheromone*, nos moldes definidos pelo formato de XML (*EXtensible Markup Language*). Essa solução foi adotada por se tratar de uma linguagem flexível e extensível, visto que XML permite criar *tags* específicas para o domínio. Além disso, é padronizada e é utilizada para armazenar informações, não sendo necessárias mudanças em bancos de dados. Um exemplo do *Schema XML* da linguagem criada por ser visto no Apêndice A. Os detalhes da linguagem, descrição das *tags* e hierarquia entre elas podem ser encontrados em [Mourão et al. (2008)].

Em relação ao novo servidor de visualização, ele é responsável por gerar a visualização dos resultados a partir das informações fornecidas pelo interpretador. O interpretador recebe do gerador as especificações da LAIU, analisa e passa a esse servidor essas informações, um *template* para que os resultados obtidos possam ser apresentados na camada de abstração.

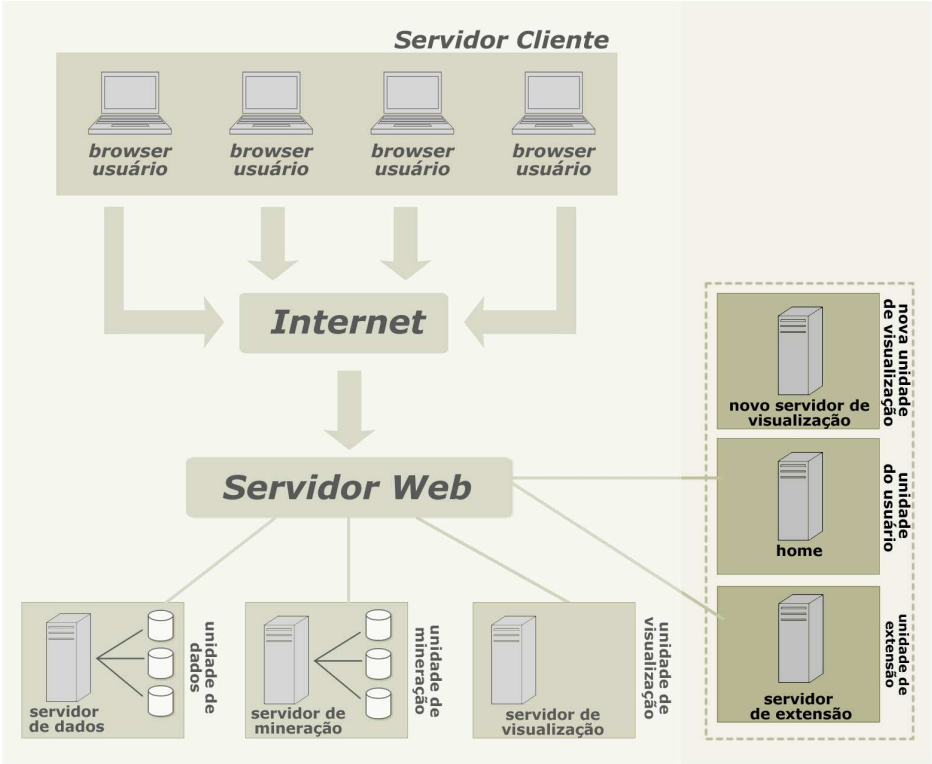

Figura 5.3: Estrutura Nova - Tamanduá

Outras modelagens e mudanças foram realizadas em relação ao Tamanduá ao desenvolver a versão do protótipo. Entretanto, algumas modificações não eram necessárias para a implantação

do modelo e sim optou-se por realizar melhorias no sistema já existente nesse momento. As modificações citadas anteriormente (apresentadas na figura 5.3) eram mesmo necessárias para o modelo. Já em relação às melhorias, por exemplo, a estrutura do Tamanduá foi remodelada de forma a conter 3 camadas. Essa modificação não era necessária, mas trouxe melhorias ao sistema como um todo. Algumas dessas modificações feitas além de trazerem melhorias para a implementação do Tamanduá também foram motivadas também pela vantagem que trariam ao acoplamento do módulo extensível. Por exemplo, em relação às três camadas, a partir dessa decisão foi criada a classe *Extension Manager* para que fosse possível tornar o modelo mais flexível e independente do sistema Tamanduá.

A próxima seção apresenta algumas modelagens e definições que foram realizadas para o desenvolvimento do protótipo e seu acoplamento ao Tamanduá. É importante ressaltar que o custo das modificações necessárias a serem feitas no sistema de segunda geração para implementação do modelo varia para cada sistema, pois depende da arquitetura e implementação de cada um. O modelo é independente do sistema de mineração a ser acoplado, mas alterações em relação à comunicação e interação são inevitáveis. O objetivo de mostrá-las aqui é dar uma idéia do custo real para um sistema, no caso o Tamanduá.

5.2.2 Modelagem e definições do protótipo

Como já citado, em relação à estrutura do Tamanduá foram feitas novas modelagens de forma a conter 3 camadas distintas (Figura 5.4). A primeira refere-se à interface, onde os usuários interagem com o sistema. Existem aqui dois componentes de interação, um que contém as funcionalidades normalmente existentes no sistema de segunda geração e outro que abrange as funções que compõem o nível de abstração criado pelo especialista. A segunda camada refere-se a de entidades, que possui como componentes *Action Manager* e *Extension Manager*. O *Action Manager* é responsável por todas as funcionalidades do sistema de segunda geração. Já o *Extension Manager* executa todas as funções relacionadas às funcionalidades de extensão e abstração. A última camada consiste nos dados persistentes. São os locais onde os dados são armazenados e de onde são consultados. Modelando dessa forma, o objetivo era deixar as camadas mais independentes, tornando a implementação do modelo a mais flexível possível e o modelo implementado mais independente.

A partir da modelagem em camadas, é possível separar as funcionalidades específicas do sistema de segunda geração, que nesse caso abrangem o próprio sistema, o *Action Manager* e os servidores de dados, mineração e visualização, das funções que são próprias do modelo, que abrangem o nível de abstração, o *Extension Manager* e o servidor de extensão. Dessa forma, os componentes que fazem parte do modelo só se comunicam em dois pontos com o sistema de segunda geração: entre nível de abstração e *Action Manager* e entre *Extension Manager* e os servidores de dados, visualizações e minerações. A vantagem dessa comunicação ser em pontos localizados envolve aspectos como organização, facilidade de mudanças, se necessárias e independência entre funcionalidades.

Ainda em relação à modelagem, foram elaborados diagrama de classes e diagrama entidade-relacionamento(ER), apresentados no Apêndice A. Nesses diagramas, é possível visualizar a organização e relação entre as entidades criadas na implementação do protótipo (grupos, consultas, tarefas, entre outras). Para o modelo, não é necessário criar um banco de dados específico, é suficiente apenas estender o existente de forma a comportar o armazenamento de entidades que devem ser criadas. Entretanto, no caso do Tamanduá uma remodelagem geral foi feita de forma a melhorar a já existente. Um dos conceitos adotados no protótipo foi o de grupo. Um grupo

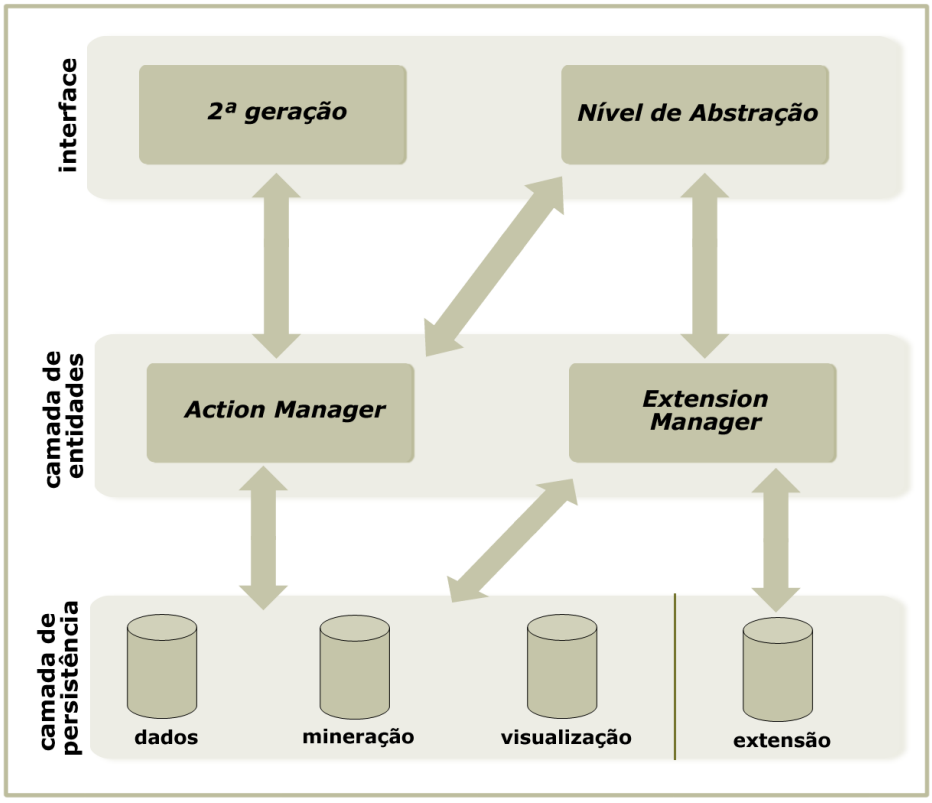

Figura 5.4: Estrutura em Camadas

representa um conjunto de usuários com os mesmos objetivos inseridos no mesmo contexto. Cada grupo é representado por seus usuários, bases e algoritmos. Por exemplo, é possível criar um grupo de auditores que podem trabalhar em bases de compras do governo e podem executar determinados algoritmos. Cada usuário no grupo deve possuir um perfil, que determina suas responsabilidades e permissões.

A criação desses perfis foi baseado no modelo, que conforme apresentado no capítulo 4, considera dois papéis possíveis para os usuários: o especialista e o leigo. Em relação a essa característica, no sistema foram modelados esses dois perfis, além do administrador. O administrador é responsável por gerenciar todos os recursos do sistema. Ele é capaz de realizar todas as funcionalidades, acrescentando e excluindo grupos, usuários, bases e algoritmos. O usuário especialista tem permissão para criar abstrações, (sendo as de entrada denominadas consultas) e especificar como serão apresentadas suas saídas, além de também ter acesso às funcionalidades disponíveis para os leigos. Já o usuário leigo participa de um grupo e pode executar as consultas criadas e compartilhadas dentro do grupo, além de visualizar os resultados obtidos, aplicando-os em seu contexto.

Cada um dos usuários, especialista ou leigo, pode possuir N consultas, sendo que cada consulta possui N instâncias de tarefas de mineração (1 para cada execução da consulta com diferentes parâmetros). Como já citado anteriormente, consultas são as abstrações de entrada dos dados presentes na interface do protótipo. Elas são as abstrações feitas pelo especialista que são acessadas pelos leigos, são questões relacionadas ao contexto de aplicação que podem ser variadas e executadas pelos usuários finais. A visualização das mesmas também é configurada pelos especialistas de forma a abstrair os resultados obtidos. Ao escolher alguns parâmetros que foram deixados pelo

especialista em aberto em uma consulta e mandar "executar", é criada uma tarefa, sendo essa uma "linha de comando" a ser executada pelo algoritmo de mineração em si. Assim, cada nova execução é denominada tarefa de mineração.

Cada consulta possui também descrições que compõem a base de conhecimento. Para cada etapa da criação, são apresentadas algumas questões a serem respondidas, que irão constituir as explicações vistas pelos leigos ao interagirem com a consulta. Para cada questão é explicitado o perfil a quem aquela informação se destina. No caso de informações técnicas são direcionadas aos especialistas e as que envolvem explicações gerais são mais direcionadas aos leigos.

Essa característica é muito importante, visto que a Engenharia Semiótica, teoria utilizada como base para o modelo, entende a interação humano-computador como uma comunicação entre designer e usuário. E a base de conhecimento possui um papel relevante em apresentar ao usuário leigo o que o designer, no caso o especialista, estava pensando no momento de construir a abstração.

Todo o desenvolvimento do protótipo foi baseado na arquitetura do modelo descrito no capítulo 4. Entretanto, nesta primeira versão, nem todas as funcionalidades desejadas foram implementadas. Algumas das funcionalidades desejadas, mas que não foram implementadas nessa primeira versão do protótipo são:

- Desenvolvimento do dicionário de forma completa, podendo em todos os casos editar os valores das dimensões léxicas, sintáticas e semânticas;
- Disponibilidade dos parâmetros de suporte e confiança para criação do corpo das consultas (exemplo, criar consultas da forma: "Quais os compradores ganharam acima de <CONFIANCA> das vezes..."). A versão atual só permite a inserção de atributos;
- Disponibilidade de operadores matemáticos, como "menor que", "igual a" para valores na criação das consultas;
- Possibilidade do especialista criar novas formas de visualização como tabular e gráfica e de filtros para interação com os resultados.

Todas essas funcionalidades estão descritas e documentadas para posterior implementação. É importante ressaltar que embora essas funcionalidades sejam desejáveis para um uso mais amplo do protótipo, a versão atual do protótipo permite a criação de consultas úteis de acordo com o modelo e permitem a apreciação tanto do modelo, como do próprio protótipo. As funcionalidades não implementadas são apresentadas como trabalhos futuros a serem desenvolvidos no protótipo. De forma a ilustrar a versão criada do protótipo, as suas principais telas estão apresentadas no Apêndice B.

Após ser implementado o protótipo, avaliações foram realizadas com o objetivo de avaliá-la. A seção a seguir apresenta as avaliações realizadas, além dos resultados obtidos.

5.2.3 Utilização do protótipo

Para demonstrar a utilização do protótipo, foram criadas consultas relacionadas aos contextos citados em 4.3.2. As consultas foram criadas pela própria autora e tinha por objetivo avaliar se o protótipo permitia a criação das consultas geradas nos cenários utilizados para uma avaliação inicial do modelo, descrita na seção 5.3. Foram utilizados dois cenários distintos, o de vestibular (descrito na seção 4.3.2) e o de criminalidade, que será utilizado como exemplo a seguir. O outro contexto, relacionado à análise de Temperatura e Consumo de energia elétrica, não foi utilizado

por não termos acesso à base de dados. Nos contextos utilizados foi possível executar tarefas e obter os resultados de forma textual. Isso porque esse tipo de visualização é considerada básica e pode ser utilizada em todos contextos e consultas.

De forma a ilustrar a utilização do protótipo, a seguir serão apresentados os passos realizados no mesmo para a criação, execução e visualização de uma consulta criada no contexto de criminalidade (Outras telas do protótipo estão apresentadas no Apêndice B). O contexto de criminalidade envolve a análise de dados relacionados à criminalidade de Belo Horizonte, que são cadastrados e analisados pelos responsáveis pela segurança pública da cidade. Esse contexto também foi utilizado nos testes com usuários, descritos na seção 5.3.

Os primeiros passos então são relacionados à criação da consulta em si. A consulta que será ilustrada consiste na seguinte questão: "Dado os crimes que ocorrem pela <|PERIODO DO DIA|> no dia da semana <DIA DA SEMANA>, o que podemos afirmar sobre os tipos de crimes e locais de ocorrência?". Para criação da mesma, são realizados os seguintes passos no protótipo:

1. **Informações:** Fornecimento de um nome e descrição para a consulta a ser criada (Figura 5.5);
2. **Algoritmo:** Escolha do algoritmo a ser aplicado e fornecimento dos valores para os parâmetros suporte e confiança (Figura 5.6;)
3. **Base:** Escolha da base e seleção dos atributos que serão utilizados no processo de mineração (Figura 5.7);
4. **Atributos:** Visualização das instâncias dos atributos, podendo selecionar instâncias específicas e dar nomes mais "claros" aos atributos para os usuários, quando necessário (Figura 5.8);
5. **Consulta:** Criação da consulta em si. Atualmente com a possibilidade de inserir atributos no corpo das mesmas, de forma a deixar a escolha dos valores para os leigos no momento da interação (Figura 5.9).

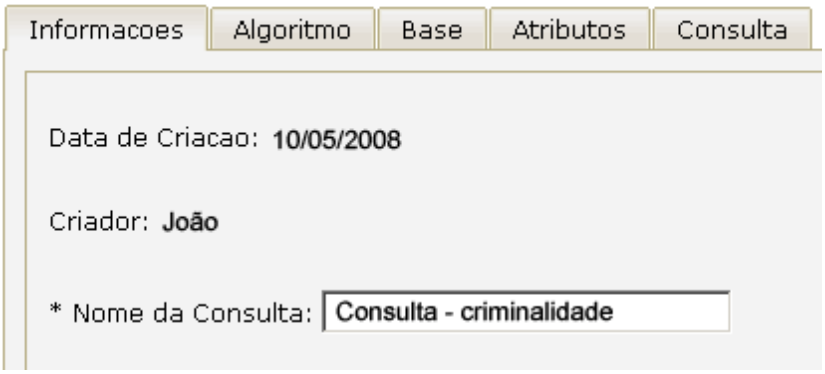

Figura 5.5: [Informação] Tela de criação da consulta

É importante ressaltar que em cada tela há perguntas relacionadas às consultas para os especialistas que visam compor a explicação que será visualizada por usuários leigos ou especialistas, dependendo do caso. Por exemplo, na tela de "Algoritmo", existem as seguintes questões a serem respondidas (nesse caso, direcionadas aos especialistas):

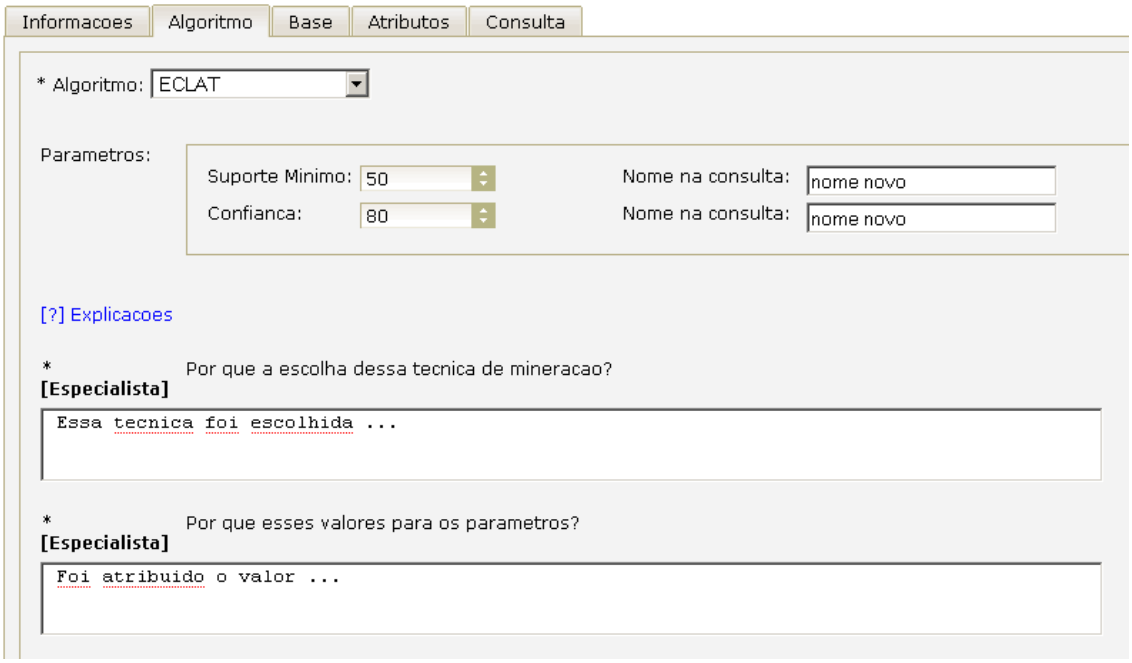

Figura 5.6: [Algoritmo] Tela de criação da consulta

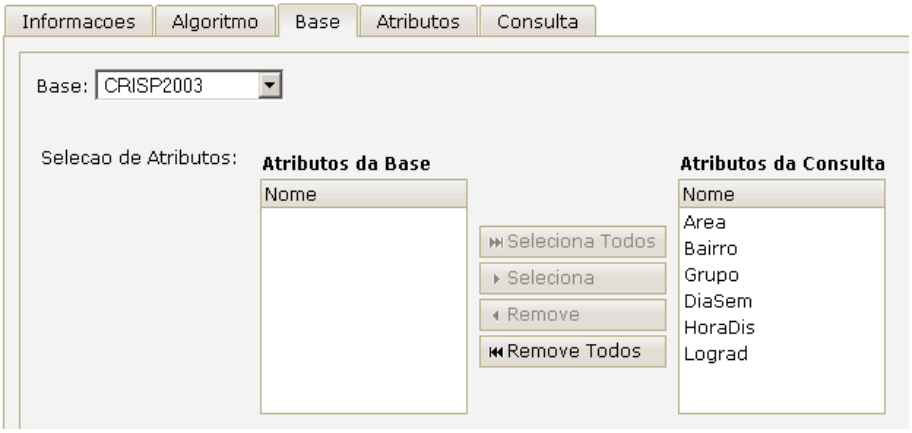

Figura 5.7: [Base] Tela de criação da consulta

- Por que a escolha dessa técnica de mineração?
- Por que esses valores para os parâmetros?

O objetivo das perguntas é guiar os usuários especialistas quanto ao conteúdo das explicações de interesse a serem fornecidas.

***

A forma de apresentação da consulta criada para os usuários leigos pode ser vista na figura 5.10. Para executar uma consulta (o que chamamos de uma tarefa), o usuário deve escolher os valores que deseja para os atributos, variando-os em novas execuções sempre que desejar. Esse

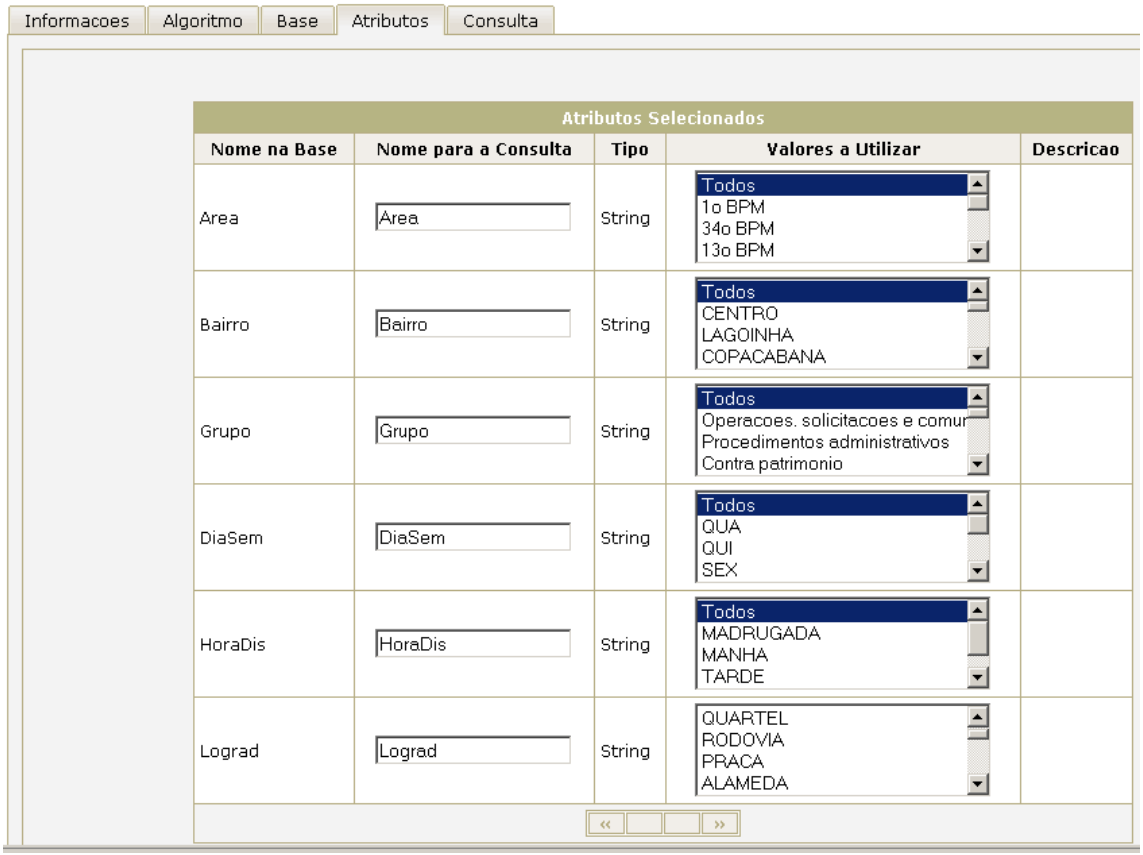

Figura 5.8: [Atributos] Tela de criação da consulta

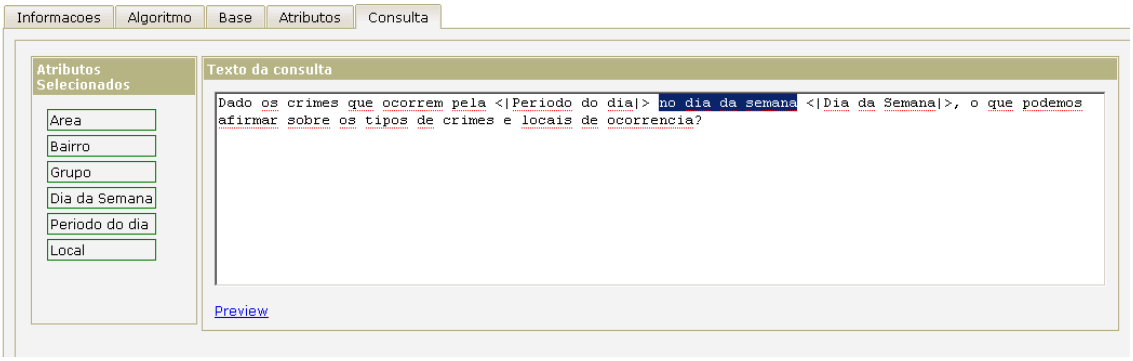

Figura 5.9: [Consulta] Tela de criação da consulta

tipo de consulta é interessante ser executada de forma periódica e para novos dados, sempre que surgirem.

O usuário especialista deve também configurar a forma de apresentação dos resultados, como eles serão vistos pelos usuários leigos, sendo que para isso utilizam um *template default* de visualização textual (Figura 5.11).

A tela apresentada na figura 5.12 representa a visualização final dos resultados pelos usuários leigos.

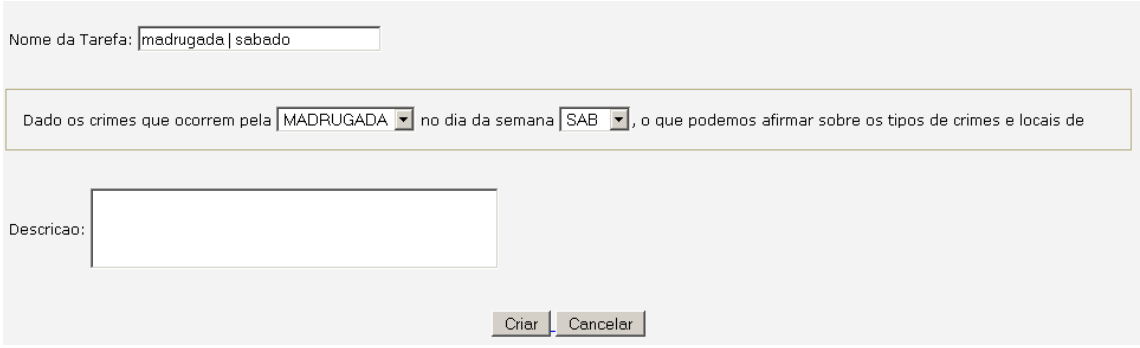

Figura 5.10: Visualização da consulta segundo visão do usuário leigo

Figura 5.11: Tela de configuração textual

5.3 Avaliação do protótipo com usuários reais

A avaliação do modelo usando o protótipo foi realizada com a participação dos usuários reais do sistema, buscando analisar se a solução atendia ao que se propunha. Assim, nesse momento, desejava-se analisar a visão dos usuários sobre a solução proposta, utilizando o protótipo. Mais uma vez, os cenários utilizados para o teste foram os do vestibular e criminalidade (descritos na seção 4.3.2). Para essa fase da avaliação, foram preparados testes em ambiente controlado, criando tarefas relacionadas aos cenários selecionados. Buscava-se analisar a experiência dos usuários, tanto especialistas quanto leigos, com o uso do sistema, coletando a opinião dos mesmos, verificando se a solução atende ao que se propõe e analisando a percepção dos usuários sobre os custos e benefícios do protótipo (e modelo).

A avaliação realizada pode ser dividida em três grandes etapas: planejamento dos testes, aplicação dos testes (incluindo entrevistas) e análise dos resultados. A seguir são descritas as atividades de cada uma dessas etapas.

5.3.1 Planejamento dos testes

O planejamento da avaliação foi a primeira etapa e teve os passos descritos a seguir.

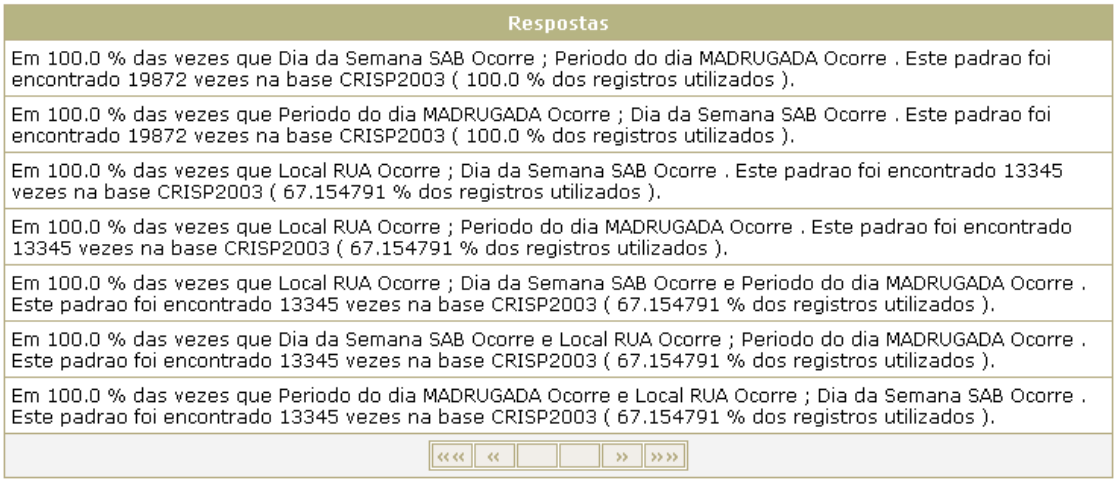

Figura 5.12: Tela de visualização textual final

- **Determinação do objetivo da avaliação:**

  O objetivo da avaliação com os usuários foi definido como sendo a resposta para as questões:

  É possível que especialistas consigam criar abstrações (consultas) úteis para os usuários finais? É possível que os leigos consigam interagir com o sistema de mineração de segunda geração sem que saibam os conceitos técnicos envolvidos?

- **Determinação do contexto de aplicação:**

  Após definido o objetivo da avaliação, foram escolhidos os contextos para aplicação dos testes: a qualidade de predição das questões em relação ao desempenho dos alunos (contexto do vestibular, descrito na seção 4.3.2) e a análise de dados sobre a criminalidade em uma cidade.

  Uma preocupação foi em limpar e organizar a base de dados de cada contexto. As bases de dados foram fornecidas pelos próprios usuários. As descrições das bases utilizadas estão mostradas na tabela 5.1.

- **Seleção dos participantes:**

  Em seguida foram selecionados os participantes para os testes. Para cada contexto, foi selecionado um usuário especialista e um leigo, onde cada um desenvolveu tarefas relacionadas ao seu perfil.

  Em relação ao perfil dos usuários, no contexto da análise da criminalidade, tivemos a participação de dois usuários leigos distintos, mas com características semelhantes. Os dois são pesquisadores do CRISP, Centro de Estudos de Criminalidade e Segurança Pública, que consiste em um órgão voltado para a elaboração, acompanhamento de implementação e avaliação crítica de políticas públicas na área da justiça criminal; ligado à Universidade Federal de Minas Gerais (UFMG) [Crisp (2008)]. Os dois participantes são sociólogos, fizeram mestrado na área e são funcionários do CRISP. Durante todo o processo de avaliação, eles participaram em momentos diferentes, onde um participou da reunião com usuários especialistas e leigos realizada para levantamento das necessidades e o outro no teste presencial realizado.

Contexto	Descrição das bases
Vestibular	A base de dados original contém as notas das provas de segunda etapa dos vestibulares da UFMG dos anos de 2002, 2003, 2004 e 2005, assim como notas dos alunos na universidade. Para melhor análise dos dados, os mesmos foram anonimizados, de forma a não permitir que os candidatos/alunos fossem identificados, sendo retiradas também informações pessoais. Para os testes, foi utilizado parte da base, com os dados do ano de 2004. Informações dos atributos e discretizações realizadas estão apresentadas no Apêndice C.2.
Criminalidade	A base de dados original foi disponibilizada pelo CRISP (Centro de estudos de criminalidade e segurança pública), órgão ligado a UFMG. A base apresenta cerca de 636.000 registros ocorridos em 2003 da cidade de Belo Horizonte, composta por 28 atributos, como natureza, descrição e código da ocorrência, além de data e hora do acontecimento. Alguns exemplos de crimes apresentados são: trânsito urbano, contra pessoa, incêndio, busca e salvamento, entre outros. Em relação a preparação da base, discretizações foram realizadas, como a criação de faixas de horários para o atributo hora: manhã, tarde, noite e madrugada. A lista completa dos atributos está no Apêndice C.2.

Tabela 5.1: Descrição das bases de dados de cada contexto

Em relação ao usuário leigo do contexto de vestibular, contou-se com a participação de um pesquisador, professor do departamento de computação no DCC, que tinha interesse em levantar aspectos interessantes sobre a análise da qualidade do vestibular.

Os usuários especialistas selecionados são alunos de graduação e pós-graduação do curso de computação da UFMG para os contextos de criminalidade e do vestibular, respectivamente. Como experiência em mineração de dados, eles cursaram a disciplina de mineração de dados no segundo semestre de 2007. Para o especialista do contexto de vestibular, esse foi o único contato que teve com o sistema Tamanduá 1.0 (versão já existente). Já o especialista do contexto de criminalidade havia participado também de um curso externo, há aproximadamente 2 anos, sendo esse de curta duração (aproximadamente 2 horas). Os dois participaram dos trabalhos descritos na seção 4.3.2 como alunos, mas não tiveram outros contatos com os usuários leigos envolvidos, além dos que ocorreram durante o curso.

- **Criação dos cenários para as avaliações:**

  Com o objetivo de descrever cenários mais reais e tarefas que seriam interessantes para os usuários, antes dos testes em laboratório, foram realizadas algumas reuniões com os usuários. A seqüência e descrição das atividades realizadas são apresentadas a seguir:

  - **Reunião com os usuários especialistas selecionados (roteiro da reunião no Apêndice C.2)**

    O objetivo da primeira reunião, com os usuários especialistas, consistia em apresentar o trabalho de forma resumida e explicar os passos que seriam executados durante todo o processo de avaliação. Além disso, levantar o interesse e disponibilidade dos participantes.

  - **Reunião com os usuários especialistas e leigos (uma para cada contexto; roteiros das reuniões no Apêndice C.2)**

    O segundo encontro ocorreu de forma separada para cada contexto. Assim, foi realizada uma reunião com o especialista e leigo do contexto de criminalidade e outra com o especialista e leigo do contexto do vestibular. Nesse momento, o objetivo da avaliação e descrição da base de dados foram apresentados de forma a igualar o conhecimento dos usuários envolvidos sobre o contexto de aplicação. O objetivo da reunião era levantar as necessidades dos usuários leigos de forma a auxiliar a criação das abstrações pelos usuários especialistas.

  - **Reunião com os usuários especialistas para geração das consultas (uma para cada contexto; roteiros das reuniões no Apêndice C.2)**

    Em um terceiro momento, foram feitos encontros separados com os especialistas, de forma a definir as consultas que deveriam ser criadas, já sendo definido os textos das mesmas. Previamente, para cada contexto de uso foi elaborado um resumo sobre as necessidades que foram levantadas na reunião anterior.

- **Geração do material para as avaliações (incluindo aspectos éticos).**

  Após as reuniões, foram elaborados os testes separadamente com cada usuário utilizando o sistema. Primeiro foram realizados os testes dos especialistas e posteriormente dos leigos. Os roteiros das avaliações estão listados a seguir:

  - **Avaliação em laboratório com os usuários especialistas (roteiro da avaliação na seção 5.3.2)**
  - **Avaliação em laboratório com os usuários leigos (roteiro da avaliação na seção 5.3.2)**

A próxima seção apresenta a descrição dos testes realizados em laboratório com os usuários.

5.3.2 Aplicação dos testes

Os testes ocorreram no laboratório próprio para avaliações do DCC/ICEx/UFMG. Cada teste contou com a participação de um usuário por vez e durou cerca de 1,5 hora para os especialistas e 1 hora para os leigos. O laboratório onde os testes foram realizados é totalmente equipado e adequado para realização das avaliações, tendo um espaço para os usuários interagirem com o sistema e outro ambiente fechado para os avaliadores observarem e fazerem as anotações necessárias. Todas as interações realizadas durante os testes foram gravadas utilizando o software *SnagIt* para documentação e posterior análise, caso necessário.

Durante os testes, ocorreram alguns problemas de desempenho nas versões do sistema Tamanduá. A versão do protótipo desenvolvida apresenta-se um pouco instável, em fase de testes e ajustes. Porém, os pequenos problemas ainda existentes não inviabilizaram os testes, que foram realizados com sucesso.

Como citado, as avaliações ocorreram separadamente para cada usuário. Os primeiros testes foram executados pelos especialistas e foram divididos em duas etapas, em um mesmo encontro. Na primeira etapa, os usuários criaram tarefas no Tamanduá 1.0[7] de forma a modelar as consultas que seriam geradas (informações como definição dos atributos e valores dos parâmetros). Em seguida,

eles utilizaram o Tamanduá 2.0 (protótipo) para criarem as consultas de acordo com as modelagens elaboradas.

Em um segundo momento, os usuários leigos realizaram os testes, onde tinham que interagir com as consultas criadas pelos especialistas, visualizando e interpretando os resultados obtidos.

Durante cada teste (com cada usuário), os seguintes procedimentos foram realizados:

- **Recepção do participante e explicação sobre o teste a ser realizado - Apêndice C.2**
- **Explicação e assinatura do termo de consentimento (tem como objetivo a autorização dos participantes dos testes para a gravação das avaliações e utilização dos resultados obtidos para a pesquisa aqui descrita - Apêndice C.2)**
- **Entrega do cenário e tarefas a serem executadas pelo participante - Apêndice C.2**
- **Condução dos testes, gravando toda a interação e fazendo as anotações necessárias**
- **Realização de entrevista pós-teste juntamente com levantamento da opinião do participante em relação à experiência de uso - Apêndice C.2.**

Como descrito, durante o teste, após execução das tarefas solicitadas, uma entrevista pós-teste foi feita, buscando coletar a opinião dos usuários. Os resultados obtidos foram então analisados, juntamente com as observações feitas durante as interações e a seção seguinte apresenta as conclusões obtidas.

5.3.3 Análise dos resultados obtidos

Como descrito anteriormente, o objetivo da avaliação consistia em:

É possível que especialistas consigam criar abstrações (consultas) úteis para os usuários finais, onde esses consigam interagir com o sistema de mineração de segunda geração sem que saibam os conceitos técnicos envolvidos?

Assim, o que se queria analisar é a proposta da solução e não o protótipo em si. Como descrito, os testes realizados contaram com a participação de 4 usuários, sendo 2 especialistas e 2 leigos, um de cada perfil para cada um dos contextos. Foram testes preliminares, não sendo possível afirmar que os resultados obtidos nos testes são conclusivos ou que abrangeram todos os aspectos possíveis. Mas a partir deles, foram detectados diversos indicadores de que a solução proposta atende ao que se propõe. Para uma primeira avaliação com os usuários, os resultados foram bastante satisfatórios.

O primeiro indicador de que a solução é satisfatória, foi que todos participantes conseguiram utilizar o sistema, realizando as tarefas propostas para cada perfil.

Os usuários especialistas consideraram baixo o custo de criação de uma consulta. Afirmaram que a modelagem pronta facilita a criação das consultas, mas não vêem problema em criá-las sem que elas sejam previamente modeladas. Eles consideraram que o trabalho não aumentou muito

[7]Tamanduá 1.0 está sendo considerado a versão existente do sistema de segunda geração, sem a aplicação do modelo. Tamanduá 2.0 consiste na versão do protótipo desenvolvida para este trabalho.

em relação ao que fazem para criar diretamente as tarefas de mineração, sendo a maior diferença em relação ao fornecimento das explicações sobre as decisões durante o processo. Entretanto, acreditam que para os usuários leigos as abstrações e explicações criadas são muito interessantes.

Para os usuários leigos, a possibilidade de interagir com as abstrações criadas é vista como uma grande oportunidade de ampliar o uso de sistemas de mineração de dados. Eles não tiveram dificuldades em executar as consultas e como não possuíam conhecimento direto em relação aos conceitos técnicos de mineração de dados, foi possível perceber que esse conhecimento não é necessário para interagir com as abstrações criadas, que consiste no que é proposto pelo modelo.

Ainda em relação ao modelo, alguns outros aspectos apontaram como fatores positivos do mesmo:

- Os usuários leigos sentiram necessidade de mais informações sobre as consultas, o que é previsto pelo modelo pela base de conhecimento e destaca a importância da mesma. Embora os especialistas tenham oferecido explicações, os usuários leigos consideraram que estavam amplas e poderiam ser mais específicas;

- Em relação aos termos utilizados, os usuários leigos também solicitaram mais "traduções", o que confirma a utilidade do dicionário proposto no modelo. Por exemplo, embora houvesse o recurso no protótipo para "tradução" dos nomes dos atributos, algumas não foram feitas, como "DiaSem" (Dia da Semana), o que gerou insatisfação dos usuários leigos.

Apesar dos indicadores positivos relacionados ao modelo, foram levantados diversos problemas e sugestões, tanto em relação às modelagens feitas quanto ao protótipo. Em relação às modelagens, os usuários leigos levantaram pontos que poderiam ser melhorados. Como citado, o processo de mineração consiste um processo interativo e iterativo que demanda que adaptações sejam feitas até que se consiga uma boa modelagem na concepção dos leigos. No caso, os especialistas não interagiram com os leigos após a criação da consulta, o que seria ideal em um contexto real. Além disso, os usuários especialistas, embora trabalhem na equipe do Tamanduá, não estavam no momento trabalhando com os projetos utilizados na avaliação. Isso pode tê-los levado a se darem por satisfeitos com a modelagem assim que identificaram uma primeira geração de padrões de interesse, sem investigar a fundo se valeria a pena refinar melhor a modelagem ou não. De todo jeito, o fato dos usuários leigos terem sido capazes de sugerir melhorias para as consultas é um indicador de que eles entenderam as decisões dos usuários especialistas, sem que tenham tido que aprender os conceitos técnicos. Desta forma, este também é um indicador interessante para esta primeira avaliação feita com usuários.

Em relação ao protótipo, alguns dos problemas levantados são relacionados à qualidade de uso do sistema. Apesar de não ser o foco da avaliação, neste momento, alguns problemas foram percebidos durante a interação dos usuários. De forma a realizar uma análise mais detalhada, dois tipos de avaliações devem ser realizadas: a de comunicabilidade e de usabilidade. Comunicabilidade visando analisar a qualidade da comunicação, visto que o modelo é baseado em Engenharia Semiótica e usabilidade por avaliar efiência e eficácia do sistema. Usabilidade é um dos fatores a ser analisado e consiste na capacidade que um sistema interativo oferece a seu usuário, em um determinado contexto de operação, para a realização de tarefas de maneira eficaz, eficiente e agradável [ISO9241 (2008)]. Entretanto, não foi feita uma análise de usabilidade detalhada, tanto que não foram medidos de forma sistemática tempos gastos nas tarefas, por exemplo, para análise da eficiência. Alguns problemas foram identificados "naturalmente" e documentados para posterior melhoria no

protótipo. No Apêndice C.2, listamos alguns tipos de problemas classificados de acordo com as diretrizes de Nilsen [Nielsen (1994)].

Ainda em relação a qualidade de interação, conforme apresentado no capítulo 3, a Engenharia Semiótica entende a interação humano-computador como uma comunicação entre designer e usuário. É fundamental então uma forma de avaliar o sucesso ou não dessa comunicação. Para isso, foi proposto o método de avaliação de comunicabilidade [Prates et al. (2000), de Souza (2005)]. A comunicabilidade é definida como sendo a capacidade da interface de transmitir ao usuário de forma eficaz e eficiente as intenções e princípios de interação que guiaram o seu projeto. O objetivo do método é o de verificar como os usuários estão recebendo as mensagens do designer através da interface e identificar rupturas na comunicação que podem ocorrer durante a interação. Entretanto, esse tipo de avaliação não foi realizada nesse trabalho, visto que, como citado, o objetivo da avaliação era analisar a solução e não o protótipo. Neste primeiro momento, o objetivo foi analisar a experiência do usuário.

Em relação às sugestões, os usuários, tanto leigos quanto especialistas, sugeriram mudanças e novas funcionalidades para o protótipo. Os principais comentários foram em relação às formas de visualização dos resultados. Foram sugeridas novas formas de visualização, como tabelas e gráficos, além de melhorias na forma atual. Algumas das sugestões já estão documentadas para as novas versões do protótipo, como a possibilidade de criação de novos *templates* e edição do apresentado atualmente como *default*. Algumas das sugestões levantadas estão apresentadas no capítulo a seguir como trabalhos futuros.

De forma geral, como vimos, a avaliação feita com os usuários foi preliminar, mas foi bastante rica e relevante. Tivemos vários fatores positivos, como apresentados anteriormente. Além disso, durante as entrevistas, os usuários afirmaram que a solução é útil e interessante. Os especialistas afirmaram que a relação *custo/benefício* para criação das abstrações é bastante positiva e os usuários leigos demonstraram bastante interesse em utilizar a solução em seus contextos de trabalhos.

Assim, foi possível obter indicadores de que a resposta para a pergunta, objetivo da avaliação, foi positiva, sendo então possível que especialistas criem abstrações (consultas) úteis para os usuários finais, onde esses consigam interagir com o sistema de mineração de segunda geração sem que saibam os conceitos técnicos envolvidos. Sendo assim, acredita-se que a solução proposta é relevante e útil aos usuários.

O próximo capítulo apresenta as conclusões deste trabalho, além de listar algumas direções futuras relevantes que podem ser desenvolvidas.

Capítulo 6

Conclusões

Mineração de dados é uma área que vem crescendo muito nos últimos anos, isso pela capacidade de lidar com grandes volumes de dados, um desafio crescente na busca por informações [Goldschmidt (2005)]. O armazenamento de dados se tornou uma prática constante em diversas áreas e a análise desses dados é algo inviável de ser realizada manualmente. A mineração em si consiste em uma fase do processo de descoberta de conhecimento (KDD), a de exploração dos dados, mas usualmente o termo é usado para referenciar todo o processo.

Os sistemas de mineração de dados podem ser divididos em quatro gerações, onde cada uma apresenta suas características específicas, descritas no capítulo 1. Nosso foco foi em relação aos sistemas de segunda geração, que engloba as chamadas "*suites*" e abrangem mais de uma etapa do processo de mineração de dados. Essa escolha foi pela ampla aplicação e utilização desse tipo de sistema, visto que não são direcionados a nenhum problema específico.

A mineração de dados apresenta diversas técnicas, que podem ser aplicadas a problemas distintos. Dentre as diversas técnicas existentes, a escolhida para ser explorada foi a mineração de regras de associação, apresentada no capítulo 3. Trata-se de uma das técnicas mais populares, que tem aplicações nas mais diversas áreas como marketing, economia e ciências sociais [Hipp et al. (2000)]. Ela tem como característica geral encontrar correlações interessantes, que não são facilmente descobertas, em bases de dados.

Entretanto, sistemas de mineração de dados de segunda geração de regras de associação apresentam diversos desafios de interação. No capítulo 2, foram apresentados esses desafios, como escolha dos atributos, escolha e manipulação de parâmetros, análise dos resultados e entendimentos de medidas de interesse. Todas essas dificuldades serviram como motivação para a solução apresentada nesse trabalho. Isso porque foi detectado que os usuários capazes de interagir com esses tipos de sistemas tinham que possuir um grande conhecimento técnico e específico de mineração de dados.

A partir dessa motivação, foi proposto neste trabalho um modelo de interface extensível para sistemas de mineração de dados por regras de associação. Ele tem por objetivo ampliar o acesso a sistemas de mineração de dados de segunda geração e, logo, ao conhecimento que esse são capazes de gerar. Para isso, ele propõe a criação de uma camada de abstração onde formas de interação possam ser executadas diretamente, sem que sejam necessários os conhecimentos técnicos citados. Com isso, o grupo de usuários desse tipo de sistema pode ser dividido em duas "categorias", leigos e especialistas. Os especialistas criam as abstrações com que os leigos irão interagir. Já os leigos, podem obter o conhecimento que necessitam sem terem os conhecimentos técnicos.

O modelo proposto aqui foi baseado na teoria da Engenharia Semiótica (apresentada no ca-

pítulo 3), que entende a interação humano-computador como uma comunicação entre designer e usuário. Assim, nossa solução permite que usuários da camada de abstração se tornem co-autores especialistas do sistema.

A estrutura do modelo é composta por três componentes: o Gerador, a LAIU (Linguagem Abstrata de Interface com o Usuário) e a Base de Conhecimento. O Gerador é utilizado pelos especialistas para criar a LAIU. A LAIU é a interface com a qual os usuários leigos irão interagir. Ela é decomposta em dois níveis, um que engloba aspectos léxicos e sintáticos (compõe a interface) e outro que aborda aspectos semânticos (chamado Interpretador). O Interpretador faz a comunicação entre o Gerador, a aplicação de segunda geração e a camada de abstração. A Base de Conhecimento, dentro do contexto da Engenharia Semiótica que considera a interação como um ato de comunicação, é um componente importante para que os especialistas, no caso os principais autores da camada de abstração, possam documentar suas decisões. Um subcomponente da base de conhecimento é o dicionário, que busca realizar a "tradução" dos termos relacionados à mineração na camada de abstração. A estrutura e demais características do modelo são apresentadas de forma mais detalhada no capítulo 4.

Para analisar a utilidade e viabilidade de implementação do modelo, foram realizadas duas grandes etapas de avaliações. A primeira parte buscava verificar a utilidade do modelo, já a segunda foi relacionada a viabilidade da implementação do mesmo, acoplado a um sistema de segunda geração.

Na primeira parte, foram criadas abstrações para um determinado contexto e utilizados cenários para verificar se o modelo poderia ser realmente utilizado. Essa etapa contou com a participação de usuários, alunos de graduação, que "simulavam" consultas, de forma a verificar se a criação das mesmas seria possível.

A segunda parte da avaliação consistiu em testar a viabilidade de implantação do modelo. Para isso, uma instância do modelo foi implementada acoplada a um sistema de segunda geração existente denominado Tamanduá, descrito no capítulo 5. Com isso foi criada uma primeira versão de um protótipo, denominada Tamanduá 2.0. Essa versão conseguiu viabilizar os conceitos descritos no modelo, permitindo a criação do que havia sido proposto. A versão do protótipo foi avaliada, inclusive com a participação de usuários reais, tanto especialistas quanto leigos.

Como resultado de todas as avaliações, foi possível perceber que o modelo é viável e útil para os usuários de diversos contextos distintos. Dessa forma, é possível que especialistas criem abstrações (consultas) úteis para os usuários finais, onde esses consigam interagir com o sistema de mineração de segunda geração sem que saibam os conceitos técnicos envolvidos.

O trabalho desenvolvido trouxe contribuições em diferentes áreas. A seguir são apresentadas as contribuições levantadas e a seguir são apresentados trabalhos futuros, mostrando que trata-se de uma pesquisa que ainda pode ser ampliada em diversos aspectos.

6.1 Contribuições

Este trabalho apresenta contribuições em áreas distintas. A principal delas relaciona-se diretamente com o objetivo do trabalho em propor o modelo de extensão para sistemas de mineração em regras de associação. A criação e avaliação do modelo levanta possibilidades de se ampliar a aplicação de técnicas de mineração em diversos contextos distintos e, o mais importante, para os usuários que não possuem conhecimento técnico necessário. Assim, um benefício seria a maior divulgação e aplicação das técnicas de mineração de dados. É importante ressaltar que, teoricamente, o modelo

é aplicado a qualquer ambiente de segunda geração. Além disso, a versão implementada inicialmente foi para regras de associação, mas teoricamente ele pode ser utilizado com outras técnicas. Assim, o modelo não é restrito nem pelo sistema nem pela técnica utilizada, o que é um grande benefício para sua utilização.

A expansão da aplicação é uma importante contribuição, isso porque a facilidade de interação em sistemas é algo relevante e extremamente importante em qualquer contexto. Em sistemas de mineração de dados, essa importância é ainda maior devido a complexidade das técnicas. Por não ser algo trivial e de fácil entendimento, muitas vezes esses sistemas ficam restritos a um grupo de usuários especialistas. Para ilustrar a importância desse aspecto, é importante citar que a usabilidade de sistemas de Mineração de Dados(MD) recentemente foi apontada em [Kriegel et al. (2007)] como um dos cinco grandes desafios da área. Entretanto, diversos trabalhos que já foram propostos relacionados a esse aspecto continuam focando nos usuários especialistas, buscando facilitar a interação e interpretação dos resultados encontrados para os mesmos. Neste trabalho, apresentamos uma proposta original no sentido de ampliar o uso desse tipo de sistemas a usuários leigos. Para isso, a teoria da Engenharia Semiótica foi utilizada como base, buscando auxiliar o entendimento dos mesmos em relação à aspectos da interação.

Em relação a Engenharia Semiótica, teoria em que o modelo foi baseado, ela argumenta que se o usuário final entende a visão do projetista, existem mais chances de ele conseguir utilizar o sistema com maior eficiência. O modelo propõe a possibilidade de que extensões permitam que sejam feitas "adaptações" do sistema ao conhecimento do usuário. Para isso, também faz uso de explicações, utilizando componentes como a base de conhecimento para melhorar a comunicação entre o usuário especialista e o leigo. Ao implementar o modelo utilizando o Tamanduá, sistema de segunda geração, foi possível avaliar e obter indicadores positivos sobre a solução e, conseqüentemente, sobre a teoria. Dado que a teoria ainda é recente, coletar dados sobre o apoio que ela é capaz de oferecer na geração de soluções práticas é importante para uma avaliação mais ampla da mesma. Entretanto, outras avaliações e análises devem ser feitas, especialmente observando o uso real do sistema em um determinado contexto, analisando como os usuários farão uso dos recursos de extensão e explicações fornecidas.

O trabalho aqui apresentado fornece um exemplo de como a Engenharia Semiótica pode apoiar a criação de modelos (e sistemas) que possam melhorar a qualidade da comunicação. No nosso caso, as melhorias da comunicação consistem em permitir que usuários especialistas apresentem uma interface que ofereça chance de usuários leigos explorarem soluções de mineração de dados para seus problemas.

Outro aspecto relevante do trabalho é o uso da taxonomia baseada na Engenharia Semiótica para caracterização de extensões. A importância na utilização da taxonomia foi permitir analisar a capacidade e poder das extensões, inclusive com visões de diferentes usuários. Assim, foi possível ilustrar como as categorias podem descrever a extensão em termos do sistema computacional e dos usuários. Vale ressaltar que a taxonomia permitiu analisar e distinguir a extensão sob o ponto de vista do usuário especialista (que produz a extensão) e do usuário leigo (que a utiliza) e também do sistema de mineração em si.

Vários autores argumentam que a extensão é fundamental para que o sistema seja realmente de alta usabilidade [Nardi (1993)][Fischer et al. (2004)][de Souza e Barbosa (2006)]. O trabalho oferece indicadores nessa direção já que permite a adequação de sistemas genéricos para situações e contextos específicos e aumenta sua usabilidade permitindo o uso por usuários leigos. Como já citado, as avaliações mostraram indicadores iniciais de que a solução foi bem recebida tanto por

usuários especialistas quanto leigos.

Existem também contribuições práticas em relação ao projeto Tamanduá, visto que uma nova versão foi desenvolvida, não perdendo as funcionalidades já existentes e englobando diversos aspectos importantes que não existiam no sistema anterior. Com isso, é possível utilizar a nova versão em diferentes contextos e com usuários que até então não eram inseridos no processo de mineração. Dado o contexto de uso atual do Tamanduá, a possibilidade de uso por parte de usuário leigos é um grande benefício, visto que existe uma grande dificuldade de interação atualmente. Para a versão criada, novas aplicações e potenciais parceiros e clientes já estão sendo levantados, visto que o público alvo pode ser ampliado.

Contextualizando as contribuições deste trabalho em relação à computação em geral, é interessante citar os grandes desafios da computação identificados pela comunidade no encontro promovido pela Sociedade Brasileira de Computação (SBC), que apontam para problemas centrais da computação [SBC (2006)]. Os problemas se caracterizam por serem visionários e ambiciosos, mas realistas. Os cinco grandes problemas identificados são decompostos em problemas incrementais que permitem a sua avaliação e eventuais mudanças no curso durante sua realização.

Este trabalho encontra-se na interseção de dois destes desafios, nominalmente, o desafio 1 de gestão da informação em grandes volumes de dados e o desafio 4 de acesso participativo e universal do cidadão brasileiro ao conhecimento [Albergaria et al. (2008b)].

O desafio 1, de gestão da informação em grandes volumes de dados, tem por objetivo "desenvolver soluções para o tratamento, a recuperação e a disseminação de informação relevante, de natureza tanto narrativa quanto descritiva, a partir de volumes exponencialmente crescentes de dados multimídia" [SBC (2006)]. O desafio 4 sobre o acesso participativo e universal do cidadão brasileiro ao conhecimento, por sua vez, apresenta como objetivo a concepção de ambientes, métodos, modelos e teorias que sejam capazes de lidar com e vencer as barreiras tecnológicas, educacionais, culturais, sociais e econômicas que atualmente dificultam ou impedem o acesso do cidadão brasileiro ao conhecimento. Para isso, busca-se soluções que envolvam o cidadão, que passaria de usuário passivo a ativo e participativo na geração do conhecimento [SBC (2006), pp 17].

Como citado, este trabalho envolve os dois desafios (1 e 4), dentre os grandes desafios identificados pela comunidade no encontro promovido pela SBC [SBC (2006)]. Em relação ao desafio de gestão de grandes volumes de dados, o Tamanduá apresenta uma arquitetura distribuída capaz de gerar novos conhecimentos a partir da recuperação e processamento de grandes volumes de dados. No entanto, sem a aplicação do modelo, o acesso a esses conhecimentos gerados fica restrito aos detentores do conhecimento técnico envolvido nos algoritmos de mineração de dados. Assim, o modelo apresentado soluciona algumas das barreiras existentes (i.e. conhecimento técnico) para o acesso à informação, ampliando esse acesso. Considerando que sistemas de mineração de dados podem ser utilizados para gerar conhecimento sobre o governo e instituições públicas (como no caso da auditoria), a possibilidade de um acesso mais amplo a este tipo de conhecimento beneficia a sociedade como um todo.

Por fim, vale ressaltar também que o trabalho traz uma contribuição relevante ao ilustrar como problemas abordados por desafios distintos podem ter uma única solução.

O trabalho aqui desenvolvido ainda pode ser ampliado e estendido a diversas áreas. A seguir apresentamos alguns trabalhos futuros que já foram levantados, apresentando a continuidade da pesquisa desenvolvida.

6.2 Trabalhos futuros

Este trabalho envolve diferentes disciplinas e aplicações, teorias distintas como Engenharia Semiótica e Mineração de Dados. Existem assim diferentes frentes de novos possíveis trabalhos. A seguir listamos esses trabalhos de acordo com os temas relacionados a eles.

6.2.1 Modelo

Em relação ao modelo propriamente, analisar a possibilidade de sua aplicação em outras técnicas de mineração, como análise de agrupamentos e classificação, é uma questão interessante e desafiadora. Nesse sentido, existem várias frentes de trabalhos, visto que são necessários estudos em relação às outras técnicas, análise da viabilidade de criação de abstrações em contextos distintos e aplicação e implementação do modelo para elas. Atualmente, já existe um projeto em andamento que consiste em implementar a técnica de mineração de exceções na versão do protótipo desenvolvida. Ainda não existem resultados preliminares, mas consiste em uma primeira possibilidade, dentre inúmeras de se avaliar a aplicação do modelo.

Neste trabalho, focamos a aplicação do modelo em uma técnica específica, a mineração por regras de associação. Seria interessante investigar um modelo que permitisse a combinação de várias técnicas através de um *workflow*. Inicialmente ele seria voltado para o usuário especialista, mas poderia ser analisada a possibilidade de, combinado ao EDeM, possibilitar o acesso dos leigos.

6.2.2 Engenharia Semiótica

Dentro do contexto da Engenharia Semiótica, devem ser feitas novas avaliações e análises observando o protótipo em uso de forma a analisar maneiras de melhorar a comunicação designer/usuário, tanto a que é feita de forma indireta quanto a direta (utilizando base de conhecimento, por exemplo).

Outra linha a ser investigada é em relação ao tipo de extensão que pode ser realizada. Atualmente ela é limitada e só permite que o usuário especialista crie extensões. Entretanto, outras formas que permitam mais ações poderiam ser úteis, ampliando a dimensão semântica do sistema computacional e a intenção do usuário especialista.

Para o usuário leigo, atualmente não é permitido que ele faça extensões, mas seria interessante permitir que ele pudesse criar novas consultas a partir das existentes, criando consultas similares ou variando atributos relacionados, por exemplo. Assim, o usuário leigo poderia se tornar coautor do especialista na geração de novas consultas. Para ilustrar essa funcionalidade, poderiam ser criadas formas do leigo variar o atributo que deseja para uma mesma consulta. Por exemplo, se ele está acessando uma consulta que questiona sobre tempo, ele pode escolher entre atributos correlacionados como dia, hora, dia da semana, mês ou ano para execuções distintas.

6.2.3 Protótipo

Outro passo na continuidade da nossa pesquisa consiste em aprimorar o protótipo desenvolvido. A seguir são apresentados trabalho de duas "naturezas" distintas, uma em relação aos protótipos, possíveis instanciações do modelo de forma geral (seção 6.2.3.1) e outra especificamente em relação a versão do protótipo desenvolvida (seções 6.2.3.2 e 6.2.3.3).

6.2.3.1 Sugestões gerais

A seguir são apresentadas sugestões para protótipos em geral que podem ser implementados pelo modelo.

- **Utilizar a linguagem PMML**

  PMML (*Predictive Model Markup Language*) é uma linguagem baseada em XML (*eXtensible Markup Language*), que fornece um modo rápido e fácil para definir e compartilhar modelos de mineração [Group (2004)]. Regras de associação podem ser representadas no formato PMML e, com isso, é possível que sejam exportadas, importadas e utilizadas por diferentes aplicações. Isso torna o protótipo mais padronizado e flexível.

  A sugestão aqui consiste em utilizar PMML em protótipos, visando torná-los ainda mais flexíveis e os dados mais portáteis, universais e extensíveis. Isso porque utilizando essa padronização é possível, por exemplo, que a mineração executada em um determinado sistema possa ser visualizada em qualquer outro também padronizado. O primeiro passo nessa direção seria utilizar a PMML no próprio Tamanduá 2.0.

- **Analisar a possibilidade de criação de uma linguagem geral e ampla (*script*) para configuração geral de consultas e/ou visualizações**

  Na seção 3.2 foram apresentadas duas análises apresentadas por Bonnie [Nardi (1993)] e Fischer [Fischer (2007)] em relação às linguagens de programação. Em relação a elas, de uma forma geral, quanto maior a complexidade das linguagens, maior o escopo de aplicação, mas aumenta também o custo de aprendizagem das mesmas. Buscando ampliar as possibilidades de configuração e aplicação, uma linguagem de abrangência maior poderia ser desenvolvida para a criação e configuração de consultas e visualizações.

  Assim, o especialista não precisaria ficar limitado pela interface gráfica e poderia configurar diversos parâmetros que não foram abordados até então. Entretanto, é importante destacar que isso torna o trabalho do especialista mais abrangente, mas mais trabalhoso, visto que ele necessitaria saber, além dos conceitos técnicos e informações sobre o contexto, também sobre a linguagem de *script* a ser utilizada.

6.2.3.2 Desenvolvimento e Melhorias para o protótipo existente

Inúmeras funcionalidades já foram levantadas para o protótipo, mas não entraram no escopo da primeira versão desenvolvida. A seguir, estão listadas algumas sugestões de funcionalidades e melhorias a serem implementadas nas futuras versões.

- **Analisar formas de trabalho cooperativo que podem ser aplicados ao Tamanduá**

  O protótipo abrange conceitos de compartilhamento de informações. Ao criar grupos de pessoas, deve ser possível que os participantes troquem informações, além de compartilharem abstrações e resultados. Atualmente a interação é apenas entre usuários especialistas e leigos, mas de forma unilateral. Assim, podem ser analisadas formas de existir um "*feedback*" dos usuários aos projetistas, possibilitando que eles possam fazer comentários em relação às

abstrações que foram criadas ou criem solicitações por novas abstrações. Além disso, poderia existir cooperação entre usuários especialistas, de forma que pudessem trocar informações em relação às abstrações criadas.

Assim, decisões sobre os tipos de solução a serem adotados em relação à cooperação entre os usuários devem ser baseadas em pesquisas e trabalhos da área, o que não foi abordado nessa dissertação.

- **Criar um *help* contextualizado e abrangente**

  A sugestão em relação ao *help* consiste em utilizar o modelo de sistema de ajuda proposto com base na Engenharia Semiótica [Silveira et al. (2004)] para projetar o sistema de ajuda do Tamanduá. Com isso, os usuários teriam acesso contextualizado e sob demanda ao conhecimento relativo aos conceitos técnicos envolvidos de forma que ajudasse o especialista durante a criação da abstração.

  Criando esse sistema de ajuda, pode ser possível coletar indicadores de como essas informações auxiliam o usuário no aprendizado ou retenção dos conceitos técnicos envolvidos. Nesse sentido, podem surgir diretrizes para ampliar a atuação como as ferramentas de quarta geração, que buscam ensinar e auxiliar os usuários no processo de mineração.

- **Funcionalidades na visualização textual dos resultados**

  Atualmente, a versão do protótipo apresenta a visualização textual dos dados, que pode ter ainda diferentes recursos para auxiliar os usuários no entendimento e seleção dos resultados obtidos.

  Assim, novas configurações em relação a saída dos resultados podem ser feitas de forma a torná-los mais claros para os usuários. Recursos como limitação das regras a serem apresentadas e escolha de parâmetro para servir de *ranking* na apresentação dos dados são exemplos de recursos. Além disso, funcionalidades como filtros podem ser interessantes para a seleção dos resultados.

- **Criação de recursos para seleção de regras**

  No protótipo atual, não foi implementado nenhum mecanismo que pudesse auxiliar os usuários na seleção das regras resultantes. Ou seja, não existem recursos que permitem a escolha por regras que sejam mais relevantes por um determinado aspecto. Nesse sentido, foram citados trabalhos na seção 2.5 que podem ser analisados e adequados dentro do contexto do protótipo.

  Além disso, formas de compartilhamento de comentários e decisões podem ser interessantes para o momento de pré-seleção das regras.

- **Implementar formas de visualização tabular e gráfica dos resultados**

  Os resultados obtidos, além de serem apresentados de forma textual, podem ser configurados em tabelas e gráficos, de forma a ajudar os usuários a compreenderem melhor as informações obtidas.

Essas visualizações, assim como a textual, podem ser disponibilizadas aos usuários leigos, sendo configuradas pelos especialistas. Esses devem definir quais valores aparecem nas tabelas e eixos dos gráficos, assim como a melhor forma de plotar os dados e melhor apresentá-los.

- **Implementar formas de visualização que permitam desenhos representativos para os problemas, como imagens, mapas e calendários**

Como citado, o protótipo passou por várias etapas de avaliação, onde uma delas consistia em verificar se outros usuários especialistas, que não participaram na criação do modelo, eram capazes de criar abstrações para diferentes domínios, onde os resultados foram apresentados na seção 4.3.2. Como já apresentado, essa parte da avaliação foi realizada como parte de um projeto de classe para o curso de Mineração de Dados, no nível da graduação. Nos trabalhos surgiram várias sugestões de visualizações como gráficas, textuais e tabulares. Uma forma de apresentação que também surgiu foi apresentando imagens. Para o cenário do vestibular, representações como a mostrada na figura 6.1 foram sugeridas para responder à pergunta: *Alunos do curso <NOME_CURSO> têm bom desempenho na disciplina <NOME_DISCIPLINA>?*. Em relação ao cenário de vitimização, um mapa foi proposto para regras que envolvam localidade, de forma a apresentar densidade de informações. A representação da figura 6.2 foi proposta para a consulta: *Quando e onde o <CRIME> do tipo <TIPO_CRIME> acontece?* Vale ressaltar que a primeira proposta consiste apenas uma representação gráfica, enquanto a segunda depende da disponibilidade de dados georeferenciados.

Essas formas propostas de visualização de imagens como parte das abstrações são extremamente interessantes, visto que auxiliam os usuários no entendimento dos resultados gerados. Entretanto, formas de configuração e geração dessas imagens é um trabalho complexo e formas de serem geradas devem ser desenvolvidas.

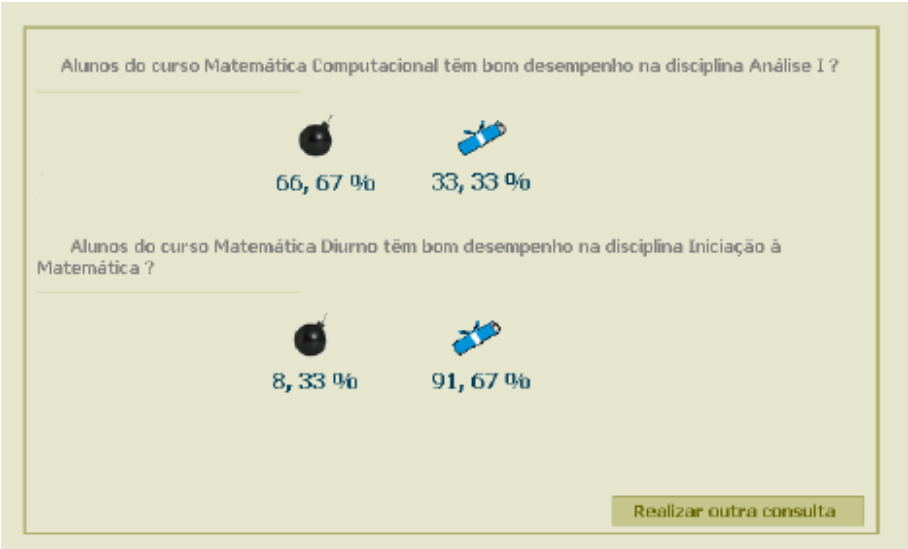

Figura 6.1: Sugestão de visualização dos trabalhos de MD utilizando imagens (contexto do vestibular)

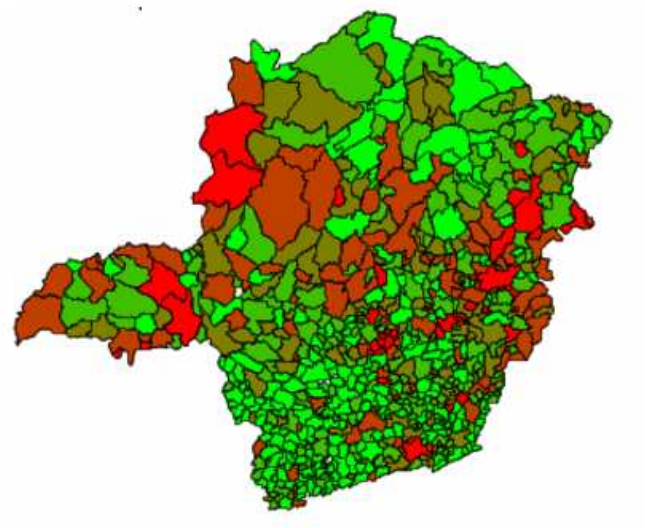

Figura 6.2: Sugestão de visualização dos trabalhos de MD utilizando imagens (contexto de criminalidade)

6.2.3.3 Avaliações do protótipo existente

A seguir são apresentados trabalhos que podem ser desenvolvidos envolvendo avaliações do protótipo.

- **Realizar novas avaliações**

  Como apresentado na seção 5.3, testes com usuários reais foram realizados. Entretanto, o foco da avaliação preliminar foi o modelo e a solução que ele propunha. Assim, buscou-se analisar a capacidade dos usuários em adotar a solução proposta e a visão dos usuários sobre ela.

  Como forma de analisar melhor a solução e protótipo, novos testes devem ser realizados, com objetivos e métodos diferentes. Inclusive testes que avaliem a interface devem ocorrer, pois ela tem impacto direto na forma como o modelo é comunicado ao usuário e o uso que ele fará deste.

  Nesse sentido, testes de usabilidade devem ser realizados, envolvendo usuários reais de contextos distintos. Além disso, todas as interações ocorridas no teste preliminar foram gravadas e um próximo passo previsto é analisar as interações de acordo com o Método de Avaliação de Comunicabilidade.

***

Assim, o trabalho aqui apresentado oferece a possibilidade de continuidade da pesquisa e desenvolvimento descritos, assim como serve de ponto inicial para novas pesquisas.

Apêndice A

Modelagem Tamanduá

O Schema XML (Linguagem Pheromone) ilustrado abaixo apresenta parte da definição realizada para o armazenamento de consultas. A criação da mesma é composta por fases (onde o usuário vai definido a consulta) e um corpo que consiste na consulta em si, em que ela pode apresentar texto, atributos, parâmetros e operadores em sua constituição.

```
<xsd:schema xmlns:xsd="http://www.w3.org/2001/XMLSchema"
version="1.0" xml:lang="EN">

  <xsd:annotation>
    <xsd:documentation xml:lang="en">
     Schema XML PHEROMONE
    </xsd:documentation>
  </xsd:annotation>

<xsd:element name="pheromone">
  <xsd:complexType>
    <xsd:all>
      <xsd:element name="body" minOccurs=0 maxOccurs=1
type="xsd:BODY"/>
      <xsd:element name="execution" minOccurs=0 maxOccurs=1
type="xsd:EXECUTION"/>
    </xsd:all>
  </xsd:complexType>
</xsd:element>

<xs:element name="CREATION_FASE">
<xsd:simpleType>
  <xsd:restriction base="xsd:string">
    <xsd:enumeration value="first"/>
    <xsd:enumeration value="second"/>
    <xsd:enumeration value="third"/>
    <xsd:enumeration value="fourth"/>
  </xsd:restriction>
</xsd:simpleType>
</xs:element>

<xsd:element name="BODY">
  <xsd:complexType>
    <xsd:all>
      <xsd:element name="text" type="xsd:string"/>
      <xsd:element name="attribute" type="xsd:ATTRIBUTE"/>
      <xsd:element name="param" type="xsd:PARAM"/>
      <xsd:element name="operator" minOccurs=0
type="xsd:OPERATOR_TYPE" use="optional"/>
    </xsd:all>
  </xsd:complexType>
</xsd:element>

<xsd:element name="ATTRIBUTE">
  <xsd:complexType>
    <xsd:all>
      <xsd:attribute name="id" type="xsd:string"/>
      <xsd:element name="name" type="xsd:string"/>
      <xsd:element name="value" minOccurs="1"
maxOccurs="unbounded" type="xsd:string"/>
    </xsd:all>
  </xsd:complexType>
</xsd:element>

. . .
```

Figura A.1: Schema XML - Pheromone

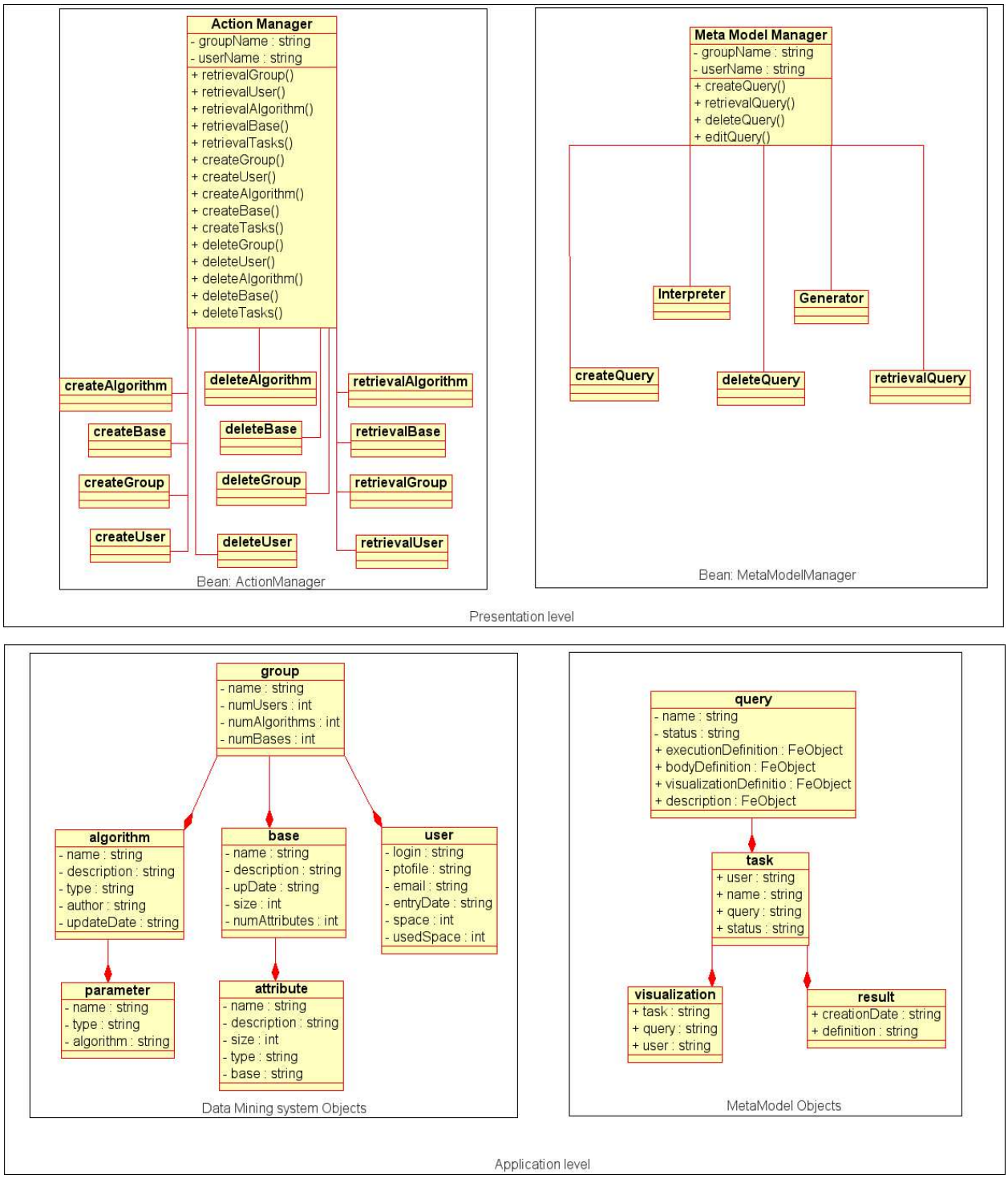

Figura A.2: Diagrama de classes - Tamanduá

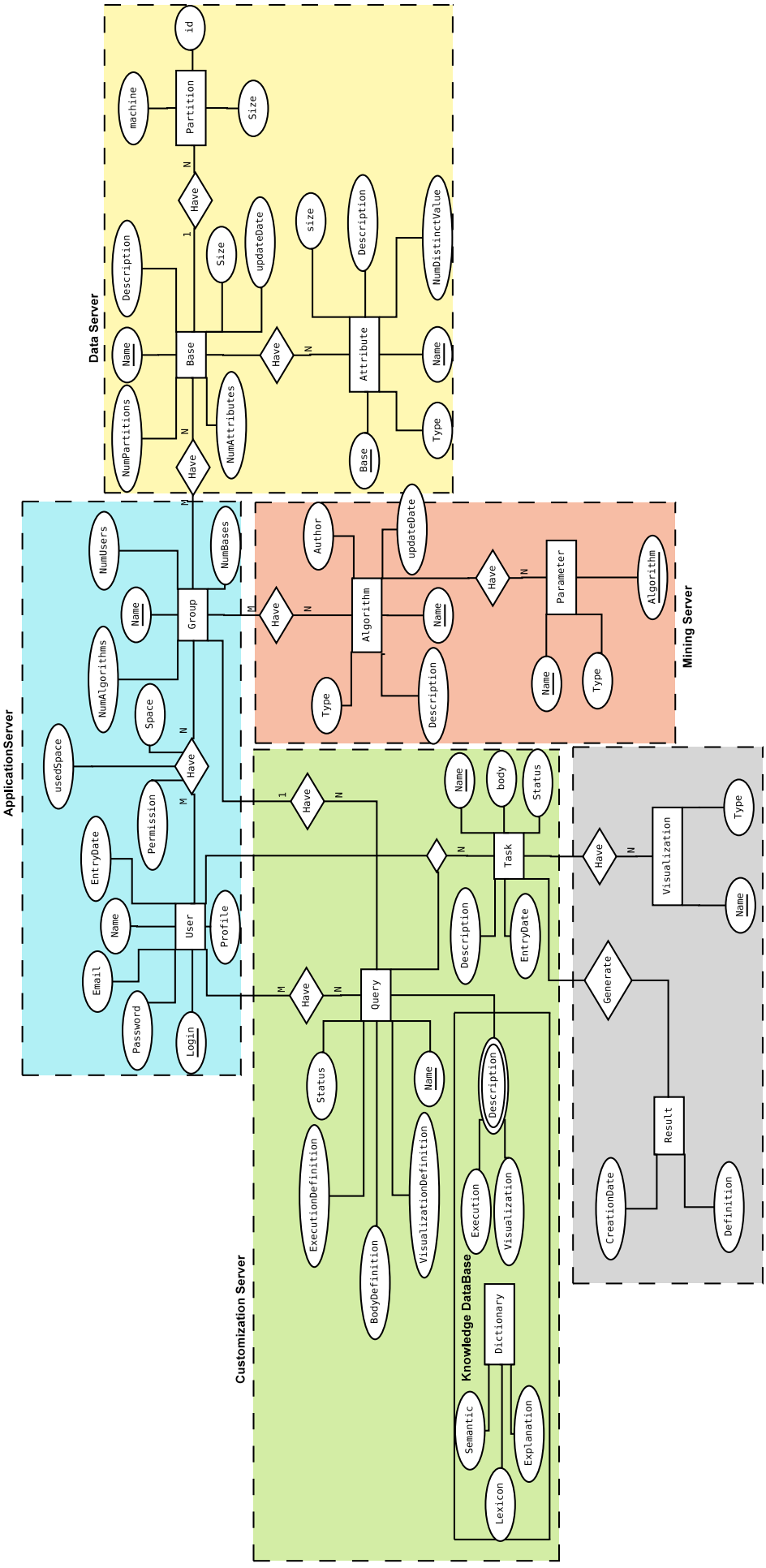

Figura A.3: Modelo de dados persistentes - Tamanduá

Apêndice B

Telas do Protótipo - Tamandua 2.0

Figura B.1: Tela de Bem Vindo

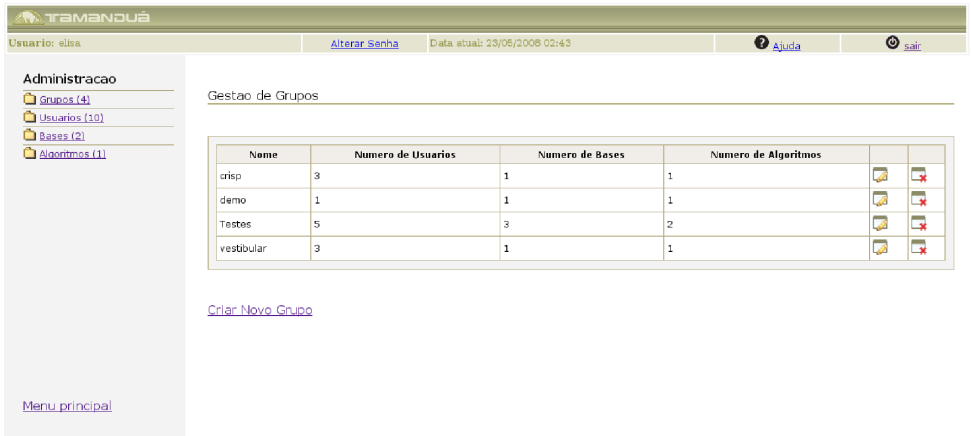

Figura B.2: Tela de Administração do Sistema

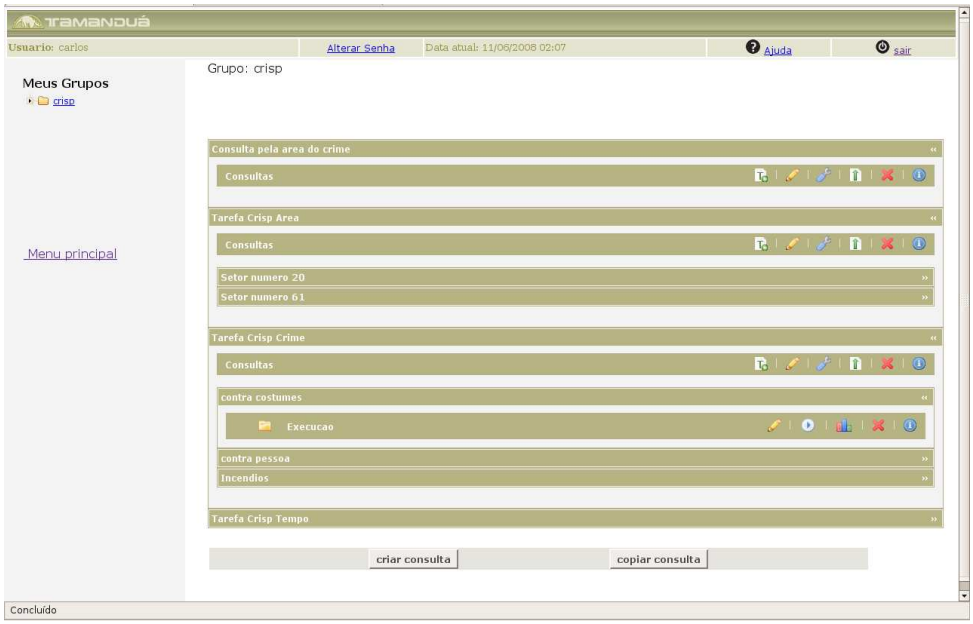

Figura B.3: Tela de lista de consultas - visão do Especialista

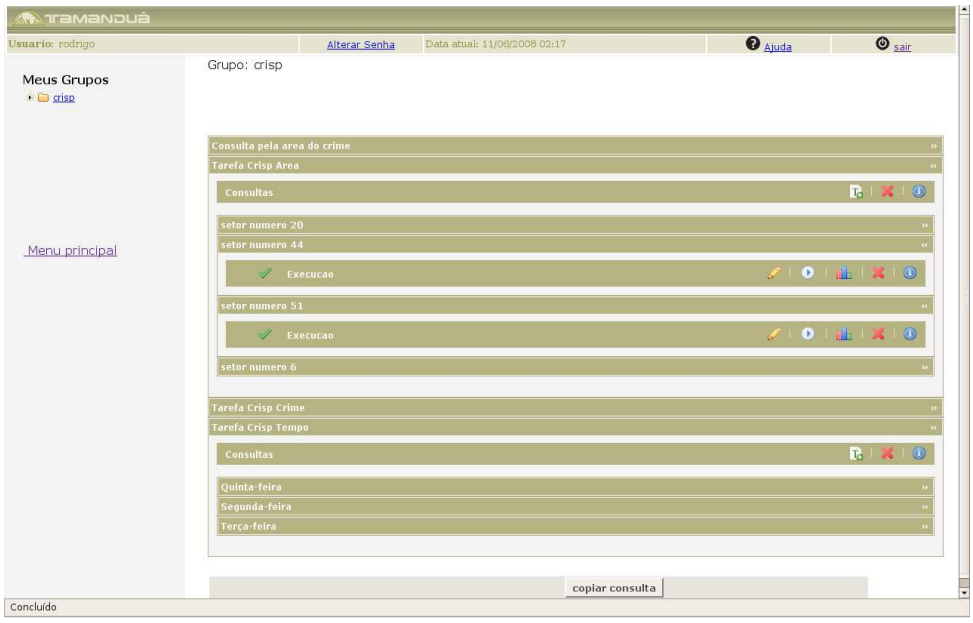

Figura B.4: Tela de lista de consultas - visão do Leigo

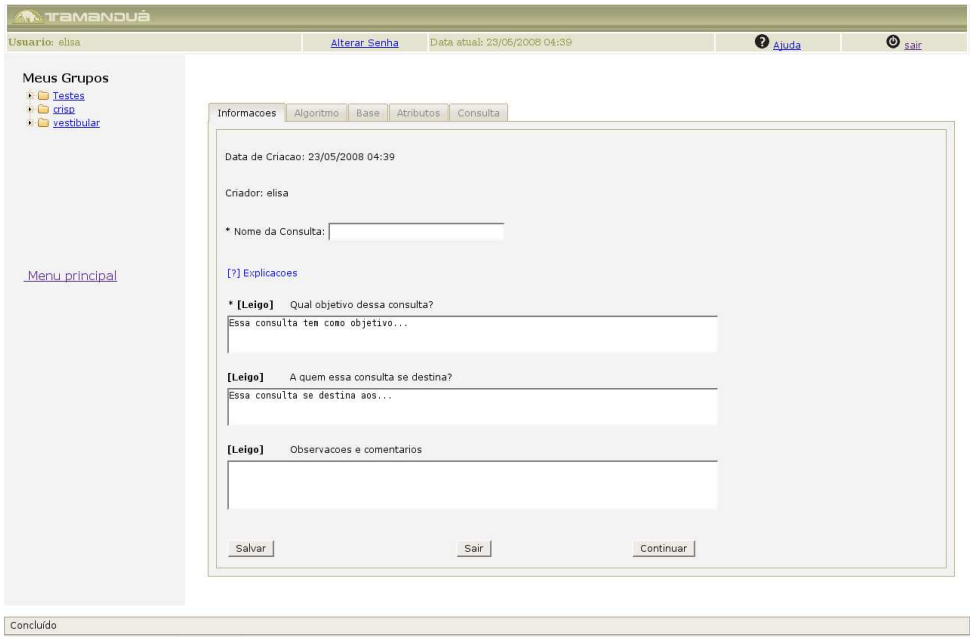

Figura B.5: Tela de Criação de Consulta (Informações)

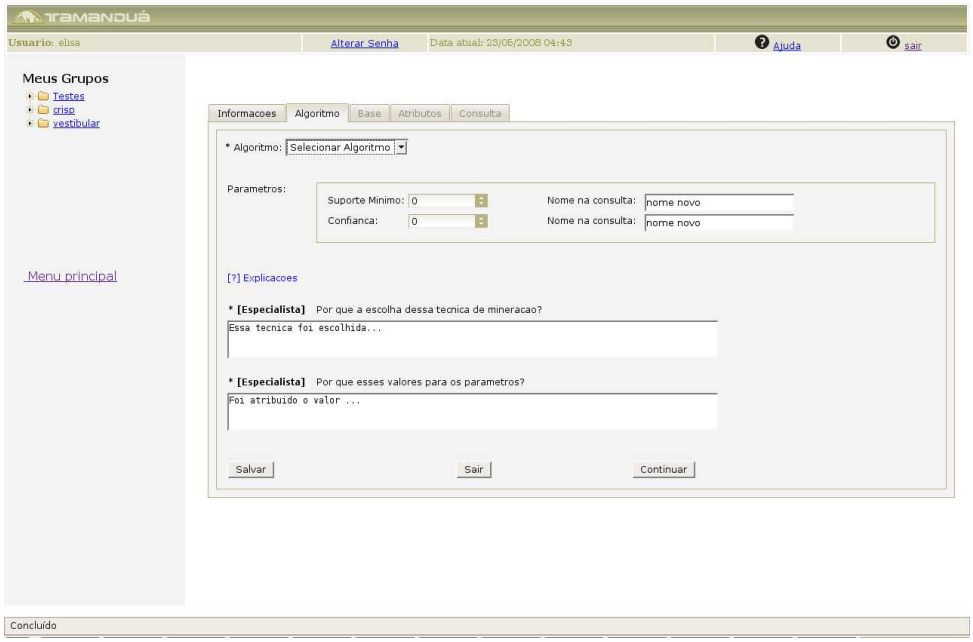

Figura B.6: Tela de Criação de Consulta (Algoritmo)

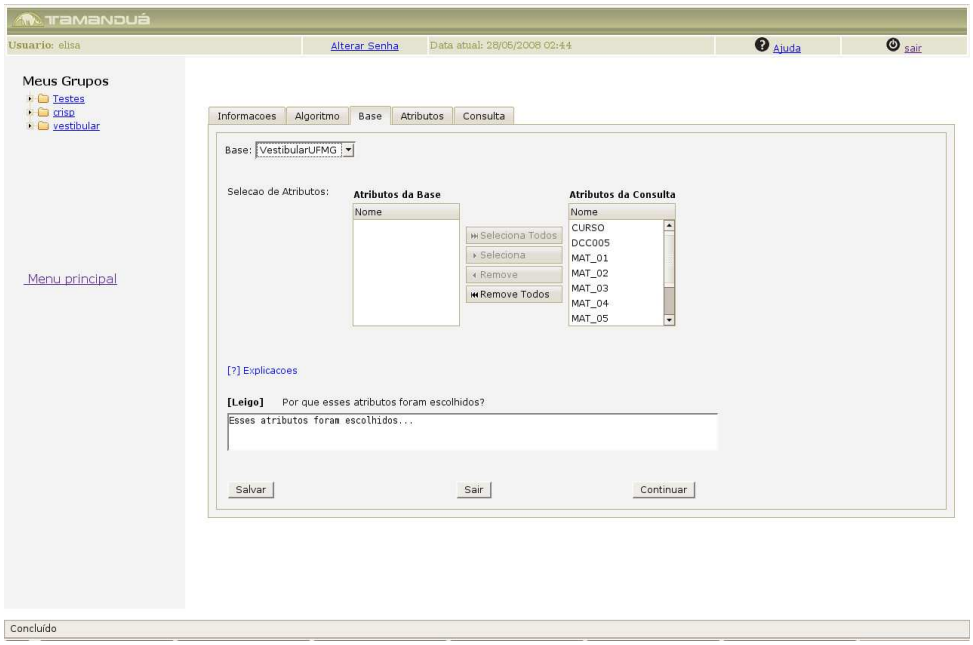

Figura B.7: Tela de Criação de Consulta (Base)

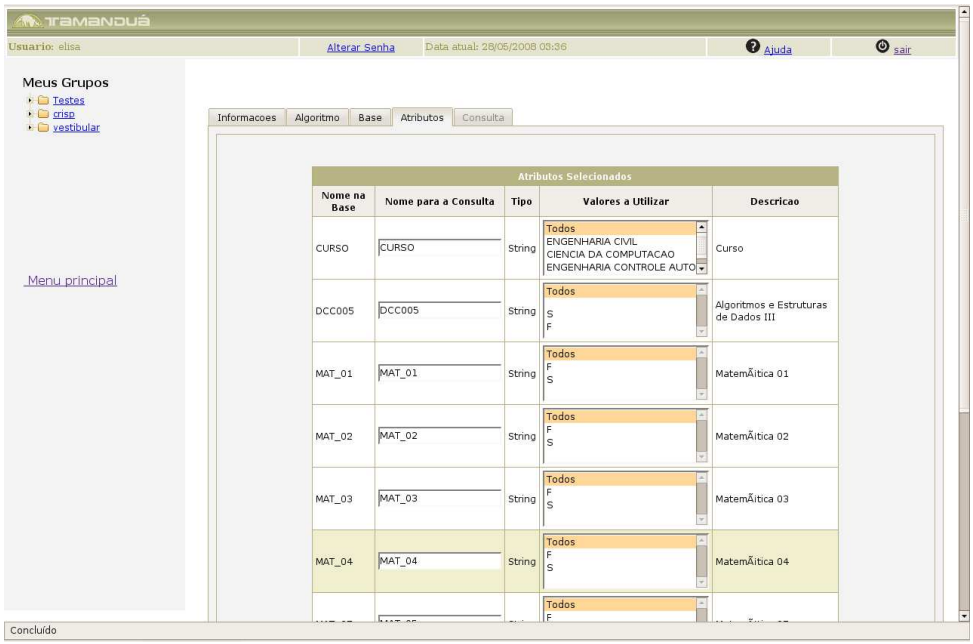

Figura B.8: Tela de Criação de Consulta (Atributos)

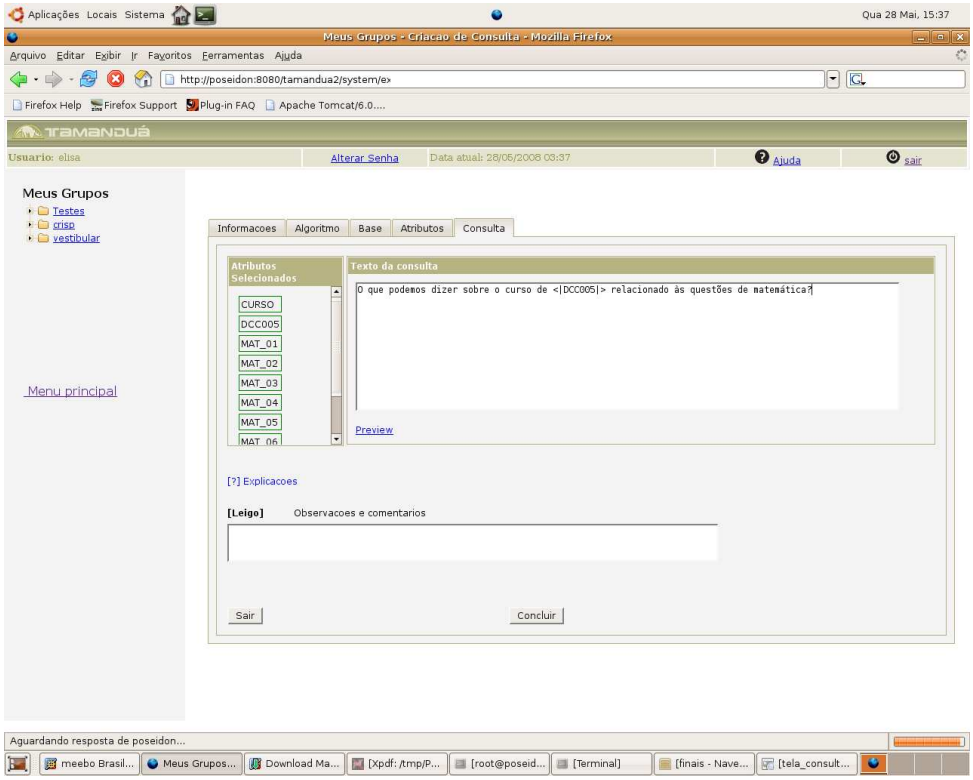

Figura B.9: Tela de Criação de Consulta (Consulta)

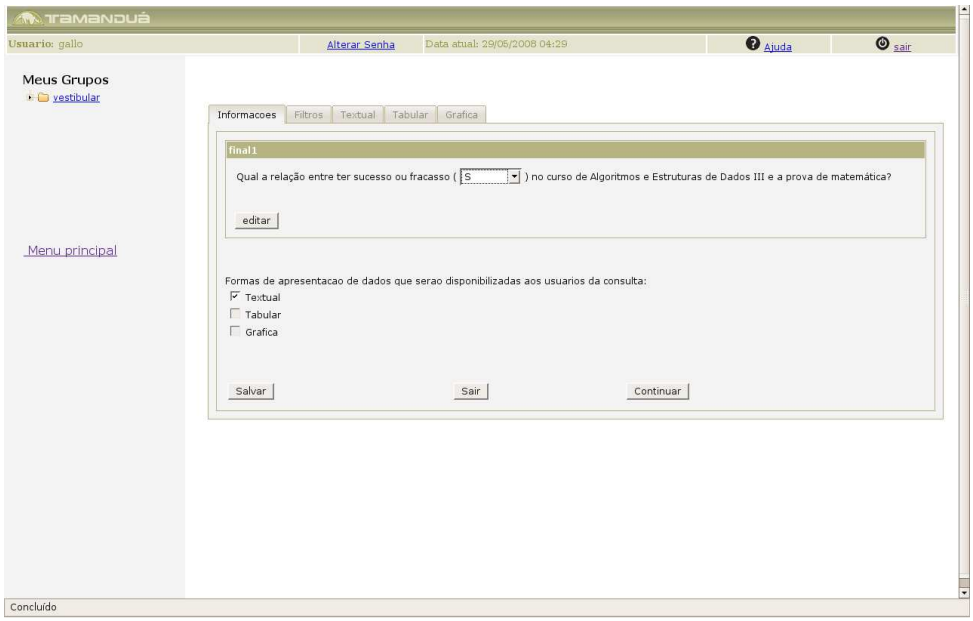

Figura B.10: Tela de Configuração da Saída da Consulta (Informações)

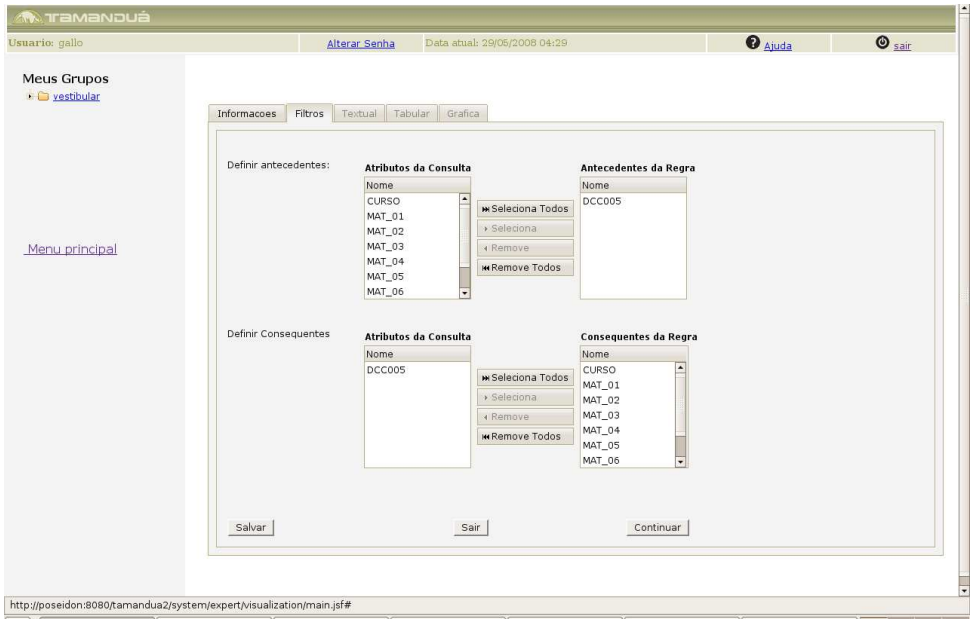

Figura B.11: Tela de Configuração da Saída da Consulta (Filtros)

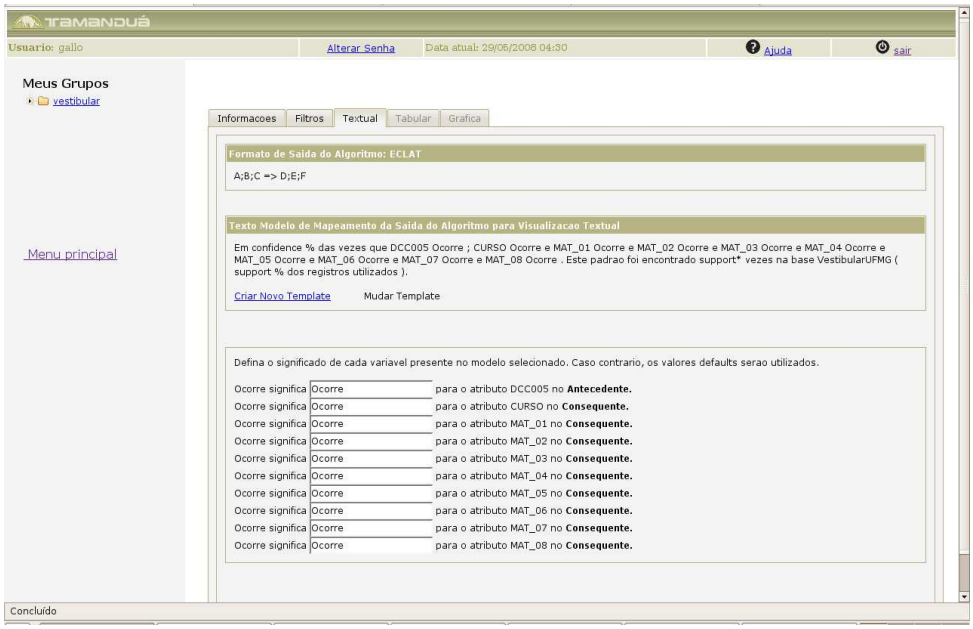

Figura B.12: Tela de Configuração da Saída da Consulta (Textual)

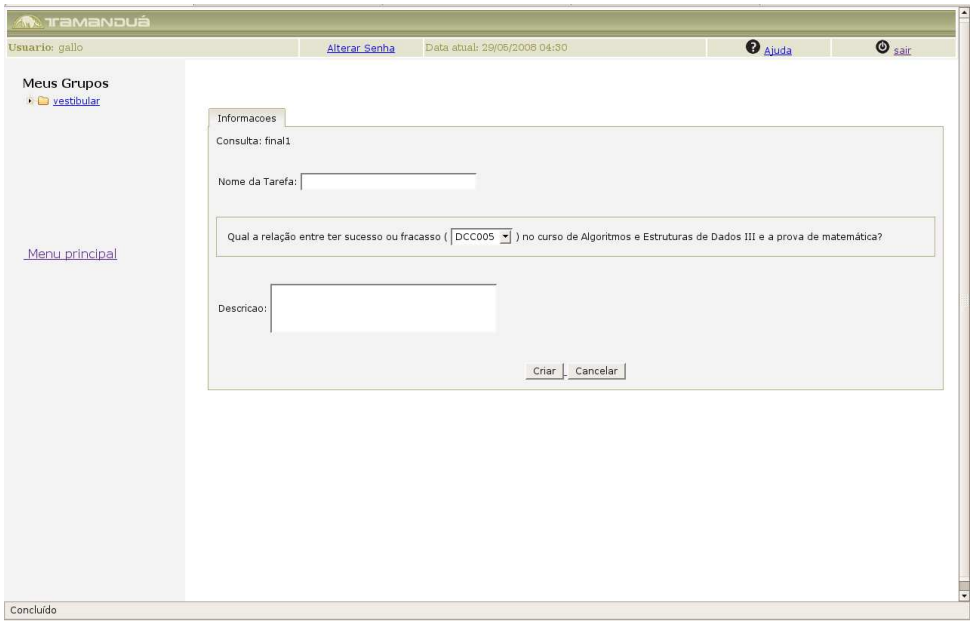

Figura B.13: Tela de Tarefa (Instância de uma Consulta)

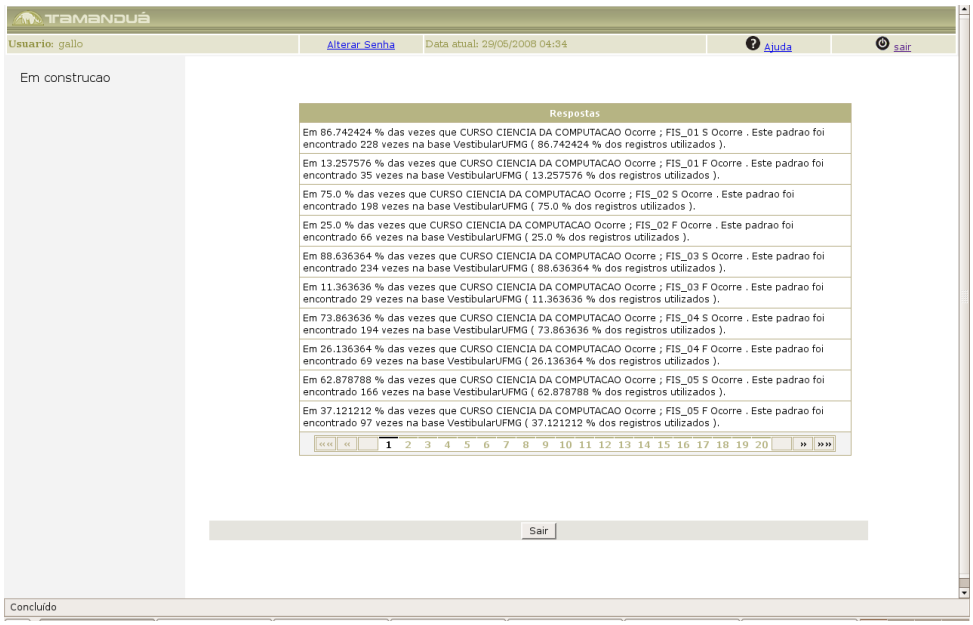

Figura B.14: Tela de Visualização de uma Consulta

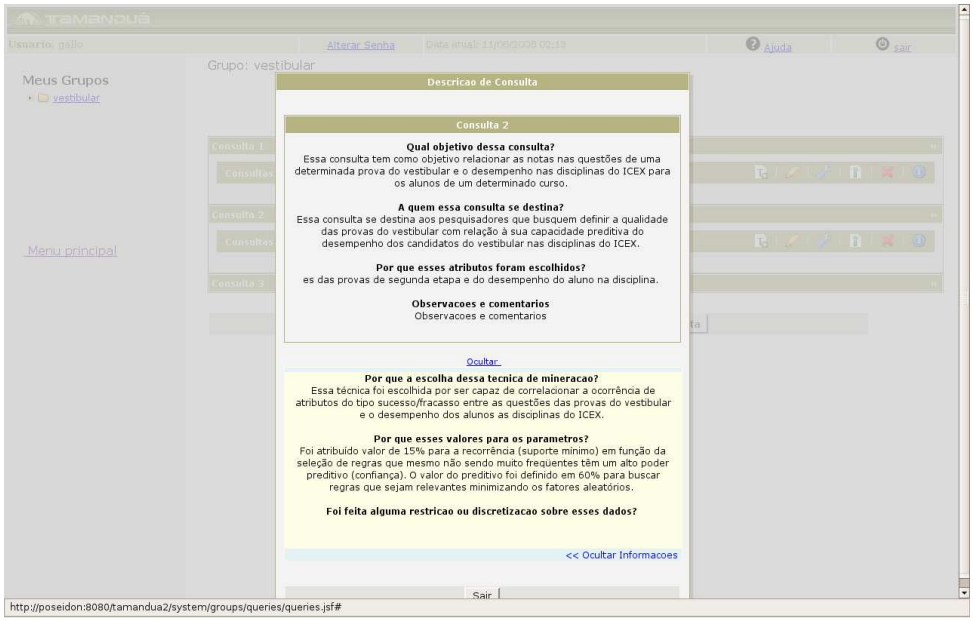

Figura B.15: Tela de Visualização das Explicações

Apêndice C

Avaliações

C.1 Avaliação com cenários

Título do Trabalho: Avaliação de modelo para aplicação em sistemas de mineração de segunda geração
Data: Janeiro/2008
Instituição: DCC / UFMG
Pesquisadores Responsáveis:
- Elisa Tuler de Albergaria (elisa@dcc.ufmg.br)
- Prof. Raquel O. Prates (rprates@dcc.ufmg.br)
- Prof. Wagner Meira Junior (meira@dcc.ufmg.br)

Introdução: Este Termo de Consentimento contém informações sobre o projeto de pesquisa indicado acima. Caso tenha alguma dúvida, não hesite em perguntar ao pesquisador responsável. Você também deverá ssinar o termo do qual receberá uma cópia.

Objetivo da Pesquisa: O objetivo desse trabalho consiste em avaliar o modelo proposto no trabalho de pesquisa de mestrado da aluna Elisa Tuler de Albergaria. O modelo visa aumentar a usabilidade de sistemas de mineração de dados por regras de associação de segunda geração.

Utilização dos projetos: Pede-se consentimento para utilização dos projetos desenvolvidos na disciplina de Mineração de Dados, ministrada e concluída no segundo semestre de 2007-2 no DCC/UFMG pelo Prof. Wagner Meira Junior. A análise a ser feita dos trabalhos apresentados não está relacionada com a disciplina concluída ou outras que os alunos possam vir a fazer. Além disso, a análise dos trabalhos não levará em consideração o desempenho dos alunos na disciplina ou nota obtida no trabalho. Os dados dos trabalhos serão utilizados para gerar cenários de uso para avaliação do modelo. Os cenários gerados podem ser utilizados em publicações técnicas e científicas sobre o modelo. Vale ressaltar a análise a ser feita e uso dos dados não serão empecilho para os autores do projeto de publicarem os resultados do mesmo, caso desejem.

Privacidade: Os alunos do curso e autores dos projetos não serão citados nominalmente, e informações que possam identificá-los publicamente não serão divulgadas.

Participação voluntária: Você é livre para decidir se quer autorizar o uso do seu trabalho de curso para fins desta pesquisa ou não. Sua decisão não afetará sua vida estudantil e nem qualquer relacionamento com os avaliadores, professores ou a Instituição por trás desta. O trabalho só será utilizado se todos os autores consentirem em sua utilização.

Compensação: A permissão para uso dos projetos é voluntária, não sendo oferecida nenhuma remuneração.

Consentimento Livre e Esclarecido (Acordo Voluntário)
Autorizo a utilização do meu projeto da disciplina de Mineração de Dados, ministrada no DCC/UFMG no segundo semestre de 2007, como parte da avaliação do modelo proposto, da pesquisa aqui descrita.

O documento mencionado acima descrevendo o termo de consentimento foi lido e explicado. Eu tive a oportunidade de fazer perguntas sobre o mesmo, que foram respondidas satisfatoriamente. Eu estou de acordo em autorizar o uso do projeto citado.

_____ Data	Assinatura do aluno: _____ Nome do aluno:
	Assinatura do pesquisador: _____ Nome da pesquisadora: Elisa Tuler de Albergaria

Figura C.1: Termo de consentimento para citação dos cenários dos alunos de mineração de dados

C.2 Avaliação com usuários

Roteiro para primeiro encontro com usuários especialistas selecionados:

- Cumprimentar os participantes
- Solicitar licença para gravar, para fins de documentação
- Explicar a pesquisa de forma resumida
  - apresentar o objetivo do modelo
  - falar do protótipo (e separação ainda existente entre os dois tamanduás)
  - falar do objetivo da avaliação
- Apresentar o contexto em que cada participante irá trabalhar
  - Perguntar que tipo de trabalho já fizeram nesse contexto e que tipo de interação tiveram com os usuários especialistas
- Explicar as etapas das avaliações que serão feitas
  1) Reunião com usuários leigos e especialistas
  2) Avaliação com usuários especialistas
  3) Avaliação com usuários leigos
- Perguntar se possuem dúvidas e/ou comentários
- Perguntar se estão interessados e irão participar
- Agradecer e ver disponibilidade para novo encontro

Figura C.2: Roteiros das etapas da avaliação com usuários - Reunião com usuários especialistas

Roteiro para reunião com usuários especialistas e leigos - VESTIBULAR

- Apresentar os usuários entre si e se apresentar
- Solicitar licença para gravar, para fins de documentação
- Explicar as etapas das avaliações que serão feitas
- Tamanduá atual
  - Falar rapidamente do tamanduá atual.
  - Perguntar a opinião do usuário especialista sobre o sistema
- Falar um pouco da pesquisa, o objetivo do modelo. Apresentar exemplos de consultas de outros contextos.
  - Perguntar sobre a experiência da disciplina, pontos positivos e negativos. O que foi obtido de interessante e o que faltou.
- Trabalho do usuário leigo em relação ao vestibular (relação com a Copeve):
  - O que deseja obter (pensar nesta base e na completa)
  - Qual objetivo
  - Quem seria o usuário final – ele mesmo ou outro (quem, que perfil)
  - Atualmente que avaliação das questões é feita. Como?
- Levar um material descritivo da base que será usada naquele contexto, quais atributos possui, que tipo de discretização foi feita, o que a base possui ou não.
  - O que você deseja saber em relação a essa base? É interessante que obtenha essa informação de tempos em tempos, ou variando algum tipo de dado?
- Solicitar que o especialista crie 3 consultas juntamente com o leigo.
- Perguntar se possuem dúvidas e/ou comentários.

Figura C.3: Roteiros das etapas da avaliação com usuários - Reunião com usuários especialistas e leigos (Vestibular)

Roteiro para reunião com usuários especialistas e leigos - CRISP

- Apresentar os usuários entre si e se apresentar
- Solicitar licença para gravar, para fins de documentação
- Explicar as etapas das avaliações que serão feitas
- Falar rapidamente do tamanduá atual, perguntando a opinião dos usuários quanto a ele (e sobre mineração?)
  - Todos conhecem o tamanduá
  - Usaram ou tiveram contato com o Tamanduá desde a última reunião
  - Ainda lembram, ou gostariam de explicação breve
  - Citar dificuldades levantadas por eles em última reunião sobre Tamanduá.
    - > *Dificuldade de identificar regras interessantes*
    - > *Dificuldade de adaptar forma de trabalho – criam teorias para serem comprovadas (ou não) pelos dados e não identificar "resultados interessantes"*
    - > *Conceitos nem sempre são claros (como suporte e confiança)*
  - Ver se tem mais algum comentário
- Falar um pouco da pesquisa, o objetivo do modelo. Apresentar exemplos de consultas de outros contextos.
  - Deixar claro que o usuário final da consulta seria a polícia ou bombeiro.
- Trabalho do CRISP sobre análise de dados para a polícia:
  - Freqüência
  - Sobre análise (tipos; periódica x única; proposta x solicitada, forma de apresentação)
  - Análise candidata a consulta – descrever melhor
  - Se não tiver já sendo feita – então passar para o próximo passo e tentar descobrir. (Pode usar exemplo dado por eles da outra vez para estimular discussão: O padrão de homicídios em BH tem mudado ou não? Por que?)
- Levar um material descritivo da base que será usada naquele contexto, quais atributos possui, que tipo de discretização foi feita, o que a base possui ou não.
- Conversar sobre (potenciais) necessidades (bombeiros/polícia), o que ele deseja saber envolvendo os dados daquela base
  - O que você deseja saber em relação a essa base? É interessante que obtenha essa informação de tempos em tempos, ou variando algum tipo de dado?
- Solicitar que o especialista crie 3 consultas juntamente com o leigo.
- Perguntar se possuem dúvidas e/ou comentários.

Figura C.4: Roteiros das etapas da avaliação com usuários - Reunião com usuários especialistas e leigos (Crisp - criminalidade)

BEM VINDO!

Obrigado por participar deste experimento! Gostaríamos que você nos ajudasse a avaliar o protótipo do Tamanduá 2.0, uma ferramenta de mineração de dados. Essa avaliação nos permitirá analisar o modelo proposto ao criarmos uma nova abstração para o sistema já existente (versão 1.0), além de identificarmos possíveis melhorias para o protótipo.

Durante esta avaliação, você deverá executar as tarefas que serão entregues a seguir. Mas lembre-se: é o sistema que está sendo avaliado e não você! E sinta-se à vontade em expressar sua opinião a qualquer momento da sessão.

Antes de começarmos, alguma dúvida?

Figura C.5: Texto de introdução aos testes

Cenário - VETIBULAR (especialista)

É fato hoje na Universidade que algumas disciplinas, por exemplo, as de grande público como os cálculos, apresentam uma alta taxa de reprovação. Análises do que poderia estar errado devem ser feitas e uma hipótese levantada foi a de que os alunos possam ter sido mal selecionados no vestibular.

Atualmente é feita uma análise em relação às questões do vestibular e a aprovação dos candidatos no mesmo. Além disso, são feitas análises de colocação do candidato no processo de seleção e seu desempenho na universidade.

Porém, não é feita uma análise em relação às questões do vestibular separadamente de forma a analisar o desempenho dos alunos em relação a elas. Assim, poderia ser descoberto, através de análises desse tipo, que questões que envolvem "fração", por exemplo, deveriam aparecer com mais freqüência, de forma a melhor selecionar alunos para as disciplinas como os cálculos, por exemplo.

Uma demanda específica de um pesquisador envolvido, o usuário leigo, foi então de tentar relacionar os desempenhos nas questões do vestibular com o desempenho dos alunos, de forma a verificar quais são boas ou más questões. (Uma boa questão seria uma que poderia "prever" o desempenho do aluno, ou seja, aquela que apresentasse a relação sucesso->sucesso ou fracasso-> fracasso na relação vestibular->disciplina)

Como você sabe, técnicas de mineração de dados podem auxiliar nesse tipo de análise. Entretanto, existem usuários que desejam "minerar" essas informações, mas não possuem o conhecimento técnico necessário em mineração. Para auxiliar esses usuários, você irá desenvolver um papel muito importante criando formas de acesso direto para eles. Para isso, você irá utilizar o Tamanduá, aplicando a técnica de mineração de regras de associação. Você e os usuários leigos pertencem a um mesmo grupo no Tamanduá 2.0, o grupo Vestibular.

As formas de acesso direto a serem criadas são denominadas consultas, que você irá criar no Tamanduá 2.0. Para isso, você deve modelá-las antes no Tamanduá 1.0 (que é o sistema que você já conhece), de forma a encontrar os parâmetros mais adequados para a mineração, como valores para suporte e confiança.

Usando o Tamanduá 1.0, crie tarefas que sejam capazes de responder as perguntas de interesse do usuário, apresentadas a seguir. Para cada tarefa, preencha o formulário em papel que será utilizado na próxima etapa com o Tamanduá 2.0.

Figura C.6: Cenário dos especialistas (Vestibular)

Cenário - CRISP (especialista)

A análise de dados relativos à criminalidade de um local, uma cidade ou todo um estado, por exemplo, é extremamente importante para a atuação do governo. O Crisp, Centro de Estudos de Criminalidade e Segurança Pública, é um órgão voltado para a elaboração, acompanhamento de implementação e avaliação crítica de políticas públicas na área da justiça criminal; ligado à Universidade Federal de Minas Gerais (UFMG) . [Fonte:http://www.crisp.ufmg.br/].

Uma demanda existente é criar formas de acessos e análise a esse tipo de dados para um maior número de pessoas, como pesquisadores do CRISP e/ou policiais (que são os responsáveis também por alimentar as bases de dados).

Tipo de crimes, informações sobre local e o momento das ocorrências dos mesmos são dados interessantes e relevantes que devem ser analisados e que podem apresentar o caminho para importantes atitudes, como aumento de policiamento em um determinado local ou horário.

Como você sabe, técnicas de mineração de dados podem auxiliar nesse tipo de análise. Entretanto, existem usuários que desejam "minerar" essas informações, mas não possuem o conhecimento técnico necessário em mineração. Para auxiliar esses usuários, você irá desenvolver um papel muito importante criando formas de acesso direto para eles. Para isso, você irá utilizar o Tamanduá, aplicando a técnica de mineração de regras de associação. Você e os usuários leigos pertencem a um mesmo grupo no Tamanduá 2.0, o grupo **Crisp**.

As formas de acesso direto a serem criadas são denominadas consultas, que você irá criar no Tamanduá 2.0. Para isso, você deve modelá-las antes no Tamanduá 1.0 (que é o sistema que você já conhece), de forma a encontrar os parâmetros mais adequados para a mineração, como valores para suporte e confiança.

Usando o Tamanduá 1.0, crie tarefas que sejam capazes de responder as perguntas de interesse do usuário apresentadas a seguir. Para cada tarefa, preencha o formulário em papel que será utilizado na próxima etapa com o Tamanduá 2.0.

Figura C.7: Cenário dos especialistas (Crisp)

Consulta 1 (C1)	*Quais locais e momentos ocorrem o tipo de crime <TIPO DE CRIME>?*
Consulta 2 (C2)	*Dada a área <AREA>, quais tipos de crime acontecem e quando eles ocorrem?*
Consulta 3 (C3)	*Considerando o dia sendo <DIASEM >, quais tipos de crimes acontecem e em quais áreas?*
Base	Crisp (2003)
Algoritmo	Eclat
Suporte	0.1
Confiança	20

C1	C2	C3	Código	Categoria	Sub-categorias
			Nrbo	N° do boletim de ocorrência	(n° do boletim)
			Nat	Natureza da ocorrência com código	Homicídio Consumado
			Descrição	Descrição da natureza da ocorrência	
			Codigona	Código da ocorrencia	
A	AC	AC	grupo	Classificação da PMMG em grupos maiores de crime	
			datain	Data do inicio da ocorrência	Data
			Lograd	Tipo de logadouro	Avenida Rua Praça
			Ende	Endereço (nome)	Nome da avenida, rua ou praça
			Num	Número do endereço	Número
			Bairro	Bairro	Nome do bairro
			Munici	Registro posterior	Nome da cidade
			quad	Quadricula de referencia geográfica	
			subsetor		
AC	A	AC	setor		
			Subarea	Sub setor ou CIA de Polícia	Número
			Area	UEOP ou BPM	Número
			Bpm	BPM que registrou a ocorrencia	Número do batalhão
			crpm	Centro Regional de Polícia Militar	
			Hora	Hora do acionamento	Hora do dia
			Turno	Turno	Número
AC	AC	A	Diasem	Dia da semana	1-Domingo 2-Segunda-feira 3-Terça-feira 4-Quarta-feira 5-Quinta-feira 6-Sexta-feira 7-Sábado
			Mês	Mês	Número do mês do ano
			Ano	Ano	Ano do acontecimento
			Bimestre	Número do bimestre de registro da ocorrencia	
			trimestre	Número do trimestre de registro da ocorrencia	
			semestre	Número do semestre de registro da ocorrencia	
			cheg	Hora de chegada para atendimento da ocorrencia	1 – Masculino 2 – Feminino
			encer	Hora de encerramento para atendimento da ocorrencia	(anos)
AC	AC		Horadis		

A= ANTECEDENTE; C=CONSEQUENTE

Figura C.8: Consultas criadas pelo especialista (Crisp - cirminialidade)

Consulta1 (C1)	***Qual a relação entre as notas das questões do vestibular da prova <<PROVA_VESTIBULAR>> e o desempenho nas disciplinas do ICEX?***
Consulta2 (C2)	***Qual a relação entre as notas das questões do vestibular da prova <<PROVA_VESTIBULAR>> e o desempenho nas disciplinas do ICEX para alunos do curso <<CURSO>>?***
Consulta3 (C3)	***Qual a relação entre as notas das questões do vestibular da prova <<PROVA_VESTIBULAR>> e o desempenho na disciplina <<DISCIPLINA>>?***
Base	• Base utilizada: Vestibular UFMG 2004 (2a. Etapa) • Provas: Redação, Física, Matemática e Química • Disciplinas com maior índice de matrículas • Questões das provas da segunda etapa do vestibular
Algoritmo	Eclat
Suporte	15
Confiança	60

C1	C2	C3	Nome	Descrição
			Id	Identificador do aluno
A	A	A	prova_2etapa	Prova de segunda etapa
A	A	A	questao1	Nota da questao 1 da prova de segunda etapa
A	A	A	questao2	Nota da questao 2 da prova de segunda etapa
A	A	A	questao3	Nota da questao 3 da prova de segunda etapa
A	A	A	questao4	Nota da questao 4 da prova de segunda etapa
A	A	A	questao5	Nota da questao 5 da prova de segunda etapa
A	A	A	questao6	Nota da questao 6 da prova de segunda etapa
A	A	A	questao7	Nota da questao 7 da prova de segunda etapa
A	A	A	questao8	Nota da questao 8 da prova de segunda etapa
A	A	A	questao9	Nota da questao 9 da prova de segunda etapa
			redacao	Nota da prova de redacao da 2 etapa
			fisica	Nota da prova de fisica da 2 etapa
			matematica	Nota da prova de matematica da 2 etapa
			quimica	Nota da prova de quimica da 2 etapa
			Sexo	Sexo do aluno
	C	AC	curso	Curso do aluno
C	C	C	disciplina	Disciplina
		C	periodo_cursado	Periodo em que a disciplina foi cursada
C	C	C	conceito	Conceito da disciplina
			ofertante	Ofertante da disciplina
			frequencia	Frequencia

= ANTECEDENTE; C=CONSEQUENTE

Observações	**Discretização** • As notas das disciplinas foram discretizadas em sucesso (S), se o aluno passou na disciplina, e insucesso (I), caso contrário. • As notas das questões foram discretizadas em sucesso (S), se o aluno tirou mais de 60% do seu valor, e insucesso (I), caso contrário. • As notas totais das provas do vestibular foram discretizadas em BEM, se o aluno tirou uma nota maior que média do seu curso, e MAL caso contrário.

Figura C.9: Consultas criadas pelo especialista (Vestibular)

Cenário - VESTIBULAR (leigo)

É fato hoje na Universidade que algumas disciplinas, por exemplo, as de grande público como os cálculos, apresentam uma alta taxa de reprovação. Análises do que poderia estar errado devem ser feitas e uma hipótese levantada foi a de que os alunos possam ter sido mal selecionados no vestibular. Dessa forma, foi sugerida uma análise em relação às questões do vestibular separadamente de forma a analisar o desempenho dos alunos em relação a elas.

Entretanto, o volume de dados a ser analisado é enorme e a análise é impossível de ser feita de forma manual. Para auxiliar nesses estudos, existem as técnicas de mineração de dados. Essas técnicas envolvem um grande conhecimento específico e técnico que nem sempre são intuitivos e de fácil aprendizagem.

Visando criar formas de acesso direto a informações importantes, foram criadas consultas no Tamanduá 2.0. Você e o usuário especialista que as criou pertencem a um mesmo grupo no sistema, o grupo Vestibular.

Imagine as três situações a seguir e veja se as consultas criadas auxiliam na análise dos dados que deseja.

1. [Consulta 1] A comissão da prova de **física** do vestibular será mudada. Porém, antes de iniciar os trabalhos para o próximo vestibular, eles gostariam de saber como é o rendimento dos alunos atualmente em relação ao desempenho no ICEX. Veja que tipo de informação pode ser interessante nessa relação.

2. [Consulta 2] O coordenador do curso de engenharia civil verificou que seus alunos possuem muita dificuldade de escrita. Em provas e trabalhos, os professores apontaram que existem muitos erros ortográficos e gramaticais. Veja a relação que existe entre a prova de **redação** e o curso de **engenharia civil.**

   Crie uma nova consulta para a prova de matemática, para o mesmo curso

3. [Consulta 3] Os professores de **AEDSI** verificaram que os alunos apresentam muita dificuldade em realizar contas matemáticas simples. Eles ficaram curiosos em saber como foi o rendimento deles na prova de **matemática** do vestibular. Verifique se encontra algum dado

Figura C.10: Cenário e Tarefas dos leigos (Vestibular)

Cenário - CRISP (leigo)

A análise de dados relativos à criminalidade de um local, uma cidade ou todo um estado, por exemplo, é extremamente importante para a atuação do governo. Entretanto, o volume de dados que pode ser armazenado é enorme e impossível de ser feita uma análise manual. Para auxiliar nesses estudos, existem as técnicas de mineração de dados. Essas técnicas envolvem um grande conhecimento específico e técnico que nem sempre são intuitivos e de fácil aprendizagem.

Visando criar formas de acesso direto a informações importantes, foram criadas consultas no Tamanduá 2.0. Você e o usuário especialista que as criou pertencem a um mesmo grupo no sistema, o grupo Crisp.

Imagine as três situações a seguir e veja se as consultas criadas auxiliam na análise dos dados que deseja.

1. [*Tarefa Crisp Area*] Uma nova campanha para prevenção de crimes será implantada. O governo irá investir em pessoal e material, mas para isso precisa saber dados em relação aos crimes. O **setor 51** foi priorizado e você gostaria de ver se encontra alguma informação relevante sobre os crimes que ocorrem nesse local.

2. [*Tarefa Crisp Crime*] É necessário analisar se há alguma medida que pode ser tomada para se evitar um tipo específico de crime. Existem alguns tipos de crimes podem estar relacionados a locais e dias distintos, podendo ou não haver correspondência entre esses fatores.

   Verifique se existe alguma informação relevante em relação ao crime "**contra pessoa**".
   Veja também se as informações da consulta que está executando lhe parecem interessantes e úteis.

3. [*Tarefa Crisp Tempo*] Uma nova escala de policiais será feita por dia da semana, dessa forma, seria interessante visualizar informações sobre o tipo de crime e local onde eles ocorrem separadamente para cada dia.

   Veja as informações relacionadas a **segunda-feira**.
   Faça uma nova consulta para **domingo**.

Figura C.11: Cenário e Tarefas dos leigos (Crisp - criminalidade)

Roteiro para entrevista Pós-Teste (especialistas)

- Experiência com o Tamanduá
  - Quais projetos
  - Modelagens feitas
  - Tempo de uso
  - Se já tiveram clientes
- Como apresentaram os resultados
- Como definiram o problema
- Formação em MD
  - Disciplina
  - Cursos
- Criação das consultas
  - Qual custo de criação
  - Dado que fez a modelagem, qual o custo de criar consultas?
  - Em que situações usaria uma consulta?
- Qual nível de formação e em que área atua

Figura C.12: Roteiro para entrevista pós-testes (especialistas)

Roteiro para entrevista Pós-Teste (leigos)

- Experiência com Mineração de Dados
  - o Conceitos
  - o Sistemas
  - o Tamanduá
- Consultas
  - o Consolidação da consulta
  - o Acha que existem problemas na modelagem ou ferramenta?
- Qual nível de formação e em que área atua

Figura C.13: Roteiro para entrevista pós-testes (leigos)

Problemas de usabilidades levantados (agrupados por heurísticas)

De uma maneira geral, os tipos de problemas de usabilidade encontrados relacionados às diretrizes de Nielsen.

> **ajuda aos usuários para reconhecerem, diagnosticarem e se recuperarem de erros:** para algumas tarefas, as mensagens para os usuários não estão claras. Para cada tipo de erro, uma mensagem deve indicar o que ocorreu e o que deve ser feito, o que não está acontecendo em todos os casos no sistema. Por exemplo, ao criar uma nova execução, se ocorrer um determinado erro, ele não é descrito para o usuário, sendo apenas indicado que um erro ocorreu.

> **correspondência entre o sistema e o mundo real:** devem ser utilizados ícones e terminologias familiares aos usuários. Foram encontrados alguns ícones que se tornaram confusos aos usuários, ou mesmo *hints* que não descreviam a ação que realmente era acionada. Por exemplo, o ícone para criar nova tarefa aparentemente não estava claro aos usuários e seu *hint* ``Definir execução'' não auxiliava no entendimento.

> **visibilidade do estado do sistema:** os usuários devem ser mantidos informados sobre o que está ocorrendo no sistema. Em relação a esse aspecto, um dos problemas levantados foi em relação ao estado de visualização dos ícones, que não apresentam diferenciação visual ao estarem desabilitados.

Figura C.14: Exemplos de problemas de usabilidade encontrados durante a avalição com usuários

Título do Trabalho: Avaliação do protótipo do Tamanduá 2.0
Data: Junho/2008
Instituição: DCC / UFMG
Pesquisadores Responsáveis:
- Elisa Tuler de Albergaria (elisa@dcc.ufmg.br)
- Prof. Raquel O. Prates (rprates@dcc.ufmg.br)

Introdução: Este Termo de Consentimento contém informações sobre o projeto de pesquisa indicado acima. Caso tenha alguma dúvida, não hesite em perguntar ao pesquisador responsável.

Objetivo da Pesquisa: O objetivo desse trabalho consiste em avaliar o protótipo do modelo proposto no trabalho de pesquisa de mestrado da aluna Elisa Tuler de Albergaria. O modelo visa aumentar a usabilidade de sistemas de mineração de dados por regras de associação de segunda geração. O protótipo consiste na versão 2.0 do sistema Tamanduá, que implementa o modelo proposto.

Sobre a avaliação: Será solicitado que você execute algumas tarefas simples utilizando o sistema Tamanduá. Essa interação será gravada para posterior análise dos resultados obtidos. Ao final das tarefas, uma entrevista será realizada para analisar sua experiência com o sistema.

Utilização dos resultados da avaliação: Pede-se consentimento para utilização dos dados resultantes da avaliação que será realizada para a pesquisa indicada acima.

Privacidade: Não será divulgada nenhuma informação pessoal, que possa identificá-lo. Todos os dados da pesquisa serão anonimizados.

Participação voluntária: Você é livre para decidir se quer participar da avaliação para fins desta pesquisa ou não. Sua decisão não afetará sua vida estudantil e nem qualquer relacionamento com os avaliadores, professores ou a Instituição por trás desta.

Compensação: A participação na avaliação é voluntária, não sendo oferecida nenhuma remuneração.

Novas condições: Caso deseje, você pode especificar novas condições que devem ser atendidas.

Consentimento Livre e Esclarecido (Acordo Voluntário)

O documento mencionado acima descrevendo o termo de consentimento foi lido e explicado. Eu tive a oportunidade de fazer perguntas sobre o mesmo, que foram respondidas

_____	Assinatura do aluno: _____
Data	Nome do aluno:

Assinatura do pesquisador: _____

Nome da pesquisadora: Elisa Tuler de Albergaria

Figura C.15: Termo de consentimento para citação dos cenários dos alunos de mineração de dados

Referências Bibliográficas

Agrawal, R.; Imielinsky, T. e Swami, A. (1993). Mining association rules between sets of items in large databases. pp. 207–216. Proc. ACM SIGMOD 93.

Albergaria, E.; Mourão, F.; Prates, R. e Wagner Meira, J. (2008a). An end user development model do augment usability of rule association mining systems. *In: 20th IFIP World Computer Congress - WCC.*

Albergaria, E.; Mourão, F.; Prates, R. e Wagner Meira, J. (2008b). Modelo de interface extensível como solução para desafios de interação em sistemas de mineração de dados. *Seminário Integrado deHardware e Software (SEMISH).*

Albergaria, E.; Prates, R. O.; Almir, F.; Rocha, L. e Wagner Meira, J. (2006). Caracterizando desafios de interação com sistemas de mineração de regras de associação. In *IHC '06: Proceedings of VII Brazilian symposium on Human factors in computing systems*, pp. 40–49, New York, NY, USA. ACM.

Analysis, M. B. (2006). http://www.megaputer.com/products/pa/tutorials/index-tut.html.

Barlow, J.; Rada, R. e Diaper, D. (1989). Interacting with computers. *Interact. Comput.*, 1(1):39–42.

Berkhin, P. (2002). Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA.

Bernstein, A. e Provost, F. (2001). An intelligent assistant for the knowledge discovery process.

Bim, S. A.; Leitão, C. F. e de Souza, C. S. (2007). The challenge of teaching hci qualitative evaluation methods: a case study on the communicability evaluation method. Technical report.

Brazdil, P.; Soares, C. e da Costa, J. P. (2003). Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. *Machine Learning*, 50(3):251–277.

Carroll, J. M. (2000). *Making Use: Scenario-Based Design of Human-Computer Interactions.* MIT Press, Cambridge, MA, USA.

Cortes, S.C; Porcaro, R. L. (2002). Mineração de dados - funcionalidades, técnicas e abordagens. Technical report.

Crisp (2008). http://www.crisp.ufmg.br.

Dama (2006). www.dis.uniroma1.it/ lembo/D2I/Prodotti/deliverable3/D3.P4.pdf.

de Souza, C. (2005). *The semiotic engineering of human-computer interaction.* MIT Press, Cambridge.

de Souza, C. S. e Barbosa, S. D. J. (2006). A semiotic framing for end-user development. pp. 401–426.

do Tamandua, E. (2005). Manual de utilização - tamanduá. Technical report.

Domingues, M. A. e Rezende, S. O. (2005). Using taxonomies to facilitate the analysis of association rules. In *Proceedings of ECML/PKDD'05 The Second International Workshop on Knowledge Discovery and Ontologies (KDO-2005)*, pp. 59–66.

Eco, U. (1976). *A Theory of Semiotics.* Advances in Semiotics. Indiana University Press, Bloomington.

Fayyad, U. M.; Piatetsky-Shapiro, G. e Smyth, P. (1996). From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining*, pp. 1–34.

Ferreira, R.; Jr., W. M.; Guedes, D. e Drumond, L. (2005). Anthill: A scalable run-time environment for data mining applications. In *SBAC-PAD'05: The 17th International Symposium on Computer Architecture and High Performance Computing*, Rio de Janeiro, Brazil.

Fischer, G. (2007). Meta-design: Expanding boundaries and redistributing control in design. In Baranauskas, M. C. C.; Palanque, P. A.; Abascal, J. e Barbosa, S. D. J., editores, *INTERACT (1)*, volume 4662 of *Lecture Notes in Computer Science*, pp. 193–206. Springer.

Fischer, G.; Giaccardi, E.; Ye, Y.; Sutcliffe, A. G. e Mehandjiev, N. (2004). Meta-design: a manifesto for end-user development. *Commun. ACM*, 47(9):33–37.

Goldschmidt, R. (2005). *Data Mining - Um guia prático.* Editora Campos.

Goldschmidt, R.; E., P. e M., V. (2002). An action plan definition assitant in kdd process. *Proceedings of the Second International Conference on Artificial Intelligence and Aplications.*

Goldschmidt, R. R. (2003). *Assistência Inteligente à Orientação do Processo de Descoberta de Conhecimento em Bases de Dados.* PhD thesis, PUC-Rio, Brasil.

Gonçalves, L. P. F. (2001). Avaliação de ferramentas de mineração de dados como fonte de dados relevantes para a tomada de decisão: Aplicação na rede unidão de supermercados, são leopoldo-rs. Master's thesis, UFRGS, Brasil.

Group, D. M. (2004). Pmml 3.0 predictive model markup language. Technical report. http://www.dmg.org.

Han, J.; Cheng, H.; Xin, D. e Yan, X. (2007). Frequent pattern mining: current status and future directions. *Data Min. Knowl. Discov.*, 15(1):55–86.

Han, J.; Fu, Y.; Wang, W.; Chiang, J.; Gong, W.; Koperski, K.; Li, D.; Lu, Y.; Rajan, A.; Stefanovic, N.; Xia, B. e Zaiane, O. R. (1996). DBMiner: A system for mining knowledge in large relational databases. In *Proc. 1996 Int'l Conf. on Data Mining and Knowledge Discovery (KDD'96)*, pp. 250–255, Portland, Oregon.

Hipp, J.; Güntzer, U. e Nakhaeizadeh, G. (2000). Algorithms for association rule mining — a general survey and comparison. *SIGKDD Explorations*, 2(1):58–64.

Hofmann, H.; Siebes, A. P. J. M. e Wilhelm, A. F. X. (2000). Visualizing association rules with interactive mosaic plots. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 227–235, New York, NY, USA. ACM.

ISO9241 (2008). http://www.iso.org/iso/en/ISOOnline.frontpage.

Jakobson, R. (1960). *Closing Statements: Linguistics and Poetics*. MIT Press, New York.

Jiawei Han, M. K. (2001). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Khabaza, T. e Shearer, C. (1995). Data mining with clementine. *IEE Seminar Digests*, 1995(21B):1–1.

Kriegel, H.-P.; Borgwardt, K. M.; Kröger, P.; Pryakhin, A.; Schubert, M. e Zimek, A. (2007). Future trends in data mining. *Data Min. Knowl. Discov.*, 15(1):87–97.

Kuranov, A. L.; Korabelnicov, A. V.; Kichinskiy, V. V. e Sheiken, E. G. (2001). Fundamental techniques of the ajax concept - modern state of research. *AIAA/NAL-NASDA-ISAS International Space Planes and Hypersonic Systems and Technologies Conference*, 10:24–27.

Lieberman, H.; Paterno, F. e Wulf, V., editores (2006). *End User Development*. Springer.

MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297.

Mei, Q.; Xin, D.; Cheng, H.; Han, J. e Zhai, C. (2006). Generating semantic annotations for frequent patterns with context analysis. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 337–346, New York, NY, USA. ACM.

Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, 42(11):30–36.

Moran, T. P. (1981). The command language grammar: A representation for the user interface of interactive computer systems. *International Journal of Man-Machine Studies*, 15(1):3–50.

Morch, A. (1997). Three levels of end-user tailoring: customization, integration, and extension. pp. 51–76.

Morik, K. (2000). The representation race – preprocessing for handling time phenomena. In *Machine Learning: ECML 2000, 11th European Conference on Machine Learning, Barcelona, Catalonia, Spain, May 31 - June 2, 2000, Proceedings*, volume 1810, pp. 4–19. Springer, Berlin.

Mourão, F.; Albergaria, E.; Graciano, B.; Teixeira, T.; Prates, R. e Wagner Meira, J. (2008). Tamanduá 2.0: Arquitetura e modelagens. Technical Report RT.DCC.009/2008.

Myers, B. A. (1992). *Languages for Developing User Interfaces*. Jones and Bartlett Publishers, Inc.

Nardi, B. A. (1993). *A small matter of programming: perspectives on end user computing.* MIT Press.

Nascimento, F. A. (2005). Visualização de regras de associação. Master's thesis, UFMG, Brasil.

Newing, R. (1996). Mineração de dados. *Management Accounting*, pp. 34–35.

Nielsen, J. (1994). Usability inspection methods. In *CHI '94: Conference companion on Human factors in computing systems*, pp. 413–414, New York, NY, USA. ACM.

Norman, D. A. (1986). Cognitive engineering. In Norman, D. A. e Draper, S. W., editores, *User Centered System Design: New Perspectives on Human-Computer Interaction*, pp. 31–61. Erlbaum, Hillsdale, NJ.

Norman, D. A. (1988). *The Psychology of Everyday Things.* Basic Books, New York.

Peirce, C. S. (1931-1958). *Collected papers of Charles Sanders Peirce*, volume 1-8. Hartshorne and P. Weiss. Cambridge, MA. Harvard University Press.

Piatetsky-Shapiro, G. (1999). The data-mining industry coming of age. *IEEE Intelligent Systems*, 14(6):32–34.

Prates, R. O.; de Souza, C. S. e Barbosa, S. D. J. (2000). Methods and tools: a method for evaluating the communicability of user interfaces. *interactions*, 7(1):31–38.

Preece, J.; Rogers, Y.; Sharp, H.; Benyon, D.; Holland, S. e Carey, T. (1994). *Human-Computer Interaction.* Addison Wesley, England.

Rainho, P. S. (2001). Mineração de dados - conceitos, técnicas e aplicações. Technical report.

Rainsford, C. e Roddick, J. (2000). Visualization of temporal interval association rules. *Proceeding of the Second International Conference on Intelligent Data Engineering and Automated Learning,*, pp. 91–96.

Ralambondrainy, H. (1995). A conceptual version of the k-means algorithm. *Pattern Recogn. Lett.*, 16(11):1147–1157.

SBC (2006). Relatório preliminar dos grandes desafios da pesquisa em computação no brasil. Technical report. http://www.ic.unicamp.br/ cmbm/desafios_SBC/.

Senator, T. E.; Goldberg, H. G.; Shyr, P.; Bennett, S.; Donoho, S. e Lovell, C. (2002). The nasd regulation advanced detection system: integrating data mining and visualization for break detection in the nasdaq stock market. pp. 363–371.

Silveira, M. S.; Barbosa, S. D. J. e de Souza, C. S. (2004). Designing online help systems for reflective users. *Journal of the Brazilian Computer Society*, 9(3):25–38.

Soares, C.; Costa, J. e Bradzil, P. (2001). Improved statistical support to matchmaking: Rank correlation taking rank importance into account. *JOCLAD 2001: VII Jornada de Classificação e Análise de Dados.*

Srikant, R. e Agrawal, R. (1997). Mining generalized association rules. *Future Generation Computer Systems*, 13(2–3):161–180.

Suchman, L. (1987). *Plans and Situated Actions*. Cambridge University Press.

Tamandua (2006). http://tamandua.speed.dcc.ufmg.br.

Thearling, K.; Becker, B.; DeCoste, D.; Mawby, W. D.; Pilote, M. e Sommerfield, D. (2002). Visualizing data mining models. pp. 205–222.

Tutorial, D. (2006). http://www.cs.sfu.ca/CC/459/han/tutorial.

Weiss, S. M. e Indurkhya, N. (1998). *Predictive data mining: a practical guide*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Weka (2006). http://www.cs.waikato.ac.nz/ml/weka/.

Wong, P. C.; Whitney, P. e Thomas, J. (1999). Visualizing association rules for text mining. In *INFOVIS*, pp. 120–123.

XLMiner (2006). http://www.resample.com/xlminer/help/Assocrules/associationrules-ex.htm.

Zaki, M. J. e Phoophakdee, B. (2003). Mirage: A framework for mining, exploring and visualizing minimal association rules.