# Modelo de Interface Extensível como Solução para Desafios de Interação em Sistemas de Mineração de Dados

**Elisa Albergaria, Fernando Mourão, Raquel Prates, Wagner Meira Jr.**

Departamento de Ciência da Computação - Universidade Federal de Minas Gerais
UFMG - Belo Horizonte - Minas Gerais - Brasil

{elisa,fhmourao,rprates,meira}@dcc.ufmg.br

***Abstract.*** *Currently, one of the great challenges of computing is the enormous volume of data generated by the increasing use of the Internet by businesses, governments and individuals. The data mining field focuses on how to generate knowledge from large volumes of data. However, data mining systems are usually difficult to use, since they require users to have technical knowledge about the technique being used. This work aims at broadening the usage of such systems. To do so, we present an extensible interface model that allows for the creation of new interfaces at a higher level of abstraction for a specific context.*

***Resumo.*** *Atualmente, um dos grandes desafios da computação é o enorme volume de dados gerado com o crescente uso da Internet por empresas, governos e indivíduos. A área de mineração de dados tem por objetivo a geração de conhecimento a partir de grandes volumes de dados. No entanto, estes 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 um modelo de interface extensível que permite criar novas interfaces de mais alto nível e específicas para um contexto, abstraindo o conhecimento técnico.*

## 1. Introdução

A Sociedade Brasileira de Computação (SBC) identificou cinco grandes problemas a serem investigados pela comunidade científica nos próximos dez anos [SBC 2006]. Este trabalho apresenta a pesquisa que está sendo feita 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.

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: pp 8]. 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, buscam-se soluções que envolvam o cidadão, que

passaria de usuário passivo a ativo e participativo na geração do conhecimento [SBC 2006, pp 17].

Mineração de dados é a área da computação que surgiu há mais de 20 anos atrás com o objetivo de lidar com grandes volumes de dados, permitindo a descoberta de novo conhecimento a partir dos mesmos [Fayyad et al. 1996]. Áreas como a bioinformática, economia, sociologia, astrofísica, dentre outras, encontraram na aplicação desse tipo de busca uma importante fonte de informações algumas vezes preponderantes. Os sistemas de mineração podem ser classificados em 4 diferentes gerações [Piatetsky-Shapiro 1999].

Sistemas de primeira geração focavam em uma tarefa específica como classificadores utilizando redes neurais, agrupamento (*clustering*) (e.g. algoritmo *K-means* [Ralambondrainy 1995]) ou mesmo a visualização dos dados. Sistemas de segunda geração representam sistemas que oferecem suporte a mais de uma etapa do processo, possibilitando realizar diversas tarefas de descoberta e apresentando mais de um tipo de análise de dado (e.g. Clementine [Khabaza and Shearer 1995], Tamanduá [Ferreira et al. 2005][Guedes Neto et al. 2006], WEKA [Weka] e DBMiner [DBMiner]). Embora de aplicação bastante ampla, sistemas de segunda geração normalmente requerem um conhecimento grande sobre técnicas específicas de mineração de dados por parte dos usuários para utilizá-las. Para lidar com este desafio surgem então dois tipos de sistemas: aqueles que são mais próximos aos usuários e específicos para um problema e contexto (sistemas de terceira geração), e aqueles que auxiliam os usuários no processo de tomada de decisão durante o processo de geração de conhecimento, através da apresentação dos conceitos envolvidos (sistemas de quarta geração).

Sistemas de 2ª geração são os mais comumente utilizados, uma vez que permitem uma ampla flexibilidade na mineração a ser feita, visto que não são específicos para um contexto em particular. Um dos desafios destes sistemas atualmente é que eles requerem um conhecimento técnico aprofundado dos algoritmos e parâmetros de mineração de dados [Albergaria et al. 2006][Kriegel et al., 2007]. Assim, atualmente, o uso destes sistemas está restrito a um grupo de pessoas que detém o conhecimento técnico necessário, conhecimento esse que normalmente é obtido em disciplinas oferecidas em cursos de Ciência da Computação ou Sistemas de Informação, ou em treinamentos e cursos sobre os próprios sistemas.

Neste trabalho, apresentamos uma solução que tem por objetivo ampliar o acesso a sistemas de mineração de dados de segunda geração e, logo, ao conhecimento que estes são capazes de gerar. Para isso, propomos um modelo de um módulo extensível a ser acrescentado à interface de um sistema de segunda geração de mineração de dados, permitindo a criação de uma nova interface de mais alto nível que oferece consultas específicas para um domínio e que abstrai o conhecimento técnico necessário. Teoricamente, o módulo pode ser acoplado a qualquer sistema de segunda geração. Como prova de conceito do modelo, neste trabalho apresentamos um protótipo desenvolvido especificamente para o sistema Tamanduá [Tamanduá] (sistema de mineração de segunda geração). Apesar do Tamanduá oferecer técnicas distintas de mineração, como primeiro passo focamos na mineração por regras de associação, uma das técnicas mais populares de geração de conhecimento por identificação de freqüência de padrões freqüentes [Hipp et al. 2000].

Sistemas de mineração de dados de 2ª. geração endereçam questões levantadas pelo desafio 1, uma vez que oferecem um conjunto de técnicas para se obter conhecimento a partir de um grande número de dados. No entanto, o conhecimento técnico necessário para se utilizar estes sistemas restringe o seu uso a poucos especialistas na área da computação, limitando então a ampla disseminação da informação, outro objetivo do desafio 1. Considerando-se que estes sistemas poderiam ser utilizados em bases de dados do governo para gerar informação de interesse da população, o requisito de conhecimento técnico iria de encontro ao desafio 4, uma vez que seu uso não seria possível pelos cidadãos brasileiros. Desta forma, este trabalho, ao propor uma solução em que se permite acesso à informação gerada por sistemas de mineração de dados de 2ª. geração pelo usuário leigo, endereça uma questão que se encontra na interseção dos desafios 1 e 4.

A seguir, na seção 2 apresentamos brevemente o conceito de mineração de regras de associação, técnica de mineração de dados que é o foco desse trabalho e os desafios de uso de sistemas de mineração que a adotam. Na seção 4 é apresentado o modelo proposto, sua arquitetura e descrição de seus componentes. A seção 3 apresenta o Tamanduá, sistema de segunda geração que foi utilizado para acoplar o modelo proposto e o protótipo, aplicação do modelo no sistema Tamanduá. Por último são apresentadas conclusões e trabalhos futuros relacionados à pesquisa aqui descrita.

## 2. Desafios da Mineração de Regras de Associação

Existem várias técnicas de mineração de dados, sendo que dentre elas algumas se destacam como: análise de agrupamento (clusters), classificação e regras de associação. Este trabalho foca na técnica de regras de associação, que tem a funcionalidade objetiva de encontrar correlações interessantes entre itens de uma base de dados. A mineração de regras de associação é uma popular técnica de mineração de dados, tendo sido introduzida por Agrawal et al. [1993].

Uma regra de associação representa uma relação entre dois ou mais itens de uma base de dados. Consideremos, por exemplo, a regra apresentada a seguir:

**[Pão], [Manteiga] => [Leite] (80.00, 50.00)**

Esta 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: 50% das compras realizadas pelos clientes da padaria incluem pão, leite e manteiga; e das compras que incluem pão e manteiga, 80% também incluem leite. 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.

O primeiro valor 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 80%. Os termos confiança, freqüência relativa e probabilidade condicional podem ser usados de forma intercambiável.

O segundo valor corresponde ao suporte da regra. O suporte 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. 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.

Os sistemas de mineração de regras de associação de 2ª. geração requerem do usuário conhecimento sobre parâmetros e estruturas de regras de associação para utilizá-lo, o que dificulta seu uso pelo usuário leigo em mineração de dados. Em [Albergaria et al. 2006] são apresentados desafios de interação em relação a sistemas de segunda geração, em relação à técnica de mineração de regras de associação. Para obterem o que desejam, é necessário que os usuários entendam os conceitos dos parâmetros de entrada, como suporte e confiança, saibam também atribuir valores a eles e conheçam o impacto de cada um durante a execução do algoritmo. Além disso, os usuários devem conhecer as outras medidas de interesse envolvidas na técnica, de forma a melhor analisar o conjunto resultante das regras. A escolha dos atributos a serem minerados também é uma tarefa que demanda conhecimento dos usuários, de forma que eles possam analisar as regras geradas posteriormente. Conceitos inerentes às regras também são importantes para a interação, como antecedente e conseqüente.

As dificuldades experimentadas pelos usuários se distribuem ao longo do processo de mineração, indo desde a definição de parâmetros de entrada até a visualização. Isso envolve configurar uma série de parâmetros, sendo 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), quanto para os responsáveis pelo sistema (o usuário pode desistir de utilizar o sistema).

Os problemas citados referem a sistemas de mineração de regras de associação de 2ª geração em geral [Albergaria et al. 2006]. A necessidade de se aumentar a usabilidade de sistemas de Mineração de Dados tem sido considerada por alguns pesquisadores como uma das principais questões de pesquisa atuais para a área de Mineração de Dados [Kriegel et al. 2007][ Mei et al. 2006]. Assim, embora a tecnologia esteja disponível para se obter a informação, o acesso a ela é restrito por alguns poucos que detêm o conhecimento necessário. Para que haja possibilidade de real disseminação da informação é preciso que se amplie o acesso a esta informação a usuários leigos.

## 3. Uma Solução através de Interfaces Extensíveis

Para permitir um maior acesso à informação gerada por sistemas de mineração por regras de associação, sem, no entanto, restringir a aplicabilidade do sistema, propomos que as interfaces destes sejam extensíveis. A idéia é permitir que novas interfaces sejam criadas em um nível mais alto de abstração que sejam específicas para um determinado contexto e que abstraiam o conhecimento técnico envolvidos na mineração. Para isso, apresentamos uma proposta de um modelo de arquitetura para um módulo de extensão a ser acoplado a sistemas de mineração de regras de associação de 2ª geração.

O modelo é baseado na teoria da engenharia semiótica. A engenharia semiótica é uma teoria de interação humano-computador que caracteriza a interação como sendo

uma comunicação do designer do sistema aos seus usuários sobre a quem o sistema se destina, que problemas pode resolver e como interagir com ele para resolvê-lo [de Souza 2005]. Assim, o modelo, ao permitir que o usuário estenda a interface do sistema, promove o usuário de receptor da mensagem a co-autor desta [Albergaria et al., 2008].

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 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 criam extensões que abstraem os conhecimentos técnicos e permitem aos usuários leigos interagirem com o sistema para resolver um problema específico.

A Figura 1 mostra a arquitetura proposta pelo modelo, onde seus principais componentes são a Linguagem Abstrata da Interface com Usuário (LAIU), o gerador e a base de conhecimento.

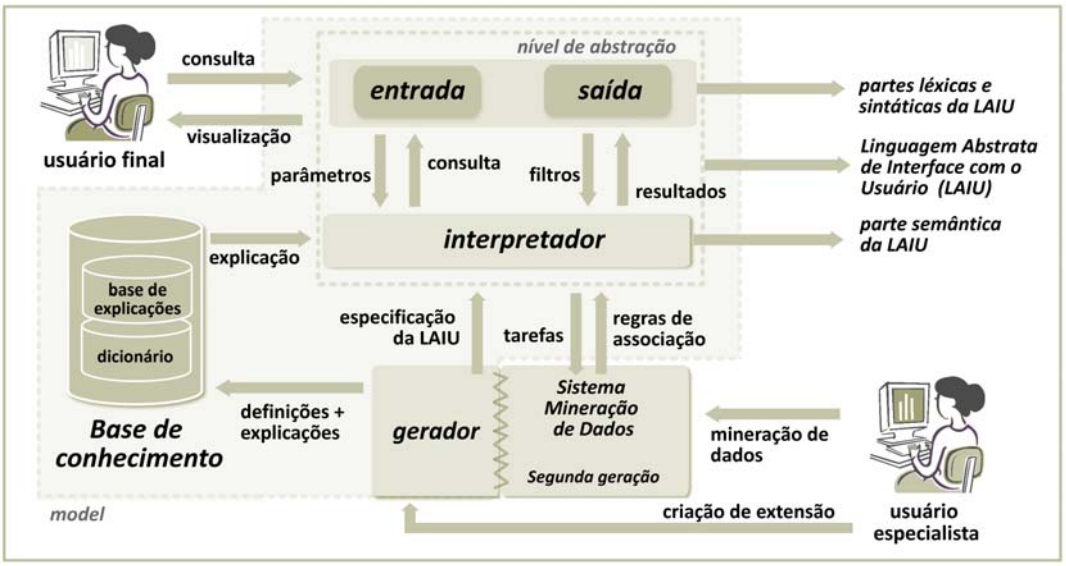



**Figura 1. Estrutura do modelo proposto**

O **gerador** é a interface disponível para o usuário especialista criar as extensões que compõem a camada de abstração da LAIU para que o usuário leigo possa interagir com o sistema. O gerador requer que o usuário especialista defina 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 os novos elementos de entrada disponibilizados na camada de abstração para usuários leigos de consultas. Nesse caso, os especialistas determinam quais parâmetros terão valor fixo e quais serão definidos pelos leigos em tempo de uso (*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.

A **base de conhecimento** tem por objetivo permitir aos usuários especialistas registrarem 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 e outras que são mais técnicas, direcionadas aos próprios usuários especialistas. O dicionário armazena o significado dos elementos da linguagem de interface criada, ou seja, o que eles representam no sistema de mineração de segunda geração. Embora a base de conhecimento não seja um componente necessário para a criação e execução de consultas, ela tem um papel fundamental na forma como as pessoas as utilizam e interpretam.

A **Linguagem Abstrata de Interface com o Usuário** é criada pelo especialista para o leigo. A parte léxica e sintática da LAIU é composta pelos elementos de entrada e saída disponibilizados para o usuário leigo. A entrada é composta por uma ou mais consultas específicas para o usuário leigo. Uma consulta denomina 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 leigos. Assim como o especialista especifica a consulta, ele também define como os resultados serão apresentados ao usuário leigo. A parte semântica da linguagem representa o seu comportamento, ou seja, a mineração a ser executada. O interpretador é o responsável por associar a semântica da LAIU à abstração criada.

O interpretador 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. 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.

Para avaliar o modelo proposto, foram definidos dois aspectos a serem considerados: (1) se os usuários especialistas poderiam de fato criar abstrações (consultas) úteis para os usuários finais; (2) se um novo módulo de extensão baseado no modelo poderia ser acoplado (implementado) a um sistema de mineração de segunda geração existente. Para avaliar a questão 1, foram usados cenários – narrativas detalhadas e plausíveis que descrevem uma determinada situação [Carroll, 2000]. Após os resultados obtidos para avaliação desta questão, usou-se o Tamanduá para avaliar-se a questão 2.

A avaliação da viabilidade das abstrações foi feita com alunos da disciplina de Mineração de Dados do Departamento de Ciência da Computação (DCC) da UFMG ministrada em 2007-2. No projeto da disciplina, os alunos tinham que desenvolver um projeto de mineração de regras de associação modelando um problema real e apresentando os resultados reais. Como parte do entendimento e planejamento da tarefa

de mineração de dados a ser executada, os alunos tiveram que definir cenários de uso do ambiente por usuários finais (não especialistas). Assim, o trabalho foi dividido em 4 etapas: (1) escolher tema com problema real; (2) levantar a necessidade dos usuários, criar propostas de consultas que pudessem atendê-las e gerar cenários que descrevessem as necessidades e consultas; (3) fazer a modelagem das tarefas subjacentes às consultas propostas no Tamanduá (interface original); (4) propor formas de visualização abstrata das regras obtidas para os usuários finais.

Os projetos foram desenvolvidos normalmente em grupos de 2 estudantes e dos 12 grupos que terminaram o projeto, 4 foram avaliados pelo professor como não tendo alcançado os objetivos propostos (em geral por não atenderem aos requisitos do trabalho ou a modelagem da tarefa de mineração feita não ter sido considerada adequada). Estes trabalhos não foram considerados para a avaliação do modelo uma vez que esta envolvia a avaliação da abstração obtida, e estes trabalhos apresentavam problemas que deviam ser considerados antes de ser possível avaliar esta abstração.

Os 8 trabalhos considerados adequados para avaliação foram então avaliados em relação à abstração proposta[1]. Estes projetos foram aplicados 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). A avaliação procurou verificar se os alunos (considerados usuários especialistas) eram capazes de criar uma abstração que pudesse ser aplicada em um determinado problema de forma que os usuários finais pudessem interagir sem terem que entender os conceitos técnicos envolvidos. Todos os 8 grupos foram capazes de criar bons níveis de abstração (entrada e saída de dados). 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 o dicionário não tenha sido apresentado aos alunos, a maioria dos trabalhos apresentou um mapeamento entre os elementos que foram criados na interface de alto nível e os elementos da interface do sistema de mineração.

Uma vez que se obteve indicadores positivos sobre a proposta do modelo de criação de abstrações para possibilitar a interação de usuários leigos, passou-se então para a avaliação de se um módulo extensível baseado no modelo poderia ser acoplado a um sistema de 2ª geração existente. Assim, foi implementado no Tamanduá um protótipo baseado no modelo proposto. A próxima seção apresenta o sistema Tamanduá, o protótipo (em fase final de desenvolvimento) e ilustra um cenário de uso deste.

## 4. Protótipo de Interface Extensível para o Tamanduá

O protótipo foi desenvolvido na plataforma Java, utilizando Java Server Faces com suporte a interfaces ricas criadas com conceito Ajax [Kuranov et al. 2001]. O sistema de 2ª geração escolhido para o modelo ser aplicado foi o Tamanduá, descrito a seguir.

[1] 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.

### 4.1. Sistema Tamanduá

O sistema Tamanduá [Ferreira et al., 2005][Guedes Neto et al. 2006] é uma plataforma de serviços de mineração de dados de 2ª geração desenvolvido pelo DCC/UFMG. Trata-se de uma ferramenta para executar tarefas de mineração simultaneamente em um ambiente web multi-usuário. Estas tarefas podem ser executadas sobre as mesmas bases de dados e garantem um bom tempo de resposta para um grande número de usuários utilizando diferentes estações de trabalho.

Podemos considerar o Tamanduá um sistema de mineração de objetivo geral, no sentido que ele procura oferecer aos usuários oportunidades de encontrar padrões interessantes em uma base de dados, sem focar nenhum domínio específico. Sendo assim, ele é considerado um sistema de segunda geração. O sistema oferece diferentes técnicas de mineração a serem utilizadas, porém o foco deste trabalho é na de mineração de regras de associação. O Tamanduá já vem sendo utilizado por algumas instituições públicas brasileiras, com aplicação 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 para cientistas sociais.

O Tamanduá visa proporcionar serviços de mineração de dados de forma escalável e eficiente e possui como componentes: uma unidade de dados (armazenamento dos dados e metadados), unidade de mineração (executa os algoritmos em si, processando os dados e produzindo os resultados) e a unidade de visualização (que recebe o conjunto de resultados e produz uma representação visual). A arquitetura do Tamanduá pode ser vista na Figura 2(a).

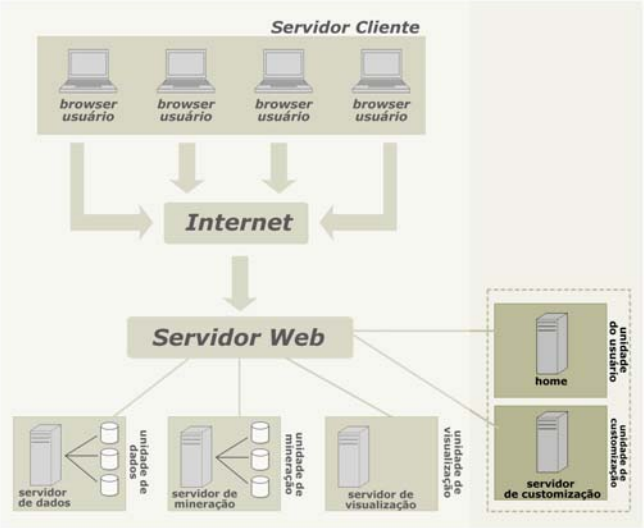
O processamento de uma tarefa de mineração no Tamanduá é realizado em um ambiente de cluster, utilizando técnicas de algoritmos paralelos baseados no paradigma de *Filter Labeled-Stream* [Ferreira et al., 2005]. O processamento em ambiente paralelo de alto desempenho é extremamente importante, vista a complexidade de minerar grandes volumes de dados.

Claramente, o Tamanduá ao permitir a geração de conhecimento a partir de grandes volumes de dados utilizando-se de processamento distribuído apresenta soluções a alguns dos problemas levantados no desafio de gestão de grandes volumes de dados. No entanto, como os demais sistemas de mineração de regras de associação de 2ª. geração ele requer do usuário conhecimento técnico para sua utilização, o que limita o acesso a este conhecimento por usuários não especialistas.

### 4.2 O Protótipo

As mudanças necessárias na arquitetura do Tamanduá para a implementação do modelo estão apresentadas na Figura 2(b). Foram necessárias duas novas unidades: servidor de usuários e servidor de customização. A unidade dos usuários (*home*) armazena os arquivos pessoais de cada usuário, como tarefas e visualizações. Já a unidade de customização possibilita ao usuário especialista a interação com o sistema para criação das extensões. Ele contém os componentes do modelo responsáveis por gerar a camada de abstração dos usuários leigos: o gerador, o interpretador e a base de conhecimento. Além disso, foi necessária a expansão do banco de dados de forma a armazenar as abstrações (consultas) criadas e em relação ao servidor de visualização, mudanças

foram necessárias para acrescentar formas de serem geradas as abstrações para as visualizações dos resultados. Isso para possibilitar que as abstrações textuais fossem criadas a partir do conjunto de regras geradas.




**(a)** **(b)**

**Figura 2. (a) Arquitetura do Tamanduá. (b) Componentes adicionados pelo modelo**

Para ilustrar o funcionamento e uso do protótipo apresentamos um dos cenários propostos para os alunos para o projeto, o do Vestibular. Neste caso, o usuário real – um pesquisador da universidade – gostaria de obter indicadores sobre a qualidade das questões de vestibular de uma determinada instituição. A base de dados existente em relação ao tema contém informações dos anos de 1995 a 2005 e apresenta dados sócio-econômicos, assim como notas nas provas dos vestibulares e em diversas disciplinas na graduação dos candidatos aprovados.

O usuário, considerado leigo nas técnicas de mineração de dados, desejava analisar a validade das questões do vestibular serem fatores preditores de desempenho do aluno na graduação. Para isso, ele gostaria de verificar se o desempenho do aluno em determinada questão do vestibular poderia estar correlacionada com seu desempenho em uma ou mais disciplinas do curso. Foi apresentada então uma classificação das questões, de forma a considerá-la como uma questão boa ou ruim.

Assim, de acordo com a classificação proposta, 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. De acordo com a classificação apresentada, o pesquisador definiu que as boas questões seriam aquelas que conseguem pré-determinar o desempenho do aluno nas disciplinas (ou seja, o erro ou acerto na questão implica em um fracasso ou sucesso na disciplina, respectivamente). Assim, um aluno que acerta uma questão (ou obtém sucesso em uma prova específica do vestibular) de física, por exemplo, e depois tem sucesso na disciplina de Fundamentos de Mecânica, pode demonstrar que a questão foi uma boa forma de seleção. Já o aluno que vai muito bem em uma prova de matemática e depois tem fracasso em várias

disciplinas relacionadas, como as de cálculo, pode demonstrar que a prova não está sendo uma boa seleção dos alunos.

Diante dessa demanda, buscou-se apresentar premissas para serem criadas algumas consultas relacionadas e que poderiam ser úteis para o usuário final, considerando, por exemplo, que "ter sucesso" em uma determinada prova ou disciplina é obter uma nota acima de 70%.

O especialista deveria então criar abstrações para os usuários leigos, de forma que eles pudessem interagir de forma direta com o sistema. No contexto do protótipo, essas abstrações são denominadas "consultas", como a apresentada a seguir:

- *Quais as melhores questões do vestibular de <ANO> da matéria <MATERIA_VESTIBULAR>?"*

Assim, <MATERIA_VESTIBULAR> é um atributo que deve ser definido pelo usuário final. A tela apresentada na Figura 3 apresenta a interface de criação de uma consulta. O especialista pode "arrastar" atributos e medidas de interesse que irão compor a consulta.


**Figura 3. Tela de criação de consulta**

Na visão do usuário leigo, a consulta é apresentada conforme ilustra a Figura 4. É possível então criar inúmeras instâncias de uma determinada consulta, que são denominadas tarefas. Assim, ao realizar a consulta: "*Quais as melhores questões do vestibular de 2005 da matéria física?"* é gerada uma demanda de mineração para o algoritmo, ou seja, uma tarefa de mineração.

De forma a analisar o desempenho dos alunos no vestibular e nas disciplinas, o especialista considerou pares de dados matéria_do_vestibular x disciplina_da_graduação. Essa relação foi considerada por ele coerente, onde alguns exemplos são: Física - Fundamentos de Mecânica; Matemática - Cálculo Diferencial e Integral I; e Química - Química Geral.

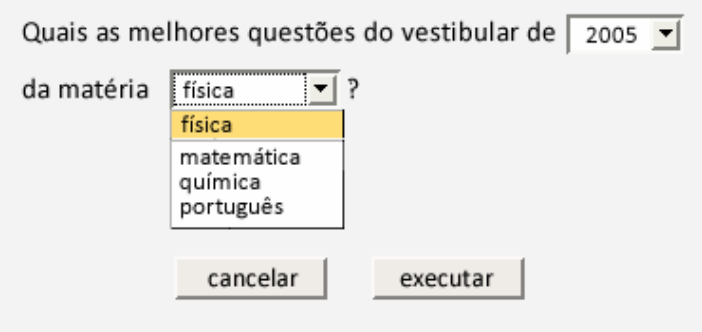



**Figura 4. Visualização da consulta segundo visão do leigo**

A explicação acrescentada pelo especialista para o pesquisador em relação à consulta, apresentada na Figura 3, foi: *Esta tarefa tentará relacionar as questões de uma das provas do vestibular de 2005 com o desempenho no primeiro período letivo dos alunos aprovados. O usuário deve selecionar uma das matérias do vestibular para realizar a consulta. O resultado será dado relacionando as questões da prova escolhida com o desempenho das disciplinas cursadas pelo aluno.* Além dessa explicação, poderia ainda conter a premissa que foi adotada em relação a considerar uma questão boa (o que o especialista considera como as melhores questões?). Assim, poderia ter a explicação sobre a consideração feita: *As relações entre as matérias do vestibular e da graduação foram adotadas da seguinte forma: desempenho na prova de física no vestibular deve ser relacionado com o desempenho na disciplina de Fundamentos de Mecânica, o da prova de matemática com ... Essa relação foi assim feita por serem consideradas disciplinas relacionadas e a relação entre elas ser coerente.*

Um outro tipo de explicação pode ser voltada mais para aspectos técnicos, além da explicação citada acima. Um exemplo seria uma análise em relação à escolha dos valores dos parâmetros: *O valor 0,10 foi atribuído ao suporte por ser considerado um valor relevante visto que .... já para confiança foi atribuído o valor 0,7 pois...*

Existem também outras premissas em relação às abstrações que poderiam ter sido registradas, em relação ao problema de considerar uma questão "boa" utilizando outros critérios, como a explicação a seguir: *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 E para notas até 50% ... 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.

O resultado de uma mineração consiste em um conjunto de regras de associação. É possível que os usuários especialistas criem abstrações textuais para os resultados obtidos. Para isso, o especialista define o *template* do texto a ser apresentado e define que atributos, parâmetros e variáveis das regras de associação serão inseridos no texto, em outras palavras define como será feita a "tradução" das regras geradas. A tela apresentada na Figura 5 ilustra essa tarefa.

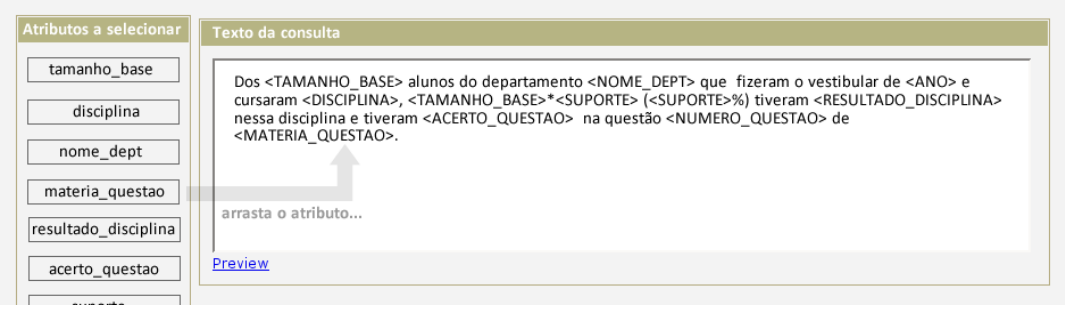



**Figura 5. Tela de configuração textual**

No cenário aqui ilustrado, as regras geradas relacionavam atributos como nome da prova, número da questão, resultado da questão e resultado da disciplina. A forma de apresentação pode variar de acordo com o contexto, abstrações e usuários. Um *template* textual é sugerido aos usuários especialistas pelo protótipo, mas eles podem criar outros. Exemplos de *templates* criados pelos especialistas podem ser vistos a seguir:

- 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 das abstrações textuais criadas, que são apresentadas aos usuários finais, são apresentadas a seguir. A tela que apresenta esses resultados pode ser visualizada na Figura 6.

- *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**.*

Vale ressaltar que a qualidade da abstração criada depende do especialista. Por exemplo, um especialista poderia utilizar termos técnicos que seriam apresentados aos usuários finais, como: "*O suporte dessa relação é...*" Desta forma, a nova interface também dependeria do conhecimento de conceitos técnicos. Assim, o modelo busca auxiliar o especialista para que isso não ocorra, como a proposta do dicionário, mas não tem como oferecer garantias da qualidade final que, em última instância, depende de decisões do especialista.

O protótipo está em fase final de desenvolvimento e todas as interfaces já estão implementadas. Em relação à funcionalidade, o gerador e interpretador já foram desenvolvidos. Como citado, a unidade de dados sofreu algumas expansões no sentido

de comportar o modelo, armazenando as abstrações criadas e geradas. Uma linguagem script (XML) foi criada para definir as abstrações de entrada e saída dos dados e a comunicação entre os componentes do protótipo e do Tamanduá. No momento o protótipo está em fase final de teste, em que estão sendo testados a parte relativa ao armazenamento das regras e de geração da visualização textual a partir destas regras. Avaliações com a participação de usuários reais, tanto especialistas como leigos estão preparadas e os passos iniciais já foram executados (reuniões com leigos e especialistas para definição de consultas de interesse). Assim que for finalizada a implementação do protótipo, os próximos passos da avaliação serão executados (criação da consulta de interesse e modelagem da mineração da tarefa pelo usuário especialista, e uso das consultas pelo usuário leigo).

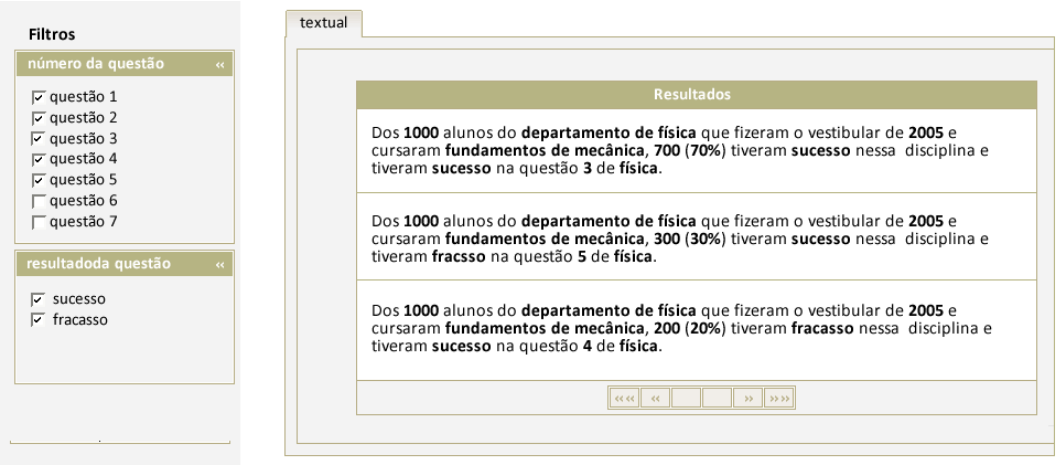



**Figura 6. Tela de visualização textual final**

As modificações necessárias representam o custo apresentado da inclusão do protótipo baseado no modelo no Tamanduá especificamente. Essa análise deve ser feita para cada sistema de mineração de 2ª. geração, uma vez que depende da sua arquitetura e implementação. O modelo tenta ao máximo ser independente do sistema a ser acoplado, mas alterações em relação à comunicação e interação são inevitáveis.

## 5. Conclusões e Próximos Passos

Nesse trabalho foi apresentado um modelo que permite que sistemas possam ser estendidos de forma a criar abstrações a serem executadas de forma direta pelos usuários finais, sem terem que possuir o conhecimento técnico até então necessário. Essa proposta abrange dois desafios, dentre os grandes desafios identificados pela comunidade no encontro promovido pela SBC [SBC, 2006]. Os desafios citados são: o primeiro, relacionado a de gestão da informação em grandes volumes de dados e o quarto de acesso participativo e universal do cidadão brasileiro ao conhecimento.

Como era de se esperar o trabalho não lida com todos os problemas envolvidos nos desafios, mas com algumas das questões levantadas. 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, como vimos, o acesso a estes fica restrito aos detentores do conhecimento técnico envolvido nos algoritmos de mineração de dados.

A disseminação da informação é um dos problemas levantados no desafio de acesso participativo e universal. Assim, o modelo proposto contribui para os esforços nesta direção, à medida que permite que usuários especialistas participem da criação de novas interfaces e geração de conhecimento, e com isso, se amplie o acesso à informação. Vale ressaltar também que o trabalho traz uma contribuição relevante ao ilustrar como problemas abordados por desafios distintos podem ter uma única solução.

Embora a solução proposta seja direcionada a um contexto específico de geração de conhecimento por sistemas de mineração de dados de 2ª. geração, ela potencialmente poderia ser aplicada a outros contextos para possibilitar ou facilitar a disseminação de informação e o acesso a ela por diversos perfis de usuários que compõem a população brasileira. Interfaces extensíveis que possibilitem aos usuários criar diferentes níveis de abstração para a interface e formas de apresentação distintas permitem que se customize o sistema para atender às necessidades de interação de usuários com diferentes níveis educacionais, culturais, tecnológicos e sócio-ecônomicos. O modelo proposto poderia ser adaptado para gerar interfaces extensíveis para outros tipos de sistemas. O primeiro passo nesta direção seria a identificação de sistemas e contextos em que o modelo poderia ser aplicado. A seguir uma investigação mais aprofundada no contexto selecionado e com uma diversidade maior de usuários seria interessante para se coletar indicadores sobre a aplicabilidade da solução apresentada neste trabalho.

A mais curto prazo, os próximos passos no andamento da pesquisa envolvem finalizar uma versão robusta do protótipo, para então executar uma avaliação mais aprofundada que nos permitiria obter indicadores tanto sobre o protótipo desenvolvido, quanto sobre o modelo em que está fundamentado. A próxima etapa da avaliação prevê a utilização do sistema em diferentes domínios de aplicação, tanto por usuários especialistas, quanto leigos. A etapa seguinte da pesquisa será estender o protótipo para permitir que se crie novas interfaces para outras técnicas de mineração de dados, além de regras de associação. A próxima técnica a ser considerada será a de classificação.

## 6. Agradecimentos

Este trabalho foi parcialmente financiado pelo CNPq, Capes, Finep e Fapemig.

## 7. Referências

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

[Albergaria et al. 2006] Albergaria, E., Prates, R., Almir, F., Rocha, L. and Meira Jr., W. 2006. Caracterizando desafios de interação com sistemas de mineração de regras de associação. Anais do IHC 2006). SBC. 40-49.

[Albergaria et al. 2008] Albergaria, E., Mourão, F. H., Prates, R., Meira Jr., W. (2008). A Meta-Design Model to Augment Usability of Rule Association Mining Systems. Proceedings of IFIP Human-Computer Interaction Symposium (HCIS 2008). A ser publicado em setembro.

[Carroll, 2000] Carroll, J. M. *Making Use: Scenario-Based Design of Human-Computer Interaction.* MIT Press, 2000.

[DBMiner] DBMiner Tutorial- Última visita: Maio/2008 - Disponível em: http://www.cs.sfu.ca/CC/459/han/tutorial/tutorial.html#association

[de Souza 2005] de Souza, C. S. The semiotic engineering of HCI. MIT Press, Cambridge, 2005.

[Fayyad et al.] 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*, pp1-34.

[Ferreira et al., 2005] Ferreira, R., W. Meira Jr., W., Guedes Neto, D., Drummond, L. Coutinho, B., Teodoro, G. Tavares, T., Araújo, R., Ferreira, G.. Anthill: A Scalable Run-Time Environment for Data Mining. Applications. Proc. of the Symp. on Computer Architecture and High Performance Computing – SBAC. IEEE, 2005.

[Guedes Neto et al., 2006] Guedes Neto, D., Meira Jr., W., Ferreira, R. A. C. Anteater: A Service-Oriented Architecture for High-Performance Data Mining. IEEE Internet computing, 10, 2006, 36-43.

[Hipp et al. 2000] Hipp, J.,Güntzer, U., Nakhaeizadeh, G. (2000) Algorithms for Association Rule Mining - A General Survey and Comparison. ACM SIGKDD Explorations Newsletter, 2, 1 58-64.

[Hofmann et al. 2000] Hofmann, H.; Siebes, A..; Wilhelm, A. "Visualizing Association Rules with Interactive Mosaic Plots". Proc. ACM SIGKDD '00, pp. 227- 235, 2000.

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

[Kriegel et al. 2007] Kriegel, H., Borgwardt, K., Kröger, P., Pryakhin, A., Schubert, M., Zimek, A. (2007) Future trends in data mining. Data Min. Knowl. Disc. 15,1,87-97.

[Kuranov et al. 2001] 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.

[Mei et al. 2006] Mei, Q.,Xin, D., Cheng, H. , Han, J., Zhai, C. (2006). Generating Semantic Annotations for Frequent Patterns with Context Analysis. Proceedings of KDD 2006, 337-346.

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

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

[SBC 2006] SBC – Relatório Preliminar dos Grandes Desafios da Pesquisa em Computação no Brasil – 2006 – 2016. Última visita: Maio/2008. Disponível em: http://www.ic.unicamp.br/~cmbm/desafios_SBC/

[Tamanduá] http://bandeira.speed.dcc.ufmg.br:8180/tamandua/ Última visita Abr/2008.

[Weka] Weka,.http://www.cs.waikato.ac.nz/ml/weka/. Accessed in February 2008.