Combinando Técnicas de *Data Warehousing* e Mineração de Dados em Avaliações Imobiliárias

Cláudia Sander
Universidade Federal do Rio Grande do Sul, Instituto de Informática,
Porto Alegre, Brasil, 91501-970
sander@inf.ufrgs.br

e

Ana L. C. Bazzan
Universidade Federal do Rio Grande do Sul, Instituto de Informática,
Porto Alegre, Brasil, 91501-970
bazzan@inf.ufrgs.br

Abstract

Real estate price evaluation intends to improve property tax administration and to achieve taxation equity. Huge volumes of real estate data are stored in municipalities, which could provide agility and precision of evaluation, if consolidated in effective data warehouses and having efficient data mining technique applied over them. This work aims to evaluate the integration possibilities of those two areas, through the implementation of a practical application. Such application consists in projecting and building a data warehouse and posterior application of a data mining tool over its data. The application domain that supports this application regards to real estate values estimation.

Keywords: Artificial intelligence, data mining, data warehousing, real estate values estimation.

Resumo

A área de avaliações imobiliárias pretende melhorar a administração de imposto sobre a propriedade e promover equidade fiscal. Enormes volumes de dados relativos a imóveis são armazenados pelas prefeituras, os quais podem prover agilidade e precisão nas avaliações, se consolidados em *data warehouses* efetivos sobre os quais sejam aplicadas técnicas eficientes de mineração de dados. Este trabalho analisa a possibilidade de integração entre estas áreas através de uma aplicação prática, que consiste no projeto e construção de um *data mart*e posterior aplicação de ferramenta de mineração de dados sobre ele, no domínio de avaliações de valores imobiliários.

Palavras chaves: Inteligência artificial, mineração de dados, *data warehousing*, avaliações imobiliárias.

1 Introdução

A constante evolução das tecnologias de armazenamento e recuperação de informação, aliada a disseminação do uso de sistemas de informações, deixa como legado enormes bases de dados nos mais variados domínios de aplicação e conhecimento. O processo de análise destes dados vem evoluindo através dos últimos anos, especialmente com as técnicas de *data warehousing*, que permitem ao usuário a análise dos dados em diferentes níveis de sumarização, e com as técnicas de mineração de dados, que investigam a informação contida em bases de dados para identificar relacionamentos significativos entre os dados.

De acordo com [18] o uso crescente de *data warehouses* é um fenômeno reconhecido, mas a experiência de aplicações de mineração de dados sobre estas bases ainda é limitada. A utilização integrada das metodologias de *data warehousing* e de mineração de dados ainda é incipiente, mas pode proporcionar documentação e metodologias consistentes para a etapa de pré-processamento da mineração de dados, bem como prover ferramentas que auxiliem na etapa de pós-processamento da mineração de dados.

Algumas questões são colocadas por [14, 8, 13], no que diz respeito às possibilidades de integração entre estas técnicas: quais requisitos devem ser seguidos ainda na etapa de projeto de um *data warehouse* para que o mesmo possa prover uma base coerente para mineração de dados efetiva? É possível definir estratégias de mineração de dados que auxiliem a encontrar melhores respostas independente do domínio? As metodologias de projeto, modelagem e documentação de *data warehouses* são eficazes no projeto e modelagem de sistemas de descoberta de conhecimento em bases de dados?

A utilização integrada destas metodologias pode reduzir a distância semântica existente entre engenheiros de mineração de dados e usuários finais, apontada por [8] como um fator que pode prejudicar o sucesso e eficácia de projetos de mineração de dados.

O objetivo deste trabalho é avaliar as possibilidades e vantagens de integração entre as tecnologias de *data warehousing* e mineração de dados, revendo as questões colocadas acima. Esta avaliação é feita a partir do desenvolvimento de uma aplicação na área de avaliações imobiliárias, junto à Secretaria Municipal da Fazenda de Porto Alegre (SMF) e à Companhia de Processamento de Dados do Município de Porto Alegre.

Através deste estudo de caso pode-se avaliar a possibilidade de estabelecimento de uma metodologia integrada de descoberta de conhecimento e soluções gerenciais, de forma a melhorar o desempenho da mineração de dados, independente do domínio. Este melhor desempenho pode ser obtido nas etapas de pré e pós-processamento da mineração de dados. A utilização das metodologias de projeto e modelagem de *data warehouse* e o estabelecimento de estratégias de classificação de atributos e modelagem de dimensões podem possibilitar a geração de uma base de dados adequada para mineração, de forma a consistir na etapa de pré-processamento da mineração de dados. As ferramentas disponíveis para análise dos dados do *data warehouse* podem auxiliar a etapa de pós-processamento, proporcionando análises detalhadas dos casos considerados pela mineração de dados.

O desenvolvimento deste trabalho consistiu em: estudo das técnicas de mineração de dados e *data warehousing* e avaliação de ferramentas disponíveis; estudo sobre a área de avaliações imobiliárias; projeto, desenvolvimento e implantação de um *data mart* com informações imobiliárias; seleção e aplicação de ferramenta de mineração de dados sobre o *data mart*; análise dos resultados e avaliação sobre a utilização integrada destas técnicas.

Este trabalho está organizado da seguinte forma: a próxima seção discute os conceitos básicos das tecnologia envolvidas; a Seção 3 apresenta uma descrição sobre o caso e a aplicação desenvolvida; a Seção 4 discute os resultados e a Seção 5 apresenta as conclusões.

2 Conceitos Básicos

2.1 *Data Warehousing*

Data warehousing é o processo de consolidação de dados esparsos e históricos em uma base de dados centralizada e consistente, que disponibiliza ao usuário ferramentas de análise e visualização destas informações [6, 11, 14]. Esta tecnologia é utilizada especialmente a nível gerencial, no apoio a tomada de decisões, por disponibilizar ferramentas de consultas analíticas sobre dados históricos.

É chamada de *data warehouse* a base de dados centralizada com todos os dados da instituição. São chamadas *data mart*s bases de dados menores, que contêm dados centralizados referentes a cada setor da instituição.

Os dados armazenados em um *data warehouse* devem ser consolidados, consistentes, orientados ao assunto, históricos e não-atualizáveis.

O processo de *data warehousing* pode ser descrito nas seguintes etapas: com base nas estruturas existentes de armazenamento dos dados operacionais é feito o projeto do *data warehouse*; rotinas de extração, transformação e carga são definidas, desenvolvidas e executadas para carregar os dados consolidados no *data warehouse*; os dados consolidados são então consultados e analisados pelos usuários de negócio.

A metodologia de projeto de bancos de dados utilizada para projetar um *data warehouse* é a modelagem dimensional, focada nos processos e questões do negócio. Esta modelagem foi introduzida por Kimball em [11]. A evolução da modelagem dimensional levou à padronização do uso do modelo denominado esquema estrela (*star schema*), que propicia tempo de resposta pequeno às consultas, através da desnormalização e particionamento dos dados.

O esquema estrela refere-se a um modelo dimensional construído em uma configuração de estrela. Este esquema tipicamente contém uma tabela no centro do modelo, denominada tabela fato, com tabelas satélites menores ao redor, denominadas tabelas dimensão, conectadas à tabela fato em um padrão radial.

A tabela fato contém dados factuais, que podem auxiliar a responder as questões do negócio. As informações nela contida residem em diversas tabelas no banco de dados relacional e a mesma pode conter milhões de linhas, dependendo do volume de informação existente e da estratégia de carga destes dados.

As tabelas dimensão são tabelas menores dentro do modelo dimensional, que contêm informação descritiva sobre o negócio. Estas tabelas permitem que o administrador navegue e se aprofunde rapidamente nas informações contidas na tabela fato.

As tabelas dimensão e fato são relacionadas entre si por relacionamentos identificadores, de forma que a chave primária das tabelas dimensão migram para a chave primária da tabela fato, como chaves estrangeiras.

O esquema estrela é extremamente adequado ao projeto de *data warehouse*, pois:

- cria uma base de dados desnormalizada que pode prover respostas rápidas às consultas;
- provê um projeto flexível que pode ser facilmente modificado ou incrementado através do ciclo de desenvolvimento, à medida em que a base de dados cresce;
- reflete a forma como normalmente os usuários pensam e usam os dados;
- reduz a complexidade para desenvolvedores e usuários finais.

Existem outras técnicas de modelagem de *data warehouses*, como *snowflake* e *star-cluster*, as quais podem ser vistas em detalhes em [10].

Cabe ainda ressaltar que antes de iniciar o projeto de um *data warehouse* é fundamental compreender e analisar o negócio, identificando as questões existentes. Dentre estas questões, deve ser identificada uma questão central pela qual será iniciada a modelagem, a qual auxiliará a identificar os dados a serem armazenados nas tabelas fato e dimensão.

2.2 Mineração de Dados

Existem divergências sobre as diferenças entre o significado de mineração de dados e Descoberta de Conhecimento em Bases de Dados (DCBD), termos que são freqüentemente utilizados como sinônimo por alguns autores [2].

Na “First International Conference in Knowledge Discovery in Databases”, realizada em Montreal em 1995, foi proposto que o termo ‘descoberta do conhecimento’ fosse utilizado como o processo completo de extração de conhecimento implícito a partir de bases de dados, representando-o de forma acessível para o usuário, e que o termo ‘mineração de dados’ fosse usado exclusivamente para o estágio de descoberta, no processo de DCBD.

No entanto, de acordo com [18], a mineração de dados não é somente uma técnica utilizada durante o processo de DCBD, uma vez que os problemas resolvidos através da mineração de dados podem ser agrupados em duas categorias gerais, quais sejam: predição e descoberta do conhecimento. A mineração de dados preditiva trabalha sobre fatos já ocorridos, com o objetivo de projetar novos casos de acordo com objetivos específicos. Já a descoberta de conhecimento é utilizada em casos onde não se tem ainda informação suficiente para predição. Sob este ponto de vista, a mineração de dados não seria apenas uma técnica utilizada em uma etapa da DCBD, mas ela mesma um processo completo, dividido nas mesmas etapas básicas da DCBD, pré-processamento, mineração e pós-processamento [18]. Neste trabalho o termo ‘mineração de dados’ é utilizado como o processo completo de identificação de tendências e padrões, baseado em fatos ocorridos.

O processo de mineração de dados é orientado à aplicação, iterativo, interativo e não-linear, onde cada etapa serve de feedback para as demais, como representado graficamente na figura 1.

O objetivo da etapa de pré-processamento é disponibilizar para a etapa seguinte uma base de dados consolidada, consistente e coerente. Freqüentemente, ao iniciar o estudo para um projeto de mineração sobre determinada base de

dados, detecta-se a necessidade ou possibilidade de enriquecer esta base com informações provenientes de outras fontes. Portanto, uma das tarefas da etapa de pré-processamento é identificar as fontes de dados disponíveis, verificar quais informações são efetivamente relevantes para a descoberta de conhecimento, definir e desenvolver rotinas que acessem e integrem estas informações.

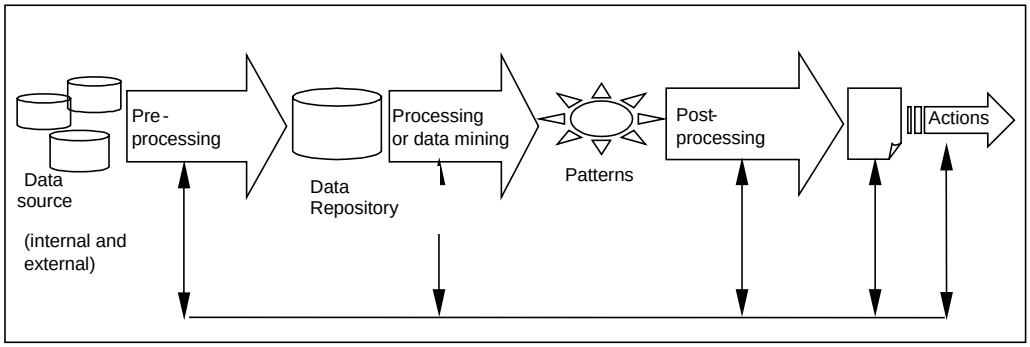


Figura 1 O processo de mineração de dados

Atualmente a maior parte das informações está armazenada em bases de dados relacionais, em tabelas normalizadas, inadequadas para aplicação de rotinas de descoberta do conhecimento. Portanto, na fase de pré-processamento deve ser projetada a forma de armazenamento das informações relevantes. Estas informações devem ser armazenadas de forma desnormalizada, para que as pesquisas executadas sobre a base sejam mais rápidas e eficientes.

Ainda nesta etapa é necessário efetuar transformações sobre os dados, de forma a permitir que sejam estabelecidos relacionamentos entre eles. Por exemplo, uma data de nascimento será mais útil se for transformada em idade ou faixa etária; datas de início e fim de um período poderão gerar descobertas mais proveitosas se forem transformadas em número de dias; valores poderão ser melhor comparados se forem transformados em faixas; etc.

A maior parte das informações disponíveis é proveniente de sistemas de informação, estando sujeitas a erros. É necessário, portanto, efetuar limpezas sobre a base de dados, eliminando registros com informações incompletas ou incoerentes, para que estes não afetem o desempenho e o resultado das rotinas de descoberta do conhecimento. Por exemplo, ao trabalhar sobre dados de um cadastro de pacientes de um hospital, devem ser desprezados registros de pessoas com data de nascimento de 1800, ou atendimentos para os quais não tenha sido registrado nenhum diagnóstico ou tratamento.

Uma vez que os dados estejam prontos, na etapa de processamento são aplicados algoritmos de mineração sobre os dados, com o intuito de descobrir relacionamentos não-explícitos entre estas informações.

A mineração de dados geralmente analisa os dados em busca de quatro tipos básicos de relacionamentos entre os dados pesquisados [2]:

- Classes: utiliza dados previamente classificados para classificar novos dados em grupos predeterminados.
- Clusters: os dados são agrupados de acordo com similaridades entre si.
- Associações: os dados são minerados para encontrar associações entre eles.
- Padrões seqüenciais: os dados são minerados para antecipar padrões de comportamento e tendências.

Existem diversas técnicas de mineração de dados consagradas para realização de cada uma destas análises [5]. Não é possível estabelecer qual a melhor técnica ou algoritmo de mineração de dados, uma vez que a eficiência de cada um deles varia de acordo com diferentes fatores como, por exemplo, o conjunto de dados disponíveis, o tipo de descoberta que se deseja, a natureza dos dados da aplicação, etc [12]. Uma das tarefas em um projeto de mineração de dados é a escolha e definição do algoritmo mais adequado, bem como o ajuste de parâmetros e atributos para obtenção de resultados efetivos. A escolha do algoritmo mais adequado deve levar em conta os objetivos que se pretende atingir com a mineração de dados e a natureza da aplicação.

Finalmente, na etapa de pós-processamento, os padrões identificados na etapa anterior são apresentados ao usuário final. Esta apresentação deve ser feita em uma forma que possa ser interpretada facilmente pelo especialista, como regras associativas, relatórios, gráficos, simulações, planilhas, entre outras. Estes padrões devem ser transformados em conhecimento que possa ser utilizado na definição de ações que melhorem os resultados do negócio.

2.3 Integração

As metodologias apresentadas trabalham sobre grandes bases de dados, com o intuito de permitir aos especialistas do negócio tomarem decisões embasadas no perfil histórico das informações armazenadas.

As técnicas de *data warehousing* permitem consultas administrativas e gerenciais sobre dados históricos, de difícil resolução em um ambiente *on-line* de atualização de dados. As técnicas de mineração de dados permitem descobrir relacionamentos ocultos entre as informações armazenadas, dificilmente encontrados pelo especialista, ao menos de forma tão rápida e automatizada.

Desta forma, pode ser colocada a seguinte questão: qual tecnologia usar em cada situação?

Em [2] os autores afirmam que ferramentas de consultas sobre bases de dados e *data warehousing* devem ser usadas quando se tem claro quais informações se deseja obter, ao passo que a mineração de dados deve ser aplicada quando se sabe vagamente o que está sendo buscado. O crescente interesse na mineração de dados deve-se ao fato de que atualmente existem mais situações nas quais a idéia do que se busca é vaga, com questões dificilmente solucionáveis mesmo com o uso de *data warehouses* [2]. A mineração de dados pode ser utilizada para descobrir hipóteses, que serão confirmadas ou refutadas através da utilização de ferramentas OLAP para análise de dados históricos armazenados em *data warehouses* [7, 9].

Atualmente estão bem estabelecidas metodologias de análise, projeto e modelagem de *data warehouses*, ao passo que não existem ainda metodologias estabelecidas para a mineração de dados. Pode-se então integrar estas áreas, utilizando metodologias de *data warehousing* para auxiliar a análise e projeto de bases para descoberta de conhecimento. Esta discussão é proposta em [9], onde ressalta-se que um dos elos fracos no processo de DCBD é a distância semântica entre os engenheiros de mineração de dados e o usuário final.

Pode-se perceber a similaridade entre as etapas de pré e pós-processamento nestas duas tecnologias. No entanto é preciso tomar cuidado para não acreditar erroneamente que, tendo um *data warehouse* em uma instituição, pode-se simplesmente aplicar sobre ele técnicas de mineração de dados. A existência de um *data warehouse* não pressupõe que a etapa de pré-processamento esteja pronta, uma vez que o mesmo não tenha sido projetado especificamente para este fim [2]. É preciso analisar as informações nele contidas, verificar a necessidade de transformações dos dados armazenados, bem como a necessidade de ampliar as informações contidas na base de dados, pois informações relevantes para a mineração de dados podem não ter sido contempladas no projeto de *data warehousing*, por não apresentarem relação evidente entre si.

Apesar destas questões não poderem ser desprezadas, é certo que as metodologias e ferramentas de *data warehousing* podem servir para facilitar as etapas de pré e pós-processamento, bem como sua estrutura pode simplificar a aplicação de algoritmos e ferramentas de mineração de dados.

3 Implementação do Estudo de Caso

O domínio escolhido para desenvolvimento deste trabalho é a área de avaliações imobiliárias. O trabalho implementado utiliza dados referentes ao Município de Porto Alegre, e foi desenvolvido em conjunto com a Secretaria Municipal da Fazenda de Porto Alegre (SMF) e com a Companhia de Processamento de Dados do Município de Porto Alegre (PROCEMPA).

De acordo com Rubens Dantas[DAN 98], a Engenharia de Avaliações é uma especialidade da engenharia que objetiva determinar tecnicamente o valor de um bem, de seus direitos, frutos e custos de reprodução. A avaliação imobiliária objetiva então estimar o valor de mercado de um imóvel, ou seja, o valor provável pelo qual o imóvel seria transacionado no mercado no qual está inserido. A correta estimativa de valores proporciona melhor arrecadação de impostos e mais justiça tributária.

Dentre os diversos métodos que podem ser aplicados na realização de um trabalho avaliatório, o mais utilizado na avaliação imobiliária, e mais facilmente justificável, é o método comparativo de dados de mercado. Este método consiste em estimar o valor de um bem através da comparação com dados de mercado assemelhados quanto a suas características intrínsecas e extrínsecas e pressupõe a existência de uma amostra significativa de dados do mercado, similares ao bem avaliado [DAN 98].

A avaliação de imóveis no Brasil não apresentou maiores avanços desde o advento do emprego de análise de regressão múltipla, que começou a ser utilizada na área no início dos anos 80. Desde então esta técnica vem sendo cada vez mais popular na área de avaliação imobiliária, muito embora possam ser percebidas dificuldades operacionais no seu emprego e eventualmente ocorram distorções nos valores estimados [DEC 98]. Diversas técnicas vêm sendo estudadas atualmente, visando aplicação na área de avaliações imobiliárias. Em [DEC 00] a mineração de dados é sugerida como uma das possibilidades a ser estudada, pressupondo-se que, aplicando técnicas de mineração de dados, possam ser identificados relacionamentos e atributos não considerados atualmente nas análises de regressão múltipla, ou mesmo para validar os atributos atualmente considerados.

Na Secretaria Municipal da Fazenda de Porto Alegre (SMF), dois setores estão diretamente interessados na precisão destas avaliações: o setor de Imposto sobre Transmissão de Bens Imóveis (ITBI) e a Equipe de Avaliações de Imóveis (EAI), responsável pela planta de valores do município. Todos os dados utilizados neste trabalho referem-se ao Município de Porto Alegre.

No setor de ITBI são realizadas estimativas de valores de imóveis, para fins de cobrança de imposto de transmissão. O setore utiliza um sistema de informações para suporte à estimativa de valores de transações imobiliárias e emissão de guias de pagamento do ITBI, apresentado em [LOR 99]. O sistema registra informações fornecidas pelos tabelionatos, referentes às transações imobiliárias ocorridas no Município. Entre estas informações constam: características dos imóveis envolvidos na transação, como área, topografia, superfície do terreno, área construída e transmitida; construções envolvidas na transação, indicando existência de piscina, cobertura, portaria, elevador; e o valor da transação, declarado pelo contribuinte. Os fiscais do ITBI, baseados nas informações cadastradas no sistema e no seu conhecimento do valor de mercado dos imóveis no município, estimam o valor dos imóveis envolvidos, podendo atribuir novo valor, ou aceitar o valor declarado pelo contribuinte. Sobre o maior valor obtido entre o declarado pelo contribuinte e o estimado pelo fiscal do ITBI, são aplicadas as alíquotas para cobrança do ITBI. Atualmente são avaliadas em média 150 transações imobiliárias por dia neste setor.

Na EAI são realizados dois tipos de avaliações de imóveis: avaliação de toda a população imobiliária em um único momento, para fins de cobrança de Imposto Predial e Territorial Urbano (IPTU), na qual a relatividade é importante, pois é necessária uma uniformidade a nível de avaliação; e avaliação individual de imóveis, para fins de compra e venda de imóveis pelo município, pagamento de desapropriações, reclamações judiciais ou de contribuintes que sintam-se lesados pela avaliação de seu imóvel para fins de pagamento de impostos e estimativa de valor de venda de solo criado.

Na EAI é realizado um trabalho de coleta de valores de mercado, a partir de anúncios de imóveis em jornais e em pesquisas de campo de imóveis à venda, e coleta de dados de venda de imóveis a partir de dados de leilões e do sistema ITBI. Os dados coletados são registrados no sistema PV e analisados no sistema REGRE, um sistema de regressões múltiplas utilizado na avaliação dos imóveis. Atualmente são feitas em média 12 avaliações por mês, sem contar as revisões de valor venal do IPTU. A avaliação leva de uma a duas semanas para estar concluída, dependendo da necessidade ou não de complementação da base de casos com casos similares ao avaliado.

Para a implementação do *data mart* foram utilizadas as ferramentas Analysis Server, Analysis Manager e banco de dados MS-SQL Server 7.0 da Microsoft.

As fontes de dados identificadas para este projeto são as seguintes:

- Dados referentes a 95.000 transações imobiliárias ocorridas em Porto Alegre no período de agosto de 1998 a agosto de 2001, armazenados em base de dados relacional MS SQL Server 7.0
- Dados de oferta e comercializações de imóveis coletados no período de 1992 a agosto de 2001 registrados em arquivos .dbf
- Dados referentes às características de infra-estrutura de cada uma das 18.500 faces de quarteirão do Município de Porto Alegre, registrados em arquivos VSAM
- Características de aproximadamente 500.000 imóveis de Porto Alegre, registradas em arquivos VSAM
- Valor venal dos 500.000 imóveis de Porto Alegre gerados por programa Cobol em arquivo texto
- Dados de valorização referentes às 404 regiões homogêneas da cidade, gerados por programa Cobol em arquivo texto
- Dados do Plano Diretor de Desenvolvimento Urbano Ambiental (PDDUA) registrado em base de dados relacional DB2 no OS390.

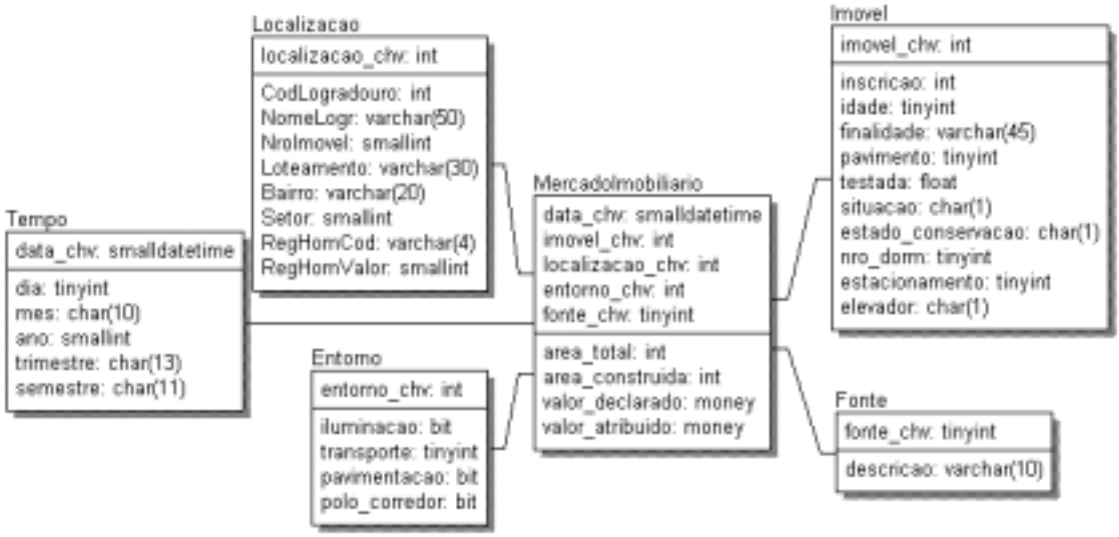
Dados provenientes destas fontes, armazenadas em estruturas de dados tão diversas, foram consolidados em uma base de dados dimensional, de acordo com o modelo apresentado na figura 2.

Foram definidos diversos critérios e transformações a serem aplicados na modelagem e carga dos dados já prevendo a etapa de mineração de dados, de forma que a implementação do *data mart* consista na etapa de pré-processamento da mineração.

Estas informações foram processadas em cubos multidimensionais permitindo análises sobre o mercado imobiliário agrupando e cruzando as informações por região da cidade, tipo de imóvel, faixa de valor, período de ocorrência, características de infra-estrutura, etc.

Sobre esta base de dados foi aplicada uma ferramenta de mineração de dados para identificação de regras que apontem para relacionamentos entre os diversos atributos que influenciem a formação do valor do imóvel.

Figura 2. Modelo dimensional para *data mart* de informações imobiliárias



Após a análise de diversas ferramentas foi selecionada a ferramenta DMS2000, apresentada em [1], que busca regularidades na base de dados, traduzidas em regras na forma se-então. Esta ferramenta implementa diversas técnicas de mineração de dados que se complementam, formando um sistema híbrido de mineração de dados. As técnicas implementadas nesta ferramenta incluem regras de associação, redes neurais, árvores de decisão, clusterização, regressão, análise fatorial e redes bayesianas. De acordo com o tipo de dados e objetivos da mineração, são utilizadas algumas destas técnicas combinadas. Dentre estas, a ferramenta baseia-se principalmente em regras de associação, apresentadas pela primeira vez por [3] e na implementação do algoritmo APRIORI, apresentado por [4], e algumas derivações e otimizações do mesmo.

A aplicação da ferramenta de mineração de dados consistiu na definição e refinamento de modelos de projetos, aplicados sobre a base de dados. O refinamento destes modelos foi determinado pela análise das regras geradas, utilizando as ferramentas de análise de regras disponibilizadas pela ferramenta e comparadas com as análises feitas a partir do cubo gerado pelo *data mart*.

Nos primeiros modelos, que consideravam toda a base de casos e tinham como parâmetro altos fatores de confiança e suporte, eram encontradas poucas regras, por volta de 2 ou 3. Além disso, as regras encontradas eram inexpressivas no contexto de avaliações imobiliárias, por serem demasiado genéricas, como exemplificadas na tabela 1.

Tabela 1. Exemplos de regras obtidas nos primeiros projetos

Regra
Se finalidade = 'Apartamento' então valor do imóvel entre R$ 454,00 e R$3.200.000,00. Suporte = 61,80%. Confiança = 100%.
Se área construída de 22 m^2 a 93 m^2 então valor do imóvel entre R$ 0,01 e R$18.172.078,00. Suporte = 62,84%. Confiança = 100%.
Se área construída < 592,50 m^2 então valor do imóvel < R$181.720,79. Suporte = 96,60%. Confiança = 96,64%.

O refinamento destes projetos incluiu o relaxamento nos parâmetros de confiança para 30% e suporte para 1%. Um fator de suporte tão baixo justifica-se pela natureza dos casos da aplicação, bastante estratificados por bairro e finalidade. Para os atributos valor e área construída, foi utilizada a estratégia de classificação de valores em intervalos equiprováveis, que agrupa os casos em função dos valores assumidos por eles. Mas pela estratificação dos casos em diversas finalidades, os intervalos gerados não ficaram adequados à natureza da aplicação e as regras geradas concentravam-se em apenas uma ou duas classes de valor de imóvel. Usou-se então a classificação informada pelo usuário para os atributos 'área construída' e 'valor de imóvel', refinada a partir da análise das regras geradas, da distribuição de casos em cada classe, e da avaliação do especialista.

A partir destes refinamentos passaram a ser encontradas mais regras. A análise das regras de baixa confiança requer mais tempo do especialista, pois se por um lado elas podem representar características emergentes nos casos em estudo, podem também ser reflexo de dados mal coletados ou falta de dados complementares. Neste sentido, as ferramentas de pesquisa sobre os casos armazenados no *data warehouse* são fundamentais para a análise dos casos

armazenados e validação das regras. Um exemplo é a regra 1 apresentada na tabela 2. Analisando a base de casos observa-se que os imóveis que dão confiança a esta regra são boxes, portanto a regra não representa uma tendência para o mercado em geral, mas uma tendência para o mercado de boxes. Mesmo uma regra com alto fator de confiança necessita de uma análise mais profunda para ser validada ou refutada. Um exemplo disso é a regra 2. Ao avaliar os casos que deram origem a esta regra verifica-se que 99,52% dos casos correspondem a apartamentos. Se o imóvel avaliado for de outras finalidade não necessariamente poderá ser enquadrado nesta regra. Não existem casos registrados no sistema que dêem este indicativo. Mas a partir deste ponto também passaram a ser encontradas regras mais específicas, que podem ser usadas para enquadrar novos casos, como a regra 3 apresentada na tabela 2.

Tabela 2. Exemplos de regras obtidas após refinamento no modelo de mineração de dados

	Regra
1	Se idade = 1 então valor do imóvel até R$ 10.000,00. Suporte = 2,17%. Confiança = 30,19%
2	Se área construída de 100 a 250 m^2 e finalidade = 'Apartamento' então valor do imóvel entre R$ 90.000,00 e R$150.000,00. Suporte = 2,21%. Confiança = 36,73%.
3	Se idade = 17 e Região Homogênea = 'K015' então valor do imóvel até R$10.000,00. Suporte = 1,65%. Confiança = 99,60%.

A análise das regras geradas no modelo anterior ressaltou dois fatores: baixo fator de suporte e falta de especificidade nas regras. Com base nessa análise foi definida outra estratégia de definição de projetos, segmentando os casos por finalidade, ou seja, cada projeto passa a considerar apenas casos referentes a boxes, ou apartamentos, ou salas, etc. Este ajuste foi feito na própria ferramenta de mineração de dados, não sendo necessário refazer a carga original dos dados do *data mart*. A partir desta alteração, regras com maior fator de confiança e mais interessantes no contexto da avaliação imobiliária passaram a ser identificadas. Esta verificação vem de encontro à afirmação de [15] de que na maior parte das vezes não faz sentido analisar todo um enorme *data warehouse*, pois os padrões ficarão perdidos e diluídos. Segundo [15], para encontrar padrões úteis em um *data warehouse* usualmente deve ser selecionado um segmento de dados adequados a determinado objetivo do negócio, prepará-los para análise e então executar a mineração de dados. Esta segmentação é feita de forma a deliberadamente focar a análise no conteúdo de um subconjunto de dados, e não em uma amostra.

O primeiro projeto definido a partir desta nova abordagem considerava o segmento de boxes. A análise dos casos que compõem este segmento, feita através das ferramentas disponibilizadas pelo *data warehouse*, levou à definição de filtros para desconsiderar casos com valores extremos nos atributos área construída e valor do imóvel. Essa análise também apontou a necessidade de redefinir dos intervalos de classificação destes dois atributos. Estes ajustes ao projeto foram aplicados diretamente na ferramenta de mineração de dados, sem interferir nos dados carregados no *data mart*. A partir destes ajustes, passaram a ser encontradas regras bastante específicas e de interesse para o domínio da aplicação, relativas ao segmento de boxes, como as exemplificadas na tabela 3.

Tabela 3. Exemplos de regras obtidas a partir do segmento de boxes

Regra
Se finalidade = 'Box comercial' e região homogênea = 'B041' e idade de 0 a 1 ano, então valor acima de R$ 13.000,00. Suporte = 1,91%. Confiança = 99,13%
Se bairro = 'Bela Vista' e idade de 2 a 6 anos e área construída < 14,50 m^2 então valor de R$ 10.600,00 a R$ 13.000,00. Suporte = 2,45%. Confiança = 86,99%.

Ao aplicar este projeto sobre outros segmentos, ajustando os intervalos de classificação de valores de imóveis e área, foram obtidas regras pouco expressivas no contexto da aplicação, conforme exemplificado na regra 1 apresentada na tabela 4. A análise destas regras aponta para a necessidade de considerar outros atributos para o estabelecimento de padrões. Na aplicação do projeto sobre o segmento de apartamentos foram acrescentados os atributos número de dormitórios e número de vagas na garagem. No entanto, a qualidade dos dados coletados, não permitiu a geração de regras relevantes, uma vez que poucos casos tinham esta informação. Desta forma, o número de casos que davam suporte a regra era muito baixo, não permitindo validar a regra. Por exemplo, a regra 2 apresentada na tabela 4 é suportada por apenas 10 casos. O mesmo problema foi encontrado ao tentar verificar a existência de relacionamento entre os atributos 'pavimento', 'existencia de elevador' e 'valor do imóvel'.

Na etapa de pós-processamento, que consiste na garimpagem de regras relevantes para a aplicação, foram utilizadas as ferramentas disponibilizadas para análises e cruzamentos de informações do *data mart*, para análise comparativa com as regras geradas. Dessa forma é possível que o especialista em avaliações imobiliárias valide as regras encontradas, refutando ou confirmando as hipóteses levantadas, e permitindo que novos casos sejam enquadrados nas regras selecionadas.

Tabela 4. Exemplos de regras obtidas a partir do segmento de apartamentos

	Regra
1	Se área construída > 250m^2 então valor acima de R$200.000,00. Suporte = 2,34%. Confiança = 90,09%.
2	Se elevador = 'Não' e nro_dorm = 2 e Região Homogênea = 'K015', então valor de R$17.000,00 a R$30.000,00. Suporte = 10 casos. Confiança = 100%

4 Resultados

A validação dos resultados obtidos, junto aos especialistas em avaliações imobiliárias, permitiu determinar a validação das estratégias definidas com o objetivo de estabelecer uma mineração de dados efetiva, independente do domínio, integrada a um projeto de *data warehousing*.

As estratégias definidas e validadas neste projeto são apresentadas a seguir e devem ser seguidas desde a etapa inicial de projeto do *data warehousing*:

- Transformação de datas iniciais e períodos em idade e/ou tempo decorrido, com base na data de ocorrência do fato.
- Segmentação da base de casos, de forma a gerar regras a partir de uma base de casos mais homogênea. Esta segmentação deve ser feita apenas na aplicação da ferramenta, e não na carga de dados do *data warehouse*.
- Agrupamento de valores contínuos em faixas, determinadas de acordo com as características de cada segmento. Esta classificação pode ser feita na etapa de mineração de dados, se a ferramenta utilizada oferecer recursos para tal, caso contrário deve ser feita ainda na etapa de carga de dados.
- Análise dos valores assumidos pelas variáveis para eliminar ou ajustar casos com valores fora do contexto. Este ajuste deve ser feito em um primeiro momento na própria carga de dados, eliminando casos com erro, mas posteriormente isto deve ser revisado de forma a aplicar mais alguns filtros na etapa de mineração de dados.
- Avaliar a possibilidade de extração de informações úteis a partir de dados armazenados em formato textual.

A aplicação destas estratégias permitiu que os dados estivessem prontos para a mineração, ao fim da carga do *data warehouse*.

O uso integrado destas técnicas também apresentou ganhos na etapa de pós-processamento da mineração de dados. Estas duas técnicas se complementam, de forma que as ferramentas de visualização disponibilizadas pela aplicação de *data warehousing* auxiliam o especialista a analisar os casos que dão suporte e confiança a uma regra, permitindo comprovar, refutar ou mesmo complementar as hipóteses levantadas pelas regras encontradas.

O desevolvimento deste projeto permitiu a observação de ganhos de produtividade e de qualidade, como segue:

- A implantação do *data mart*, que implementa rotinas de carga automática de dados, aumentou o universo de casos selecionados para compor a base de pesquisa. O número de casos considerados aumentou para 83,33% das ocorrências, contra 5% dos casos considerados anteriormente.
- A seleção e carga automática dos casos reduziu em 70% o tempo gasto na busca de casos e na tarefa de coleta e preenchimento dos dados.
- O tempo gasto pesquisando outros sistemas foi reduzido em 50%.
- A necessidade de consultas a fonts externas, devido a falta de casos na base, tende a reduzir-se a zero..
- O tempo gasto na pesquisa de casos similares foi reduzido em torno de 20%.
- Como a pesquisa ficou mais simples e efetiva, os fiscais passaram a fazer mais pesquisas de casos similares, bem como melhor verificação dos dados informados, melhor qualificando a estimative de valores de imóveis.
- A identificação de atributos e relacionamentos que influenciam a avaliação de imóveis, não computadas nas técnicas utilizadas anteriormente, apontam para a possibilidade de aumentar a precisão da avaliação.

Erros que costumavam ocorrer na verificação de casos de interesse, bem como na transcrição da informação proveniente de diversos sistemas, foram eliminados.

5 Conclusões

Este estudo de caso permitiu a revisão das questões apresentadas na seção 1, relativas à possibilidade de integração das técnicas de *data warehousing* e mineração de dados. A análise desta integração e avaliação dos resultados possibilita caminhar em direção ao estabelecimento de uma metodologia para projetos de mineração de dados, especialmente quando integrados a projetos de *data warehousing*.

O resultado deste projeto consiste em uma poderosa ferramenta de suporte à tomada de decisão, qualificando os avaliadores, promovendo maior precisão e agilidade nas avaliações, baseando-as em justificativas concretas e não apenas baseadas no conhecimento particular de especialistas sobre o mercado imobiliário. Considerando esta motivação, este trabalho propôs uma metodologia para realização destas tarefas.

Como trabalho futuro, sugerimos o desenvolvimento de aplicações similares a aplicação discutida aqui, em outros domínios de problema, permitindo que as estratégias aqui apresentadas sejam comparadas, validadas e ajustadas com o objetivo de estabelecer metodologias de projeto que integrem as duas áreas.

Em relação à aplicação na área de avaliações imobiliárias, sugere-se a complementação dos dados a serem coletados, com a inclusão de novas variáveis, e o preenchimento mais criterioso das informações, enriquecendo o modelo e os casos a serem estudados, de forma que novas análises possam ser realizadas, gerando resultados mais precisos e regras mais específicas. Finalmente, a possibilidade de mineração de dados textuais deve ser melhor explorada, uma vez que muitas informações relevantes sobre as transações imobiliárias são registradas na forma de texto livre.

Em relação à aplicação de mineração de dados observa-se que em alguns casos são geradas muitas regras, que devem ser analisadas e confirmadas antes que seja possível enquadrar novos casos nelas. A ferramenta utilizada apresenta diversos recursos para descartar, filtrar, analisar regras, e implementa técnica para descartar regras sem interesse. Mas outras técnicas e abordagens podem ser utilizadas e implementadas, no sentido de auxiliar ainda mais o especialista nesta tarefa. A aplicação desenvolvida neste trabalho pode servir de base para pesquisas de técnicas que trabalhem neste sentido. O desenvolvimento de ferramentas que auxiliem na determinação de medidas de interesse das regras geradas simplificará sua seleção e utilização no enquadramento de novos casos de avaliação e estimativa de imóveis.

References

[1] Acabit Tecnologia. Acabit Tecnologia - DMS 2000. Available at: <http://www.acabit.com.br/dmsPrincipal.asp>. Accessed in: 10 Aug. 2001.

[2] Adrianns, P; Zantinge, D. Data mining. Edinburgh, UK:Addison-Wesley Longman, 1996.

[3] Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. In: ACM SIGMOD CONFERENCE ON MANAGEMENT OF DATA, 1993. Conference on Management of Data: Proceedings. Washington, USA: 1993. p. 207-216.

[4] Agrawal, R. et al. Fast discovery of association rules. In: Fayyad, U. M. Advances in knowledge discovery and data mining. Menlo Park, USA: AAAI Press, 1996. p. 307-328.

[5] Anand, S. S.; Hughes, J. G. Hybrid data mining systems: the next generation. In: PACIFC-ASIA CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, PAKDD, 2., 1998. Research and development in knowledge discovery and data mining: Proceedings. Berlin: Springer-Verlag, 1998. p. 13-24. (Lecture Notes in Artificial Intelligence, 1394).

[6] DbServer, Assessoria em Sistemas de Informação Ltda. Data warehousing usando SQL Server 7.0. Porto Alegre: 1999.

[7] Edelstein, H.A. Introduction to data mining and knowledge discovery, third edition. [S.l.]:Two Crows, 1999. Disponível em: <http://www.twocrows.com/booklet.htm>. Acesso em: 13 jul. 2000.

[8] Fayyad, U. M., Piatetsky-Shapiro, G., Smith, P. From data mining to knowledge discovery: an overview. In: Fayyad, U. M. et al. Advances in knowledge discovery and data mining. Menlo Park, USA: AAAI Press, 1996. p. 1-34.

[9] Feldens, M. A. et al. Towards a methodology for the discovery of useful knowledge combining data mining, data warehousing and visualization. In: CONFERÊNCIA LATINOAMERICANA DE INFORMÁTICA, 24., 1998, Quito, Equador. Memorias... Quito, Equador: Pontifícia Universidad Catolica del Ecuador, 1998. v. 2, p. 935-947.

[10] Inmon, W. Building the Data Warehouse. John Wiley & Sons, Inc. 1996.

[11] Kimball, R. The Datawarehouse Toolkit. John Wiley & Son, Inc., 1996.

[12] Lavrac, N. Selected techniques for data mining in medicine. In: Artificial Intelligence in Medicine, n.16, p 3-23. 1999.

[13] Microsoft Corporation. Designing and implementing a data warehousing using Microsoft SQL Server 7.0. Argentina: Docuprint S.A., 1999.

[14] Moody, D. Kortnik, M. From Enterprise Models to Dimensionals Models: A Methodology for Data Warehouse and Data Mart Design. DMDW'00, Sweden, 2000.

[15] Parsaye, K. Datamines for data warehouses. Data mining above, beside and within the warehouse to avoid 'The paradox of warehouse patterns'. Information Discovery, 1996. Disponível em: <http://www.dmreview.com/portal_ros.cfm?NavID=92&WhitePaperID=61&PortalID=9>. Acesso em: 30 out. 2001.

[16] Sander, C. Mineração de dados na identificação de padrões e tendências. 2000. Trabalho individual (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

[17] Sander, C. Combinando técnicas de data warehousing e mineração de dados em avaliações imobiliárias. 2001. Dissertação (Mestrado em Ciência da Computação) - Instituto de Informática, Universidade Federal do Rio Grande do Sul, Porto Alegre, RS.

[18] Weiss, S. M., Indurkhya, N. Predictive Data Mining: a Practical Guide, San Mateo, USA: Morgan Kaufmann Publishers, 1998.