Mineração de regras de associação em dados georreferenciados

Marina Abichabki Pivato

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: / /

Assinatura:_____

Mineração de regras de associação em dados georreferenciados

Marina Abichabki Pivato

Orientadora: *Prof^a Dr^a Solange Oliveira Rezende*

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, como parte dos requisitos para obtenção do título de Mestre em Ciências de Computação e Matemática Computacional.

USP - São Carlos
Fevereiro/2006

Este documento foi preparado utilizando-se o formatador de textos $\text{\LaTeX}$. Sua bibliografia é gerada automaticamente pelo $\text{\BibTeX}$, utilizando o estilo Chicago. O formato padrão utilizado foi desenvolvido e cedido por Ronaldo Prati.

Agradecimentos

A Deus por ter me dado condições de construir esse trabalho.

Aos meus pais José Luiz e Salma, minhas irmãs Débora e Paula, meus sobrinhos, Laura e Pedro e toda minha família, pelo amor, carinho e apoio durante o período que me dediquei a este trabalho.

À professora Solange, minha orientadora sem a qual este projeto de pesquisa não seria concluído, pela amizade, apoio, não deixando de acreditar neste trabalho e por todo o conhecimento acadêmico e principalmente, as experiências de vida que me ensinou nesses anos de convivência.

Ao professor Edson, praticamente meu co-orientador, pelas idéias e sugestões, por sua grande ajuda neste projeto de pesquisa, pela estrutura oferecida na sua sala na UFSCAR e pela amizade nesses anos.

À Clarissa, a especialista em engenharia urbana, por toda a ajuda no estudo de caso realizado, toda paciência nos meus atrasos, por fazer os momentos de realização dos experimentos uma atividade divertida, não me deixando desanimar quando as coisas davam errado.

Aos professores, Agma, Alneu, Caetano, Graça Nunes, Maria Carolina e Rosane, pelas conhecimentos passado e sugestões para este trabalho.

Aos meus amigos Camila, Daniel, Fernanda, Giselle, Jaqueline, Magaly, Marco Aurélio, Marcos, Márcio, Roberta, Rodrigo Bianchi, Veronica e Walter, pela convivência, pela ajuda no meu texto e projeto, por tudo que aprendi sobre pesquisa e texto científico, pelos comentários, sugestões, amizade, viagens, jantares e bons momentos juntos.

Aos meus amigos Ana Carolina, Edson, Eduardo, Katti e Ronaldo pela companhia no laboratório, pela amizade e pela paciência em administrar o laboratório oferecendo toda a estrutura disponível.

Aos meus amigos Christiane, Fábio, Fabrício, Ivana, Jorge, Maria Beatriz, Maria Carolina, Marie, Melina, Miquela, Raquel e Sullivan por tudo que vivi e aprendi em nossa convivência, pelas horas alegres, pela diversão, pelo incentivo nos momentos em que pensei em desistir e por entender a minha ausência durante a conclusão dessa pesquisa.

Aos meus amigos Álvaro, Claudia, Emílio, Fábio Antunes, Helena, Maria Gabriela, Patrícia Dontal, Rodrigo e Vanessa pelo apoio e amizade.

À todos do LABIC pelos bons momentos e por fazer do laboratório um ambiente agradável.

À todos os funcionários do ICMC pela atenção, respeito, dedicação e educação.

À CAPES pelo apoio financeiro ao projeto de pesquisa.

À Universidade de São Paulo e Universidade Federal de São Carlos pela estrutura necessária para o desenvolvimento do projeto de pesquisa.

À prefeitura de Jaboticabal por disponibilizar os dados para realizar o estudo de caso apresentado neste trabalho.

Resumo

Sistemas de informações geográficas permitem armazenar, manipular e analisar dados espaciais e aspectos descritivos desses dados. A análise de dados espaciais pode ser realizada por meio de técnicas de extração de regras de associação, ou seja, regras que descrevem relacionamentos entre os dados. Porém, a mineração de regras de associação não considera as relações topológicas existentes entre dados georreferenciados. Para solucionar esse problema, Koperski and Han (1995) e Malerba et al. (2001) propuseram um processo de extração de regras integrado ao algoritmo de mineração e utilizavam predicados lógicos para representar as regras. Como alternativa a essa solução, este trabalho propõe pré-processar os dados georreferenciados para encontrar relações topológicas em separado e aplicar um algoritmo de mineração de regras de associação disponível pela comunidade acadêmica. As regras geradas devem apresentar características descritivas dos dados e relações topológicas. Para atingir esse objetivo foi especificado um processo de extração de regras em dados georreferenciados e implementado um módulo de pré-processamento que extrai relações topológicas. O módulo foi avaliado por meio de um estudo de caso utilizando o sistema de informação geográfica da cidade de Jaboticabal, no contexto de planejamento urbano. As regras encontradas foram analisadas por um especialista utilizando as medidas de suporte e confiança. Além disso, uma análise sobre o tempo de processamento e consumo de memória para encontrar as relações topológicas foi realizada, mostrando que é possível extrair padrões utilizando o processo e o módulo proposto neste trabalho.

Abstract

Geographic information systems are used to store, manipulate, and analyze spatial data and its descriptive aspects. Spatial data analysis can be done by searching association rules that describe relationships between the data. However, georeferenced data present topological relations unknown to traditional mining association rule algorithms. To solve this problem, Koperski and Han (1995) and Malerba et al. (2001) proposed a topological relation extraction process integrated to a mining association rule algorithm. This process requires all data to be translated as logical predicates. As an alternative to this solution, this work proposes to break down this process by pre-processing the georeferenced data to find topological relations, then executing traditional mining association rule algorithms. The resulting rules must present descriptive characteristics of the data and topological relations. To reach this objective, a process of rule extraction in georeferenced data was specified, in addition to a pre-processing module implementation. This module was evaluated by using a case study that uses a geographic information system of the city of Jaboticabal, in the context of urban planning. The generated rules were analyzed by a specialist using the measures of support and confidence. In addition, an analysis regarding the processing time and memory consumption was provided to find the topological relations, which shows that it is possible to extract the patterns with the proposed process and module.

Sumário

1	Introdução	1
2	Sistema de informação geográfica	5
2.1	Objetos espaciais	5
2.2	Operações espaciais e relações topológicas	11
2.3	Sistemas de informações geográficas	14
2.3.1	Banco de dados geográficos	17
2.3.2	Spring	18
2.3.3	TerraLib e TerraView	19
3	Mineração de dados	23
3.1	Descoberta de conhecimento	23
3.1.1	Pré-processamento	25
3.1.2	Extração de padrões	26
3.1.3	Pós-processamento	27
3.2	Associação	28
3.2.1	Regra de associação	28
3.2.2	O algoritmo <i>Apriori</i>	31
3.3	Regras de associação espacial	34
3.3.1	Algoritmo para extração de regras de associação espacial	35
3.3.2	Exemplo de mineração de regras de associação espacial	38
3.3.3	Estudos relacionados a regras de associação espacial	39
4	Extração de regras de associação em dados georreferenciados	41
4.1	Objetivo	42
4.2	Processo de extração de regras de associação em dados georreferenciados . .	43
4.3	Módulo de extração de relações topológicas	45
4.3.1	Regras de associação espacial	48

4.3.2	Regras de associação em espaço limitado	50
4.3.3	Regras de associação espacial utilizando lógica relacional	51
4.4	Extração de regras de associação com informação espacial	52
4.5	Pós-processamento das regras	53
4.6	Considerações finais	54
5	Estudo de caso	57
5.1	Sistema de informação geográfica da cidade Jaboticabal	57
5.2	Modelo do banco de dados	58
5.3	Identificação do problema	61
5.4	Pré-processamento	63
5.4.1	Preparação dos dados para mineração de regras de associação	63
5.4.2	Extração de relações topológicas	69
5.5	Extração de padrões	70
5.6	Pós-processamento	71
5.6.1	Seleção de regras	72
5.6.2	Regras avaliadas pelo especialista	73
5.7	Avaliação da metodologia e do módulo de pré-processamento	77
6	Conclusão	83
6.1	Principais contribuições	84
6.2	Trabalhos futuros	85
	Referências	90

Lista de Figuras

2.1	Separação em universos na modelagem de dados geográficos	6
2.2	Representação geométrica de objetos espaciais	10
2.3	Exemplos de relações topológicas	15
2.4	Sistemas de informações geográficas	16
2.5	Consulta espacial	18
2.6	Modelo conceitual adaptado do Spring	20
2.7	Exemplo de função do TerraLib	21
3.1	Etapas do processo de extração de conhecimento	24
3.2	Tipos de atividades de mineração de dados	27
3.3	Dados de entrada para o processo de extração de regras de associação.	29
3.4	Combinações possíveis dos itens comprados.	30
3.5	Etapas para geração de regra de associação espacial	35
3.6	Consulta de entrada do algoritmo de mineração de regras de associação espacial	36
3.7	Descrição das tabelas sobre British Columbia	38
4.1	Adaptação do processo de mineração de dados para mineração de dados ge- orreferenciados	44
4.2	Módulo de extração de regras de associação espacial	46
4.3	Planos de informação e mapa de objetos	49
4.4	Exemplo do formato de saída do módulo de extração de relações topológi- cas para gerar regras de associação espacial	49
4.5	Exemplo do formato de saída do módulo de extração de relações topológi- cas para gerar regra de associação em espaço limitado	50
4.6	Exemplo do formato de saída do módulo de extração de relações topológi- cas para viabilizar a extração de regras de associação espacial utilizando lógica relacional	51

5.1	Modelo simplificado do banco de dados espaciais da cidade de Jaboticabal	60
5.2	Modelo das tabelas de IPTU e lote	65
5.3	Modelo das tabelas de lotes de escolas	65
5.4	Modelo das tabelas de lotes que disponibilizam serviços	66
5.5	Modelo das tabelas do IBGE	67
5.6	Modelo com tabelas pré-processadas finais	68
5.7	Regras de associação espacial encontradas utilizando lotes e escolas	73
5.8	Regras de associação com informação espacial encontradas utilizando lotes e escolas	74
5.9	Regra de associação com informação espacial encontrada utilizando lotes e IBGE	74
5.10	Regras de associação com informação espacial encontradas utilizando lotes e rodovias	75
5.11	Regras de associação espacial encontradas utilizando lotes e setores	75
5.12	Regras de associação espacial encontradas utilizando serviços e escolas	75
5.13	Regra de associação com informação espacial encontrada utilizando serviços e IBGE	76
5.14	Gráficos sobre tempo de processamento	79
5.15	Gráficos sobre uso de memória	80
5.16	Gráficos sobre tamanho do arquivo gerado	81

Lista de Tabelas

2.1	Relação entre tipos de dados geográficos e o modelo de representação	10
3.1	Conjunto de dados no formato atributo-valor	25
3.2	Representação das transações e itens comprados.	30
3.3	<i>Itemsets</i> freqüentes e seus respectivos suportes.	31
3.4	Exemplos de relações de proximidade entre objetos	39
3.5	Relacionamento espacial detalhado para os predicados freqüentes	39
5.1	Principais tabelas do banco de dados de Jaboticabal	59
5.2	Aspectos procurados na extração de padrões	62
5.3	Relação entre os aspectos procurados e as tabelas do banco de dados de Jaboticabal	63
5.4	Experimentos de extração de relações topológicas	70
5.5	Experimentos de extração de regras de associação	71
5.6	Arquivos e regras de associação após aplicação do filtro	72

Lista de Algoritmos

3.1	Apriori	32
3.2	Função apriori-gen	33
3.3	Geração de regras de associação	34
4.1	Algoritmo para extração de relações topológicas.	47

Introdução

O emprego da computação no armazenamento e processamento de imagens e dados geográficos gerou uma nova expectativa: explorar conhecimentos e relações que estejam contidos nesses dados e então aproveitá-los para auxiliar no processo de tomada de decisão. O conhecimento encontrado pode facilitar a compreensão das influências do ambiente na vida do ser humano e por exemplo, melhorar o planejamento do crescimento para determinada região. Além disso, é importante extrair a maior quantidade possível de informações e conhecimento, a partir dos dados disponíveis pois a aquisição desses dados, normalmente, possui um alto custo.

Para armazenar, manipular e analisar dados geográficos existem sistemas computacionais denominados sistemas de informações geográficas (SIG). Os dados geográficos são georreferenciados, isto é, são associados a uma posição no globo terrestre em função de coordenadas espaciais. Além dos dados georreferenciados, um sistema de informação geográfica pode conter informações sobre características descritivas a respeito desses dados. No geral, o volume de informações é elevado dificultando a análise do comportamento dos dados georreferenciados, sendo importante analisar, descobrindo padrões e tendências para facilitar a compreensão e solução de problemas sobre as regiões pesquisadas.

Diante da deficiência para analisar e compreender grande volume de dados, diversos estudos têm sido direcionados ao desenvolvimento de tecnologias de extração automática de conhecimento de banco de dados. Esse campo de pesquisa é chamado de extração de conhecimento em banco de dados, geralmente referenciado na literatura como *knowledge discovery in database* (KDD), *data mining* (DM) ou mineração de dados (MD). Alguns

autores consideram os termos KDD e MD referentes a processos distintos (Fayyad et al. 1996). Entretanto, neste trabalho, os termos KDD e MD serão tratados indistintamente referenciando o processo de extrair conhecimento a partir de dados.

O processo de extração de conhecimento em banco de dados tem o objetivo de encontrar conhecimento a partir de um conjunto de dados para ser utilizado em um processo decisório. Portanto, um requisito importante é que esse conhecimento descoberto seja compreensível por humanos, além de útil e interessante para os usuários finais do processo, de forma que forneça suporte aos usuários na tomada de decisão (Fayyad et al. 1996). A mineração de dados é uma área multidisciplinar que incorpora técnicas utilizadas em diversas áreas como banco de dados, inteligência artificial e estatística. Assim, as técnicas utilizadas em MD não devem ser vistas como substitutas de outras formas de análises (por exemplo, OLAP), mas como práticas para melhorar os resultados das explorações feitas com as ferramentas atualmente usadas.

Entre as técnicas disponíveis para mineração está a utilização de Associação. Uma regra de associação representa o quanto a presença de um conjunto de itens implica a presença de outro conjunto de itens (Agrawal et al. 1996). A utilização de regras de associação vem sendo intensamente pesquisada e vários algoritmos têm sido propostos para a geração de regras, como o algoritmo *Apriori* (Agrawal et al. 1996). Contudo, nem sempre é possível aplicar esses algoritmos diretamente sobre os dados disponíveis, devido a particularidades existentes nesses dados, como é o caso da aplicação em dados espaciais.

Dados espaciais, também conhecidos como dados georreferenciados, representam informações sobre a localização no espaço, como latitude e longitude, caracterizando relações topológicas, como inclusão, adjacência e proximidade. Porém, os algoritmos tradicionais de extração de regras não são capazes de gerar esses relacionamentos, uma vez que esses dados apresentam valores de ponto, linhas e polígonos, definidos em um espaço de coordenadas que por si só não traz o significado topológico. O desafio de trabalhar essas relações no processo de mineração de dados começou a ser estudado há alguns anos, considerando as técnicas de *Clustering*, Classificação, Associação e Detecção de padrões (Ester et al. 2001).

No caso de extração de regras de associação, o trabalho pioneiro de Koperski and Han (1995) apresenta um algoritmo para extrair regras espaciais, baseado no algoritmo *Apriori*, utilizando um processo de refinamento progressivo, com aproximações para determinar regiões próximas. Após o refinamento é realizada a extração de padrões. Malerba et al. (2001) utilizaram programação lógica indutiva (ILP, do inglês *Inductive Logic Programming*) para extrair regras de associação espacial. Nesses trabalhos a grande diferença entre a mineração de dados tradicional e a espacial é a extração de relações topológicas, sendo necessário conhecer as estruturas de dados específicas para dados espaciais e mecanismos de extração de relações topológicas. Porém, os dados utilizados estão armazenados em bancos de dados relacionais, não estando no formato de entrada dos algoritmos propostos.

Para extrair as regras, é necessário converter os dados do banco de dados em predicados ou cláusulas lógicas. Além disso, as relações topológicas não são encontradas diretamente. Primeiro é extraída a relação de proximidade. A partir dessa relação todas as outras relações topológicas são pesquisadas. Essa divisão foi proposta por Koperski and Han (1995) para melhorar o tempo de processamento na obtenção de relações topológicas. Porém dividir essa tarefa implica aplicar mais de uma vez o algoritmo de extração de regras de associação. Essa tarefa talvez não seja necessária se todas as relações fossem encontradas diretamente. Para não ser preciso mais de uma aplicação do algoritmo, extrair relações topológicas na etapa de pré-processamento dos dados e organizar esses dados para serem utilizados pelos algoritmos tradicionais de extração de regras de associação pode ser uma solução.

Assim, o objetivo deste trabalho é propor uma metodologia que pré-processa os dados georreferenciados viabilizando a extração de regras de associação em dados georreferenciados. Além disso, propor um módulo de pré-processamento de relações topológicas para dados georreferenciados com a finalidade de permitir a aplicação de algoritmos de mineração de regras de associação para encontrar padrões considerando o relacionamento espacial entre os dados. Os padrões devem apresentar características comuns dos dados georreferenciados em função de relações topológicas, considerando as características descritivas dos dados, além das relações. O intuito é pesquisar e especificar uma disposição dos dados no pré-processamento para aplicar o algoritmo *Apriori* e implementar um módulo para realizar esse pré-processamento. Esse módulo será avaliado aplicando dados do sistema de informação geográfica da cidade de Jaboticabal e os resultados analisados por um especialista em planejamento urbano.

A metodologia proposta é baseada no processo de mineração de dados, o qual possui as etapas de pré-processamento, extração de regras e pós-processamento. Esse trabalho propõe incluir a atividade de extração de relações topológicas na etapa de pré-processamento e definir um módulo para pré-processar essas relações de acordo com a metodologia sugerida.

Na etapa de pré-processamento os dados são organizados e selecionados. Atividades como remoção e transformação de dados são executadas, além da obtenção de relações topológicas. Como última tarefa da etapa de pré-processamento é necessário colocar os dados no formato adequado para o algoritmo de extração de regras de associação. A interpretação das regras de associação depende de como os relacionamentos espaciais estão representados nesse formato. Foram propostos três formatos, o primeiro para extrair regras de associação espacial de acordo com a definição proposta por Koperski (1999). O segundo formato para extrair regras de dados que obedecem uma determinada condição espacial, uma relação topológica. O terceiro formato utiliza cláusulas lógicas para ser aplicado em algoritmos de programação lógica indutiva.

Após pré-processar os dados, é necessário aplicar os algoritmos de extração de regras de associação, na etapa de extração de padrão. Para isso é executado um algoritmo tradicional de mineração de regras de associação, sem a necessidade de alteração do mesmo. Ao se

obter as regras, o resultado encontrado é avaliado na etapa de pós-processamento. Medidas como suporte e confiança podem ser utilizadas para verificar qual a frequência da regra no banco de dados. Porém, as vezes o volume de regras é elevado dificultando essa avaliação. Para facilitar a análise das regras pode-se utilizar filtros que eliminam regras desnecessárias, segundo um critério estabelecido. No caso da utilização de relações topológicas, esse critério pode, por exemplo, obrigar a presença de uma relação topológica como item da regra.

Para a avaliação geral da metodologia e do módulo de extração de relações topológicas, foi realizado um estudo de caso da cidade de Jaboticabal, utilizando um micro-computador com processador pentium 4 de 2,8 Ghz e memória de 512 Mb. No banco de dados encontram-se informações sobre lotes, quadras, setores, hidrografia, vegetação, estabelecimentos comerciais, edificações e a cobrança de IPTU. O tempo de processamento para extrair as relações topológicas foi elevado mas não tornou inviável a aplicação do módulo. O consumo de memória desse procedimento não ultrapassou 40 Mb. Foram considerados objetos espaciais referentes a lotes e escolas estaduais, sendo que os lotes possuem informações de benfeitoria, categoria do proprietário, uso do terreno, topografia, e pedologia e escolas com informações sobre o tipo de ensino. O módulo foi executado para a relação de proximidade buscando encontrar padrões em lotes próximos a escolas, considerando a proximidade de 300 metros. Os formatos de saída do módulo utilizados no estudo de caso foram os formatos para encontrar regras de associação espacial e para encontrar regras de associação em dados que obedecem uma determinada condição espacial. Após extrair as relações topológicas, o algoritmo *Apriori* foi executado utilizando as configurações padrões de suporte mínimo e confiança mínima. As regras de associação encontradas que não possuíam relações topológicas foram eliminadas e o conjunto de regras apresentado para um especialista avaliar o conhecimento extraído. Foram encontradas regras como $\text{BUFFER=Estadual P1_Lote_IPTU_USOTERRE=EDIFICADO} \Rightarrow \text{P1_Lote_IPTU_BENFEITO=CALÇADA_+_MURO}$ (11.6, 94.0). Essa regra pode ser interpretada como 11,6% de lotes que possuem terreno edificado e estão próximos de escola estadual tem como benfeitoria calçada e muro. Dos lotes que estão próximos de escolas e possuem uma construção, 94% também tem calçada e muro. As regras encontradas mostram que é possível obter regras em função de relações topológicas, considerando a extração das relações em uma etapa de pré-processamento dos dados.

Para a compreensão do estudo realizado, no Capítulo 2 é apresentado um estudo de sistemas de informações geográficas, com informações sobre o tipo dos dados armazenados e a organização dos mesmos. Esses dados são a base para o processo de mineração. No Capítulo 3, uma revisão do processo de mineração de dados, regras de associação e aplicação de regras de associação espacial são abordados. Essas informações dão suporte ao projeto realizado, apresentado no Capítulo 4. Para comprovar a viabilidade do módulo proposto foi realizado um estudo de caso, descrito no Capítulo 5. Por fim, as conclusões no Capítulo 6.

Sistema de informação geográfica

Estudar, entender e descrever a Terra e a sociedade é um recurso que permite melhorar a qualidade de vida da população, auxiliando na compreensão e previsão de eventos geográficos como variações climáticas, uso-ocupação do solo e distribuição populacional. Os dados provenientes desse estudo podem ser definidos, de uma maneira geral, como dados geográficos ou dados georreferenciados. Existem diferentes meios de coletar esses dados, por exemplo, levantamentos em campo, censos, imagens de satélite, imagens aéreas e mapas analógicos. Esses dados reunidos formam um poderoso e volumoso conjunto que muitas vezes o ser humano, sem uma ferramenta tecnológica, não é capaz de organizar e analisar. Assim, sistemas computacionais são desenvolvidos para melhorar o armazenamento, o gerenciamento e a análise de dados geográficos.

Nesse capítulo, são abordados conceitos e tecnologias voltados a dados geográficos apresentando o contexto em que este trabalho foi desenvolvido.

2.1 *Objetos espaciais*

Para utilizar dados geográficos é necessário definir um espaço E , entendido como um conjunto de objetos e relações entre esses objetos. Observe que é possível existir tipos diferentes de espaços dependendo de como a relação entre os objetos é definida. Para dados geográficos, é interessante considerar a distância como relação e a característica principal dos objetos, a localização no espaço. O espaço que tem como característica apresentar relação de distância é o espaço métrico (Maguire et al. 1991).

De uma maneira geral, os elementos pertencentes ao espaço podem ser definidos como objetos espaciais, formados por um conjunto de características descritivas situadas em uma localização do espaço. No caso específico da aplicação utilizando dados geográficos também é possível nomeá-los como objetos geográficos. Nota-se que esses objetos são entidades que têm tanto localização espacial como características de atributos independentes e podem ser influenciados por objetos vizinhos, sendo descritos por propriedades geométricas e de topologia. As propriedades geométricas incluem por exemplo, localização espacial, área, perímetro e as topológicas, inclusão, intersecção e adjacência (Malerba et al. 2001). A localização pode estar representada por coordenadas bi-dimensionais (x, y) ou tri-dimensionais (x, y, z) . Essas coordenadas são valores que representam um ponto no sistema de coordenadas utilizado no qual $x, y, z \in E$.

Tendo definido a localização do objeto espacial, é necessário encontrar como representar e tratar a geometria presente nos mesmos (Schneider 1997). Assim, deve-se mapear os dados espaciais do mundo real em classes específicas para um dado fenômeno. Por exemplo, um modelo de variação contínua da elevação topográfica de um terreno pode ser modelado em função de um modelo numérico de terreno. Uma vez definido o modelo conceitual é preciso encontrar uma forma eficiente de representação digital ou uma estrutura de dados. Dados espaciais apresentam problemas únicos para representação digital. A representação da realidade geográfica é complexa e um grande número de itens de informação espacial pode ser gerado de uma simples variável. Por exemplo, assim como é necessário representar todo o contorno de um mapa, pode também ser necessário a representação de casos pontuais variando a complexidade do modelo desejado (Maguire et al. 1991).

Para mapear os dados geográficos pode-se utilizar o paradigma dos quatro universos, em que são definidos o Universo do Mundo Real, o Universo Conceitual, o Universo da Representação e o Universo da Implementação (Camara and Monteiro 2002). Na Figura 2.1 pode-se observar o relacionamento entre os modelos considerados para incluir um objeto espacial do mundo real em um sistema digital, apresentados na sequência.

Figura 2.1: Separação em universos na modelagem de dados geográficos. Fonte: (Camara and Monteiro 2002)

No **Universo do Mundo Real** encontram-se os eventos estudados e que serão mapeados para os outros universos. Os problemas referentes a como os dados serão obtidos, medidos e

caracterizados são solucionados nessa etapa, além da decisão sobre qual projeção cartográfica e escala a ser utilizada.

No **Universo Conceitual**, também chamado de universo matemático, os dados geográficos são divididos em dados contínuos e dados individualizados. Camara and Monteiro (2002) apresentam as seguintes definições para modelar os dados nesse universo.

Região geográfica: é um dado contínuo. Uma região geográfica R é uma superfície qualquer pertencente ao espaço em que se localizam os objetos geográfico. Essa região pode ser representada em um plano ou reticulado.

Geo-campo: é um dado contínuo e representa uma distribuição espacial de uma variável que possui valores em todos os pontos pertencentes a uma região geográfica, num dado tempo t .

Geo-objeto: é um dado individualizado, definido como um elemento que possui localização espacial e atributos descritivos que o caracterizam, em uma região geográfica.

Objeto não-espacial: é um dado individualizado, não possui localização espacial diretamente como parte do objeto, mas pode ou não estar relacionado com um geo-objeto.

Essa classificação ainda pode ser especializada em tipos de dados geográficos.

Dados temáticos: são dados contínuos que podem ser classificados como geo-campo e representam temas ou classes como um todo, sendo possível visualizar a distribuição espacial de uma região que apresenta uma característica geográfica. Dados temáticos qualificam determinadas porções do espaço utilizando as classes para um evento geográfico. Um exemplo é o mapa de vegetação de uma região com as diversas classes de vegetação.

Dados cadastrais: são dados individualizados classificados como geo-objetos, descritos na forma de atributos. Podem estar associados a representações geométricas atribuindo valores e características a estas. Um exemplo é a característica de um lote como proprietário, endereço e valor.

Redes: são semelhantes aos dados cadastrais porém os elementos estão conectados, podendo apresentar informações de direção ou fluxo. É utilizada para representar dados como rodovias, sistemas de água, luz e telefone.

Modelos numéricos de terreno: são dados contínuos classificados como geo-campos, porém com informações quantitativas a respeito de determinada região e característica geográfica. Um exemplo são os dados de altimetria, representados em mapas topográficos que contém curvas de nível para demonstrar a variação da elevação do local.

Camara and Monteiro (2002) definem um modelo numérico de terreno como um modelo matemático que reproduz uma superfície real a partir de algoritmos e de um conjunto de pontos (x, y) , em um referencial qualquer, com atributos denotados z , que descrevem a variação contínua da superfície.

Imagem: não representa um dado geográfico em si, mas por meio das imagens é possível visualizar várias características geográficas e obter informações por meio de processamento da imagem ou pela visualização da mesma. É classificada como um geo-campo. Exemplos de imagens são aerofotos e imagens de satélite ou radar.

No **Universo de Representação** é feita a escolha quanto a forma de simbolizar dados, como representar os dados. A implementação desses dados em sistemas digitais estará diretamente ligada ao modelo de representação escolhido. Os modelos utilizados são o modelo matricial ou o modelo vetorial. No modelo matricial os dados são representados por uma matriz, isto é, uma grade quadriculada na qual cada quadrado da grade representa um elemento. No modelo vetorial, a geometria do objeto espacial é considerada para representar os dados, sendo que o objeto é mapeado para uma representação geométrica, ponto, linha e polígono. Para definir essas geometrias, dado um plano coordenado E no espaço, considere:

- um ponto é um par ordenado (x, y) de coordenadas, $(x, y) \in E$;
- um segmento de linha é formado por dois pontos distintos, denominados pontos extremos, conectados por um conjunto de pontos não coincidentes;
- uma reta é um segmento de linha que dado dois pontos quaisquer pertencente ao segmento, é satisfeita a equação da reta $(x - x_1)(y_2 - y_1) = (y - y_1)(x_2 - x_1)$;
- pontos colineares são pontos pertencentes a mesma reta;
- segmentos colineares são segmentos pertencentes a mesma reta;
- segmentos consecutivos são segmentos de linhas que possuem um ponto extremo em comum;
- conjunto de segmentos de linhas é um conjunto finito, com segmentos consecutivos conectados de maneira que quaisquer dois pontos de segmentos adjacentes compartilhem um ponto extremo e pontos extremos não pertencem a mais do que dois segmentos;
- um conjunto de segmentos de linhas é denominado simples se nenhum ponto, além dos pontos extremos, é compartilhado por dois segmentos, isto é, não existe intersecção dos segmentos além dos pontos extremos;

- um conjunto de segmentos de linhas é denominado fechado se todos os pontos extremos são compartilhados por dois segmentos, isto é, quando partindo de qualquer ponto, é possível percorrer todos os segmentos e regressar ao ponto de partida;
- uma linha é definida como um conjunto de segmentos de linhas simples e não fechado;
- linha poligonal é formada por um conjunto de segmentos de linha consecutivos e não-colineares;
- um polígono é formado por uma linha poligonal simples e fechada e por toda a região demarcada por essa linha.

As representações geométricas estão definidas a seguir.

Ponto: um ponto é um par ordenado (x, y) de coordenadas espaciais, $(x, y) \in E$. É o elemento sem dimensão, que representa a localização de um elemento, mas não informa o espaço ocupado, como uma capital de um país em um mapa-mundi.

Linha: uma linha é definida como um conjunto de segmentos de linhas simples e não fechado. É o elemento com uma dimensão, o comprimento. Pode representar, por exemplo, rua, estrada, rio e rota de transporte.

Polígono: um polígono é formado por uma linha poligonal simples e fechada e por toda a região demarcada por essa linha. É o elemento com duas dimensões, representando comprimento e largura. O polígono delimita fronteiras, elementos e regiões. Pode ser utilizado para descrever uma quadra, um bairro ou uma cidade.

Na Figura 2.2, pode-se observar o mesmo objeto representado em diferentes modelos. A esquerda encontra-se o objeto a ser mapeado. A direita estão os dois modelos utilizados. No modelo matricial, cada quadrado da matriz é um elemento. Nota-se que não existe mais o elemento individual caracterizado, somente pedaços desse elemento nos diferentes quadrados da grade. No modelo vetorial, o objeto é caracterizado por seu contorno, pelos polígonos e linhas formadas.

Dependendo do tipo de dado geográfico, pode ser necessária a opção por um determinado modelo ou pode ser feita a escolha por aquele que melhor represente os dados. Na Tabela 2.1, é apresentada a relação entre o tipo de dado e o modelo de representação correspondente. Os tipos de dados temático e modelo numérico de terreno podem ser representados tanto no modelo vetorial como no modelo matricial. Para esses tipos de dados, neste trabalho, foi considerado somente a representação usando modelos matriciais.

Figura 2.2: Representação geométrica de objetos espaciais. Adaptado de Ortiz (2005)

Os dados que não tem localização espacial ou são informações descritivas de um objeto espacial são caracterizados utilizando uma representação de atributos, independente do tipo de dado geográfico, geralmente seguindo modelos de banco de dados relacionais ou orientados a objetos.

Tabela 2.1: Relação entre tipos de dados geográficos e o modelo de representação. Extraído de Camara and Monteiro (2002)

Tipo de dado geográfico	matricial	vetorial
Dado temático	sim	sim
Dado cadastral	não	sim
Rede	não	sim
Modelo numérico de terreno	sim	sim
Imagem	sim	não

Por fim, no **Universo de Implementação**, os dados são codificados usando uma linguagem de programação. As estruturas de dados e os algoritmos necessários são escolhidos de acordo com a representação adotada no universo de representação.

Assim, ao se trabalhar com dados espaciais é necessário utilizar um modelo de dados que representa o espaço, uma descrição a partir da compreensão humana do espaço, formal ou informal e de objetos contidos nesse espaço. Além disso, é importante compreender como esses objetos se relacionam. Existem diferentes tipos de relacionamentos, abordados a seguir.

2.2 Operações espaciais e relações topológicas

Objetos espaciais apresentam relacionamentos em função da localização espacial. Esses relacionamentos podem ser definidos por meio de operações espaciais. Schneider (1997) classifica as operações espaciais em cinco categorias.

A primeira categoria inclui relações espaciais que comparam dois objetos, considerando alguma propriedade de localização e retornam um valor booleano. Entende-se que essas relações são relacionamentos entre dois ou mais objetos espaciais e também podem ser definidas como predicados espaciais. Basicamente é possível agrupar as relações espaciais em relações topológicas, métricas, de ordenação espacial, *fuzzy* e direcionais. As **relações topológicas** consideram a topologia entre os objetos e incluem conceitos como igualdade, disjunção, intersecção, toque, cruzamento, inclusão, sobreposição e cobertura, não dependendo de distância. Por exemplo, dois lotes são adjacentes se um toca o outro, estando lado a lado. As **relações métricas** usam medidas como distâncias e direções, por exemplo a distância em quilômetros entre uma cidade e a capital de um estado. As **relações de ordenação espacial** são baseadas em definições de ordem. Cada relação de ordem tem uma relação inversa. Por exemplo, uma casa está atrás de uma escola e possui a relação inversa em_frente_de. As **relações fuzzy** tratam de propriedades imprecisas, como definir perto, próximo ou longe. Essas propriedades são subjetivas e variam de indivíduo para indivíduo. Por exemplo, definir se um lote está próximo de um rio. As **relações direcionais** comparam posições relativas entre dois objetos, como norte e sul. Por exemplo, informar se uma rodovia está ao norte de uma cidade.

Na segunda categoria as operações retornam uma parte de um objeto espacial, extraindo as geometrias que constituem o objeto. Por exemplo, considere um rio representado por um conjunto de linhas, poderia ser encontrado um trecho do rio, isso é, uma linha específica pertencente ao conjunto.

Na terceira categoria, as operações espaciais retornam um valor escalar, isto é, computam uma propriedade métrica de um objeto espacial. Por exemplo, dado o objeto espacial quadra em uma cidade, definido por um polígono, uma operação dessa categoria calcula e retorna a área total da quadra.

Na quarta categoria, as operações retornam como resultado novos objetos espaciais, podendo ser originados de outros objetos espaciais por meio da junção, divisão ou transformação geométrica como redução do tamanho. Por exemplo, um setor em uma cidade pode ser definido selecionando diversas quadras e concatenando todas as quadras para formar um novo objeto, o setor.

Na quinta categoria, as operações incluem consultas aos objetos espaciais, podendo utilizar as relações definidas na primeira categoria. Essas consultas podem ser a seleção de

objetos espaciais e não espaciais a partir de uma determinada condição. Por exemplo, encontrar todas as cidades com uma distância de oitenta quilômetros de um rio.

Dentre todas essas relações espaciais, neste trabalho foram consideradas somente relações topológicas. Existem diferentes definições para essas relações, não existindo uma unanimidade na utilização das mesmas. Uma possível maneira de definir relações topológicas é a partir dos conceitos de fronteira, interior e exterior. Uma fronteira pode ser definida, no caso do polígono, como a linha poligonal que o define, no caso da linha como os dois pontos extremos e o ponto não possui fronteira. O interior pode ser definido como o conjunto de pontos pertencente ao objeto espacial exceto a fronteira. No caso do ponto o interior é o próprio ponto. O exterior pode ser definido como os pontos não pertencentes ao objeto, isto é, nem à fronteira e nem ao interior. Utilizando os trabalhos de Egenhofer and Herring (1990) e Clementini et al. (1993), pode-se encontrar oito relações topológicas (Terralib 2005), sendo duas ainda subdivididas, totalizando dez relações. Para defini-las, considere:

O - um objeto espacial

A, B - números inteiros

$B(O)$ - o conjunto de pontos pertencentes a fronteira de O .

$I(O)$ - o conjunto de pontos pertencentes ao interior de O .

$E(O)$ - o conjunto de pontos pertencentes ao exterior de O .

$\max(A, B)$ - função que retorna o maior valor entre os inteiros tal que

$$\max(A, B) = \begin{cases} A, & \text{Se } A > B; \\ B, & \text{Se } A < B. \end{cases}$$

$\dim(O)$ - a função que retorna a dimensão tal que

$$\dim(O) = \begin{cases} -, & \text{Se } O = \emptyset; \\ 0, & \text{Se } O \text{ contém ao menos um ponto e não contém linhas ou polígonos;} \\ 1, & \text{Se } O \text{ contém ao menos uma linha e não polígonos;} \\ 2, & \text{Se } O \text{ contém ao menos um polígono.} \end{cases}$$

A relação de **igualdade** é uma relação simétrica que define se um objeto espacial é igual ao outro. Também pode ser denominada de *equals* e *TeEquals*. Essa relação é aplicada aos relacionamentos geométricos ponto com ponto, linha com linha e polígono com polígono. Dado dois objetos espaciais X e Y :

$$\text{igual}(X, Y) = (X \cap Y = X) \wedge (Y \cap X = Y)$$

A relação de **disjunção** é uma relação simétrica que define quando um objeto espacial não está relacionado com o outro objeto por meio de qualquer outra relação topológica. Também

pode ser denominada de *disjoint* e *TeDisjoint*. Essa relação é aplicada aos relacionamentos geométricos ponto com ponto, ponto com linha, ponto com polígono, linha com ponto, linha com linha, linha com polígono, polígono com ponto, polígono com linha e polígono com polígono. Dado dois objetos espaciais X e Y :

$$\text{disjunto}(X, Y) = (X \cap Y = \emptyset)$$

A relação de **intersecção** é uma relação simétrica que define quando um objeto espacial possui qualquer ponto em comum com outro objeto. Também pode ser denominada de *intersect* e *TeIntersects*. Essa relação é aplicada aos relacionamentos geométricos ponto com ponto, ponto com linha, ponto com polígono, linha com ponto, linha com linha, linha com polígono, polígono com ponto, polígono com linha e polígono com polígono. Dado os objetos espaciais X , Y :

$$\text{intersecção}(X, Y) = (X \cap Y \neq \emptyset)$$

A relação de **toque** é simétrica e definida quando um objeto espacial possui pontos somente na fronteira do outro objeto espacial. Também pode ser denominada de *touch* e *TeTouches*. Essa relação é aplicada aos relacionamentos geométricos ponto com linha, ponto com polígono, linha com ponto, linha com linha, linha com polígono, polígono com ponto, polígono com linha e polígono com polígono. Dado dois objetos espaciais X e Y :

$$\text{toca}(X, Y) = (I(X) \cap I(Y) = \emptyset) \wedge (X \cap Y \neq \emptyset)$$

A relação de **cruzamento** é uma relação simétrica que define quando um objeto espacial cruza outro objeto espacial de maneira que os pontos em comuns formem uma geometria com uma dimensão menor do que a maior das geometrias. No caso de duas linhas, somente existirá um ponto em comum, o ponto onde ocorreu o cruzamento. No caso de um polígono e uma linha, o cruzamento irá ser uma linha. Essa relação também pode ser denominada de *cross* e *TeCrosses* e é aplicada aos relacionamentos geométricos linha com linha, linha com polígono e polígono com linha. Dado dois objetos espaciais X e Y :

$$\text{cruza}(X, Y) = (\dim(I(X) \cap I(Y)) = (\max(\dim(X), \dim(Y)) - 1) \wedge (X \cap Y \neq X) \wedge (X \cap Y \neq Y))$$

A relação de **inclusão** define se um objeto está completamente no interior de outro. É composta por duas relações, contém e contido. A relação contido também pode ser denominada de *within* e *TeWithin*. A relação contém pode ser denominada *contain* e *TeContains*. Além disso, a relação contido é aplicada aos relacionamentos geométricos ponto com ponto, ponto com linha, ponto com polígono, linha com linha, linha com polígono e polígono com polígono e a relação contém é aplicada aos relacionamentos geométricos ponto com ponto, linha com ponto, linha com linha, polígono com ponto, polígono com linha e polígono com polígono. Dado dois objetos espaciais X e Y :

$$\text{contido}(X, Y) = (X \cap Y = X) \wedge (I(X) \cap I(Y) \neq \emptyset) \wedge (B(X) \cap B(Y) = \emptyset)$$

$$\text{contem}(X, Y) = \text{contido}(Y, X)$$

A relação de **sobreposição** é uma relação simétrica definida quando um objeto espacial relacionado com outro objeto espacial possui o objeto resultado da relação na mesma dimensão que os objetos originais. Isto é, uma linha em intersecção com outra linha deve ter como pontos comuns uma linha. Um polígono em intersecção com outro polígono deve ter como resultado um polígono. Essa relação também pode ser denominada de *overlap* e *TeOverlaps* e é aplicada aos relacionamentos geométricos linha com linha e polígono com polígono. Dado dois objetos espaciais X e Y :

$$\text{sobrepor}(X, Y) = (\dim(I(X)) = \dim(I(Y)) = \dim(I(X) \cap I(Y))) \wedge (X \cap Y \neq X) \wedge (X \cap Y \neq Y)$$

A relação de **cobertura** define se um objeto está no interior de outro e também possui algum ponto em comum entre as fronteiras. É composta por duas relações, cobre e coberto. A relação cobre também pode ser denominada de *cover* e *TeCovers*. A relação coberto pode ser *covered* e *TeCoveredBy*. Além disso, a relação cobre é aplicada aos relacionamentos geométricos linha com linha, polígono com linha e polígono com polígono e a relação coberto é aplicada aos relacionamentos geométricos linha com linha, linha com polígono e polígono com polígono. Dado dois objetos espaciais X e Y :

$$\text{coberto}(X, Y) = (X \cap Y = X) \wedge (I(X) \cap I(Y) \neq \emptyset) \wedge (B(X) \cap B(Y) \neq \emptyset)$$

$$\text{cobre}(X, Y) = \text{coberto}(Y, X)$$

Na Figura 2.3, pode-se observar exemplos de geometrias e relações entre essas geometrias. Os objetos espaciais com geometrias e relações espaciais são tratados por sistemas específicos, sistemas de informações geográficas, apresentados na próxima seção.

2.3 Sistemas de informações geográficas

O propósito de manipular de forma digital dados espaciais é dar ao usuário um método de consulta e análise da variação geográfica. Geralmente, a parte cadastral dos dados espaciais estão armazenadas em banco de dados e a informação espacial em ferramentas gráficas (CAD do inglês *Computer Aided Design*), em imagens obtidas por sensoriamento remoto ou em cartografia apoiada por computador. Um sistema de informação geográfica busca auxiliar na recuperação, classificação e análise de dados espaciais, tratando todas as funcionalidades

Figura 2.3: Exemplo de relações topológicas. Adaptado de (Clementini et al. 1993)

dos sistemas anteriores em um único sistema e incorporando também a análise topológica. Pode ser visto como a integração das diferentes ferramentas para trabalhar dados geográficos (Maguire et al. 1991). Na Figura 2.4 é possível visualizar como um sistema de informação geográfica está relacionado com as diversas ferramentas e outras áreas

Em Câmara and Davis (1999) um sistema de informação geográfica é definido como uma ferramenta computacional para automatizar o processo de gestão de informações geográficas constituída de módulos como banco de dados, interface com o usuário, visualização e plotagem de dados, consulta e análise espacial, integração e gerência dos dados. O valor potencial de um sistema de informação geográfica é sustentado por sua habilidade em permitir a análise de dados espaciais usando técnicas de análise espacial (Maguire et al. 1991).

Figura 2.4: Sistemas de informações geográficas. Fonte: (Maguire et al. 1991)

Os primeiros sistemas de informações geográficas possuíam uma arquitetura simples e foram evoluindo para os sistemas complexos atuais. Basicamente existem oito diferentes arquiteturas (Ribeiro and Câmara 2002).

Arquitetura tradicional: o sistema armazena separadamente os componentes, gráficos e alfanuméricos, com estruturas de armazenamento proprietárias e banco de dados integrado ao sistema. Além disso, possui interface com o usuário e um núcleo para integrar e controlar todos os elementos do sistema. O núcleo pode implementar as funcionalidades de geoprocessamento ou podem existir módulos externos que forneçam essas funcionalidades.

Arquitetura dual: semelhante a anterior, com o banco de dados para armazenagem de dados alfa-numéricos constituído de um banco de dados relacional externo ao sistema. A armazenagem de componentes gráficos é feita pelo sistema.

arquitetura baseada em CAD: o gerenciamento e armazenamento de componentes espaciais é externo, realizado por ferramentas CAD. Os outros elementos do sistema são semelhantes a arquitetura anterior.

Arquitetura relacional: os componentes espaciais e alfanuméricos estão armazenados em um mesmo elemento da arquitetura, um banco de dados relacional e muitas vezes já oferece mecanismos específicos para trabalhar com dados espaciais, como a utilização de estrutura de dados espaciais e consultas topológicas.

Arquitetura orientado a objetos: o banco de dados utilizado é orientado a objetos. Com isso, toda a modelagem do sistema também está inserida no contexto de orientação a objetos.

Arquitetura *desktop mapping*: possui foco na interface com o usuário e na produção de mapas. No geral as estruturas de armazenamento são arquivos, mas permite comunicação com outras aplicações.

Arquitetura baseada em imagens: o foco principal são as imagens e as funcionalidades são implementadas visando a melhor performance do processamento das imagens. Os dados que não são espaciais são armazenados em um banco de dados relacional.

Arquitetura integrada: semelhante a arquitetura dual, permitindo o tratamento de dados espaciais considerando o modelo matricial e vetorial, com formas de armazenamento separados para cada modelo e feitas, em geral, por estruturas proprietárias.

Existem diversos sistemas de informações geográficas como *ARCGIS*, GeoMidia e Spring¹. Em qualquer sistema de informação geográfica o ponto principal a ser tratado é como organizar os dados no banco de dados espaciais e como recuperá-los com eficiência. Na próxima seção serão introduzidas algumas características de um banco de dados espaciais.

2.3.1 Banco de dados geográficos

O armazenamento de dados é um campo muito estudado na área de computação por ser, na maioria dos casos, o foco principal dos sistemas e também o gargalo, pois acesso a disco é uma atividade dispendiosa. Os bancos de dados atuais provêm diversas funcionalidades para facilitar esse armazenamento mas muitas vezes não são suficientes para abranger as diferentes aplicações e são necessárias alterações e extensões para atingir as funcionalidades desejadas. Banco de dados espaciais pode ser visto como uma extensão de um banco de dados padrão e tem como característica predominante ser composto por dados descritivos associados a uma representação geométrica que possui localização espacial. No caso de utilização de dados geográficos, o banco de dados espaciais pode ser nomeado como banco de dados geográficos.

Para tratar o acesso aos dados espaciais, são utilizadas estruturas de dados específicas para melhorar o desempenho na manipulação desses dados como Quad-tree, Kd-tree e R-tree (Samet 1990). Além do uso de estruturas de dados específicas, pode-se encontrar extensões da linguagem SQL para comportar alguns comandos relacionados a dados georreferenciados. Na Figura 2.5 é mostrado um exemplo de uma consulta na sintaxe SQL com extensão para incluir relações topológicas, baseada na sintaxe utilizada pelo banco de dados MySQL.

¹ Desenvolvido pelo Instituto Nacional de Pesquisas Espaciais (INPE) em parceria com o grupo de Computação Gráfica da Pontifícia Universidade Católica do Rio de Janeiro e a Fundação de Ciência, Aplicações e Tecnologia Espaciais.

Nesse exemplo, são definidos dois polígonos e a consulta verifica se o polígono 1 está dentro do polígono 2. Nem sempre essas extensões apresentam um padrão e podem variar de um banco de dados espaciais para outro. Também, dependendo do sistema de informação geográfica, algumas funcionalidades podem ser implementadas no programa principal ou ser disponibilizadas no banco de dados.

```
mysql> SET @g1 = GeomFromText('Polygon((0 0,0 3,3 3,3 0,0 0))');  
mysql> SET @g2 = GeomFromText('Polygon((0 0,0 5,5 5,5 0,0 0))');  
mysql> SELECT MBRWithin(@g1,@g2);
```

Figura 2.5: Consulta espacial.

A modelagem do banco de dados espaciais está diretamente relacionada com a modelagem conceitual do sistema a ser implementado. Existem diversos modelos para organizar e representar os dados, como os tradicionais modelo relacional e orientado a objetos ou modelos específicos para aplicações geográficas como MGEO, GISER, GMOD e GEO-OMT. A descrição e comparação desses diferentes modelos pode ser encontrado em Borges and Davis (2002). Para compreender melhor as idéias de uma modelagem de dados espaciais, na próxima seção será apresentada a modelagem de dados do sistema de informação geográfica Spring. A partir dessa modelagem será descrita a modelagem de dados do visualizador TerraView e a biblioteca TerraLib, ambos utilizados neste trabalho.

2.3.2 *Spring*

O Spring é um sistema de informação geográfica com licença livre. As principais funcionalidades são o gerenciamento de banco de dados espaciais, consulta e análise de dados, tratamento de modelos numéricos de terreno e processamento de imagens, trabalhando com modelos matriciais e vetoriais (Spring 2005a). Esse aplicativo não foi adotado neste trabalho, mas o modelo de dados e alguns termos referentes a esse modelo foram utilizados. O modelo conceitual do Spring foi adaptado e pode ser visto na Figura 2.6. Nesse modelo foi incluída a separação de projetos e categorias em um banco de dados.

Na figura pode-se observar a existência de diversos projetos no banco de dados, de forma que para cada projeto é necessário ter um retângulo envolvente delimitando a região representada. Além disso, cada projeto contém diversas categorias associadas. Uma categoria representa um conjunto de informações comuns, organizadas logicamente de acordo com a aplicação, e possui o mesmo tipo de dado geográfico ou é constituída por geo-objetos e elementos não espaciais, isto é, elementos que não possuem qualquer referência espacial. Por exemplo, um projeto de uma cidade pode ter categorias de cadastro urbano, rede de logradouro e cobertura de solo. Contudo as categorias são mais uma divisão lógica e definição

de um tipo de dados do que a representação dos dados reais. Esses são representados de fato nos elementos associados a categoria, no geo-objeto, no plano de informação (PI) e nos dados não espaciais, descritos a seguir (Spring 2005b).

Geo-objeto: é um elemento que possui localização espacial e atributos descritivos, representando um elemento individualmente. Por exemplo, em um mapa de uma rede de esgoto, cada segmento da rede é um geo-objeto, sendo representado por linhas conectadas. O conjunto de geo-objetos de um mapa pode ser definido como um mapa de geo-objetos e pode conter informações cadastrais ou referente a redes.

Plano de informação: representa um conjunto de dados geográficos, em uma determinada região definida por um retângulo envolvente e que apresenta os mesmos atributos descritivos. O plano de informação está associado a uma categoria e por isso apresenta o mesmo modelo de dado atribuído a categoria, isto é, cadastral, rede, temático, imagem ou modelo numérico de terreno. Pode representar um mapa de geo-objetos ou um geo-campo, dependendo do modelo atribuído.

Dados não espaciais: são tabelas alfa-numéricas que não são georreferenciadas. Esses dados são utilizados associados aos geo-objetos por meio de algum campo alfa-numérico comum. Por exemplo, pode-se ter em uma prefeitura informações de proprietários de lotes em uma tabela não espacial e ter um mapa de geo-objetos dos lotes da cidade. Existindo um campo em comum pode-se associar essas informações para eventuais consultas sobre os proprietários dos lotes de uma determinada região.

Na biblioteca TerraLib, e conseqüentemente no software TerraView, as idéias apresentadas anteriormente também estão presentes. Assim, esse modelo conceitual pode ser utilizado para compreender os modelos de dados do TerraLib e TerraView apresentados na próxima seção.

2.3.3 *TerraLib e TerraView*

O TerraLib é uma biblioteca para auxiliar na implementação de sistema de informação geográfica composta por um conjunto de classes e funções para acessar banco de dados espaciais, organizar e manipular dados georreferenciados. Possui licença livre para o uso e foi implementada utilizando ANSI C++ (Terralib 2005). Essa biblioteca possui classes que representam o banco de dados como a classe **TeDatabase**, o plano de informação **TeLayer**, a representação utilizada **TeRepresentation**, os atributos de uma tabela do banco de dados **TeAttributeList** e **TeAttribute**, e a geometria do objeto **TeGeometry**. A classe **TeDatabase**, em

Figura 2.6: Modelo conceitual adaptado do Spring. Fonte: (Spring 2005b)

específico, disponibiliza funções para encontrar relações topológicas entre os objetos como a função espacial apresentada na Figura 2.7.

As operações topológicas disponíveis são: **TeDisjoint**, **TeTouches**, **TeCrosses**, **TeWithin**, **TeOverlaps**, **TeContains**, **TelIntersects**, **TeEquals**, **TeCovers**, **TeCoveredBy**.

A biblioteca TerraLib foi utilizada para implementar o visualizador de dados TerraView, desenvolvido pela mesma equipe que desenvolveu a biblioteca. Esse aplicativo permite, entre outras funcionalidades, criar ou importar dados georreferenciados para o padrão definido pela biblioteca TerraLib e trabalhar com os modelos de representação vetorial ou matricial. O TerraView possui um modelo de dados com algumas diferenças em relação ao Spring. As principais entidades desse modelo estão definidas a seguir (TerraView 2005).


```
spatialRelation( nome_da_primeira_tabelaPi,  
                 tipo_de_representação_da_primeira_tabela,  
                 vetor_de_objetos_de_entrada,  
                 nome_da_segunda_tabelaPi,  
                 tipo_de_representação_da_segunda_tabela,  
                 vetor_de_objetos_de_saida,  
                 operação)
```

Figura 2.7: Exemplo de função do TerraLib.

Banco: é a entidade utilizada para representar um banco de dados relacional ou com tratamento de dados espaciais, definido pelo usuário. Todos os dados estão armazenados no banco de dados, tanto os dados não espaciais como espaciais.

Plano de Informação: é a entidade que representa cada característica geográfica para acesso ao dados georreferenciados. Um banco de dados está organizado de acordo com planos de informação, apresentando as mesmas características descritas para o Spring. Porém a visualização dos dados georreferenciados não é possível diretamente a partir de um plano de informação.

Vistas: podem ser entendidas como as categorias do Spring com informações sobre projeção e localização. Em TerraView (2005) vistas são definidas como mapas interativos que permitem mostrar, consultar e analisar os dados geográficos. Uma vista contém um conjunto de temas e cada tema é apresentado na tela de visualização em função dos parâmetros cartográficos definidos para a vista.

Temas: são utilizados para exibir os dados de um plano de informação. Em TerraView (2005) encontra-se a definição - temas são definidos para exibir o conteúdo de um PI que está no banco ativo. Um tema mostra um PI na projeção cartográfica da vista a qual está associado.

Tanto a biblioteca TerraLib quanto o visualizador TerraView foram utilizados neste trabalho, permitindo importar, armazenar e manipular dados georreferenciados.

Assim, esse conjunto de ferramentas disponibiliza uma grande quantidade de informações. Existem métodos de análise espacial, principalmente referentes a análise estatística para auxiliar na compreensão e análise dos mesmos. Além disso, pode-se utilizar técnicas de extração de conhecimento para melhorar e aproveitar o conhecimento intrínseco aos dados. Dentre essas técnicas, pode-se utilizar mineração de dados, apresentada no Capítulo 3.

Mineração de dados

A utilização de banco de dados geográficos proporciona um elevado número de informações de natureza diferente, associando características sociais com aspectos geográficos de uma determinada região. Em geral, a quantidade de dados disponíveis dificulta a percepção dessas associações sendo utilizado, na maioria das vezes, métodos estatísticos para encontrar padrões em banco de dados geográficos. Além de métodos estatísticos, é possível automatizar a extração de padrões com o processo de mineração de dados.

3.1 Descoberta de conhecimento

Mineração de dados busca extrair padrões, regras ou modelos que não são claros ou visíveis à percepção humana, isso é, estão escondidos no conjunto de dados. Fayyad et al. (1996) definem extração de conhecimento em banco de dados como o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis que estejam embutidos nos dados.

Existem diferentes abordagens para a divisão em etapas do processo de descoberta de conhecimento. Weiss and Indurkha (1998) dividem o processo nas etapas de preparação dos dados, redução dos dados, definição dos experimentos e execução seguida da análise dos resultados. Fayyad et al. (1996) propõem uma abordagem em nove etapas, envolvendo conhecer o domínio da aplicação, obter o conjunto de dados, limpar e pré-processar os dados, reduzir o volume de dados, escolher a função de mineração, escolher o algoritmo, minerar os dados, interpretar os resultados e utilizar o conhecimento descoberto. Rezende et al.

(2003) consideram uma abordagem que divide nas etapas de identificação do problema, pré-processamento, extração de padrões, pós-processamento e utilização do conhecimento. O processo é iterativo, podendo ser executado diversas vezes, com diferentes configurações, pois nem sempre os melhores resultados são obtidos com uma única execução. É possível recomeçar desde o início do processo, selecionando melhor os dados, alterando algoritmos e configurações para encontrar o melhor resultado, ou simplesmente voltar a uma etapa anterior. Além disso, esse processo é centrado na interação entre as diversas classes de usuários e o seu sucesso depende, em parte, dessa interação. Os usuários do processo podem ser divididos em três classes: especialista do domínio, que deve possuir amplo conhecimento do domínio da aplicação e deve fornecer apoio para a execução do processo; analista, que deve conhecer profundamente as etapas que compõem o processo de mineração de dados e ser responsável por sua execução; e usuário final, que representa a classe de usuários que utiliza o conhecimento extraído no processo para auxiliá-lo em tomada de decisão. Essa abordagem é mostrada na Figura 3.1 e adotada nesse trabalho.

Figura 3.1: Etapas do processo de extração de conhecimento. Fonte: (Rezende et al. 2003)

Na etapa de identificação do problema, é necessário estudar o domínio da aplicação, o tipo de informação que será trabalhada, os aspectos relevantes da área da aplicação, especificando os objetivos e as metas a serem alcançados no processo de mineração de dados. Requer uma grande interação entre o especialista e o analista, visando uma adequada compreensão do problema. As três etapas seguintes serão abordadas nas próximas seções. Ao fim do processo, têm-se a utilização do conhecimento, na qual o usuário final tem disponíveis os conhecimentos extraídos no processo.

3.1.1 Pré-processamento

Na etapa de pré-processamento os dados são obtidos e preparados para serem utilizados. Essa etapa trata da organização para a execução do algoritmo de extração de padrão. Diversas tarefas são necessárias para garantir um bom resultado (Batista 2003). Essas tarefas podem ser divididas em fases as quais são descritas a seguir.

A fase de **extração e unificação** tem como objetivo integrar os dados que podem ser originados de diversas fontes de dados como banco de dados, *data warehouse*, arquivos textos ou planilhas. No processo de mineração de dados, não é possível trabalhar com todos esses formatos, sendo que cada algoritmo de aprendizado tem um padrão de entrada. O formato mais utilizado é o formato atributo-valor, na qual um atributo representa uma característica de um exemplo. Um exemplo é constituído de um conjunto de atributos relacionados que descreve um registro. No formato atributo-valor tem-se um conjunto de exemplos T com n exemplos e m atributos, sendo a linha i referente ao i -ésimo exemplo ($i=1,2,...n$) e a entrada x_{ij} ao valor do j -ésimo ($j=1,2,...m$) atributo X_j do exemplo i . Cada exemplo corresponde a uma tupla $(x_{i1}, x_{i2}, \dots, x_{im}, Y_i)$, em que, em alguns casos, Y_i é um atributo especial denominado atributo-meta, ou seja, a classe referente a esse exemplo. Nem sempre o atributo-meta é um atributo obrigatório, isto depende da atividade de mineração desejada (Monard and Baranauskas 2003). Pode-se observar um exemplo do formato atributo-valor na Tabela 3.1. Em mineração de regras de associação, os termos itens e transações são utilizados para se referir aos atributos e exemplos, respectivamente. Na fase de **transformação**

Tabela 3.1: Conjunto de dados no formato atributo-valor.

	X₁	X₂	...	X_m	Y
T₁	x_{11}	x_{12}	...	x_{1m}	Y_1
T₂	x_{21}	x_{22}	...	x_{2m}	Y_2
...	...	...	...	...	...
T_n	x_{n1}	x_{n2}	...	x_{nm}	Y_n

é necessário trabalhar os dados para melhor representar um conceito. Algumas transformações possíveis são: normalização, discretização de atributos quantitativos, transformação de atributos qualitativos em quantitativos e conversão de tipo de dados (Batista 2003).

Também pode ser preciso realizar uma **limpeza dos dados**. O bom desempenho da mineração de dados está diretamente relacionado com a qualidade dos mesmos. Porém, o processo de coleta de dados nem sempre consegue garantir que todos os dados estejam corretos e que todos os atributos possuam valores preenchidos. Com a finalidade de evitar um resultado alterado por esses fatores é necessário realizar uma limpeza nos dados. O conhecimento do domínio obtido na etapa de identificação do problema pode ser aplicado para identificar registros com valores inválidos ou de granularidade incorreta em algum atributo. Também pode-se utilizar técnicas que não se baseiam no conhecimento do domínio, como

estratégias para tratamento de atributos incompletos.

As vezes pode ser necessário uma **redução de volume** dos dados devido ao tempo de processamento e consumo de memória. Basicamente, pode-se reduzir o volume de dados por meio da exclusão de atributos, exemplos ou da diminuição do número de valores possíveis de um atributo. Weiss and Indurkha (1998) e Lee (2005) propõem estratégias para selecionar atributo, exemplo ou valores para serem eliminados, que gerem poucas influências no resultado da mineração de dados.

3.1.2 *Extração de padrões*

Após a etapa de pré-processamento ocorre a extração de padrões, sendo necessário escolher a atividade e o algoritmo a serem empregados para obter os resultados. A primeira fase dessa etapa é a escolha da atividade de mineração. Pode-se optar entre predição e descrição dependendo do objetivo da mineração (Rezende et al. 2003).

Atividades preditivas utilizam indução para construir o modelo que representa o conhecimento aprendido. Os exemplos de entrada já possuem atributos-metas definidos, ou seja, os exemplos são rotulados e são generalizados para a obtenção do modelo. Assim será obtido um modelo para ser usado na predição de novos exemplos. Essas atividades podem ser divididas em classificação e regressão. Ambas buscam encontrar conjuntos de relacionamentos nos dados (exemplos) que determinam a classe de um objeto de acordo com seus atributos. Esses relacionamentos poderão ser utilizados para classificar novos exemplos que ainda não possuem atributo-meta associado. A diferença entre essas duas tarefas é que classificação trabalha com valores discretos e regressão, contínuos.

Por outro lado, pode-se ter dados para serem minerados que não possuam atributo-meta especificado, ou seja, sem rótulos pré-definidos. Para esses casos, os padrões obtidos no processo de mineração de dados auxiliam na descrição e compreensão dos dados. Essa atividade de mineração é conhecida como atividade descritiva e possui como objetivo principal identificar comportamentos dos dados. Algumas atividades de descrição são apresentadas a seguir:

- Associação é utilizada para caracterizar o quanto à presença de um conjunto de itens está relacionada com a presença de outro conjunto de itens;
- *Clustering* permite mapear os dados em agrupamentos, encontrando grupos naturais baseados em métricas de similaridade;
- Sumarização oferece mecanismos para obter uma descrição compacta para um subconjunto dos dados;

- Análise sequencial permite encontrar tendências nos dados, auxiliando na detecção de ocorrências de padrões sequenciais num fluxo de identificação de categorias.

Na Figura 3.2 são apresentados os tipos de atividades de mineração de dados.

Figura 3.2: Tipos de atividades de mineração de dados. Fonte: (Rezende et al. 2003)

Após a definição da atividade é necessário escolher o algoritmo a ser utilizado, além de especificar a sua configuração. A escolha do algoritmo está relacionada com a capacidade de manipular tipos de dados, o formato dos dados de entrada do algoritmo, compreensibilidade do modelo de saída, nível de familiaridade do usuário, escalabilidade, requisitos de software e hardware (Cabena et al. 1997). Por fim, aplica-se o algoritmo escolhido para encontrar os padrões nos dados.

3.1.3 Pós-processamento

No pós-processamento, os resultados são analisados, verificando se o conhecimento encontrado é compreensível, se condiz com a realidade ou se não está representando somente os dados utilizados. Os algoritmos podem extrair uma grande quantidade de padrões, mas nem todos são interessantes para o usuário e essa grande quantidade dificulta a compreensão e utilização do conhecimento (Rezende et al. 2003).

Em um primeiro momento o conhecimento pode ser filtrado, dependendo da atividade de mineração escolhida e do algoritmo aplicado, utilizando técnicas de poda, restrição de atributos ou ordenação de regras com o auxílio de métricas. Também é possível realizar a interpretação e explanação, comparando, documentando, visualizando ou modificando o conhecimento, tornando-o compreensível para o usuário.

Outra importante tarefa dessa etapa é a avaliação do conhecimento extraído segundo medidas de desempenho ou de qualidade. As medidas de desempenho buscam avaliar as regras de acordo com os dados originais, verificando o quanto a regra é representativa para

o conjunto, como a cobertura, precisão e erro. Essas medidas representam a fidelidade com que uma regra representa os dados.

As medidas de qualidade têm foco no usuário, tentando avaliar a compreensibilidade e interessabilidade dos padrões. A compreensibilidade está relacionada com a facilidade de interpretação dessas regras pelo ser humano, por exemplo considerando o significado sintático de uma regra, pelo número de regras e pelo número de condições por regra. Já a interessabilidade busca estimar o quanto um conhecimento é interessante, útil ou inesperado para o usuário, por exemplo, buscando descrever o grau de utilidade das regras para o usuário final do processo de mineração de dados (Horst 1999).

Além dessas atividades, no pós-processamento, é possível integrar conhecimentos obtidos por meio de diferentes abordagens para melhorar a precisão e o desempenho do sistema. Após essa etapa, pode-se decidir se o conhecimento encontrado atendeu aos objetivos propostos. Caso isso não ocorra é possível realizar novamente algumas etapas ou todo o processo de mineração de dados.

Todas as etapas de mineração de dados apresentadas podem ser aplicadas às diferentes atividades descritivas e preditivas, como em Associação, atividade selecionada para ser aplicada nesse trabalho por encontrar relações entre os dados auxiliando na detecção de novos padrões.

3.2 Associação

Associação é uma técnica de atividade descritiva de mineração de dados que extrai o conhecimento a partir de exemplos construindo representações compreensíveis ao ser humano. A representação utilizada é uma regra indicando o quanto a presença de um conjunto de itens s está relacionada com a presença de outro conjunto de itens distintos u nos mesmos registros.

O formato da regra pode ser apresentado como $LHS \Rightarrow RHS$, sendo LHS e RHS são respectivamente, o lado esquerdo (*Left Hand Side*) e o lado direito (*Right Hand Side*) da regra, definidos por conjuntos disjuntos de itens pertencentes aos exemplos indicando que as transações que contém LHS tendem a conter RHS (Agrawal et al. 1996). O lado esquerdo da regra é denominado antecedente da regra e o direito como conseqüente da regra.

3.2.1 Regra de associação

Uma regra de associação, de acordo com Agrawal and Srikant (1994), é definida como:

Seja $A=\{a_1, a_2, \dots, a_m\}$ um conjunto de literais chamados itens, D um banco de dados composto por um conjunto de transações $T=\{t_1, t_2, \dots, t_n\}$, onde cada tran-

sação t_i é composta por um conjunto de atributos de A , isto é, $t_i \subseteq A$. Os itens de cada transação t_i estão em ordem lexicográfica e n representa o número total de transações (exemplos). Um conjunto de itens $I \subset A$ é chamado de um *itemset*, ou seja, um *itemset* é um subconjunto I de A , ordenado lexicograficamente, representando a idéia de conjunto de itens adquiridos em uma transação. Uma transação t_n contém um *itemset* I se $I \subseteq t_n$. Considerando que LHS e RHS são *itemsets*, uma regra de associação é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$ e $LHS \cap RHS = \emptyset$. A regra $LHS \Rightarrow RHS$, ocorre no conjunto de transações em D com confiança = $conf$ se em $conf\%$ das transações em D que ocorre LHS , ocorre também RHS . A regra $LHS \Rightarrow RHS$ tem suporte = sup se em $sup\%$ das transações em D ocorre $LHS \cup RHS$.

Nessa definição, o suporte quantifica a incidência da regra no conjunto de dados, ou seja, indica a frequência com que LHS e RHS ocorrem juntos no conjunto de dados. É equivalente à probabilidade $p(LHS, RHS)$, isto é, a probabilidade de que LHS e RHS ocorram simultaneamente. A confiança indica a frequência com que LHS e RHS ocorram juntos em relação ao número total de registros em que LHS ocorre. É equivalente à probabilidade condicional $p(LHS|RHS)$, ou seja, representa a probabilidade de ocorrência de LHS , dado que RHS ocorre.

Em um processo de mineração de dados usando Associação, um elevado número de regras pode ser encontrado. As regras podem representar alguma particularidade de uma transação ou não serem relevantes diante do número total de exemplos fornecidos. Assim, para verificar a validade de uma regra, somente são consideradas as que apresentam suporte acima de um valor mínimo, denominado suporte mínimo. O mesmo também ocorre para a confiança. Um *itemset* é dito freqüente se possui valor de suporte maior ou igual ao suporte mínimo. Para realizar a mineração de regras de associação, é necessário, basicamente, encontrar as combinações de itens que possuem suporte acima do suporte mínimo e usar essas combinações para gerar as regras (Agrawal et al. 1996).

O exemplo a seguir ilustra esses conceitos. Considere o conjunto A de itens comprados em uma loja e o conjunto de transações T , representando as compras ocorridas, suporte e confiança mínima 50% como mostrado na Figura 3.3.

$A = \{\text{bermuda, calça, camiseta, sandália e tênis}\}$
 $T = \{\{\text{calça, camiseta, tênis}\}, \{\text{camiseta, tênis}\}, \{\text{bermuda, tênis}\}, \{\text{calça, sandália}\}\},$
 Suporte Mínimo = 50% (2 transações),
 Confiança Mínima = 50%

Figura 3.3: Dados de entrada para o processo de extração de regras de associação.

A representação dos itens comprados, referente a cada transação é mostrada na Tabela 3.2. Por exemplo, na transação 1 foi comprado *calça*, *camiseta* e *tênis*.

Tabela 3.2: Representação das transações e itens comprados.

Transações	Itens Comprados
1	calça, camiseta, tênis
2	camiseta, tênis
3	bermuda, tênis
4	calça, sandália

Tendo as transações disponíveis, extrai-se os *itemsets* a partir de combinações dos elementos do conjunto de itens A e de *itemsets* já encontrados. A formação de *itemsets* varia de acordo com os algoritmos utilizados. Nesse exemplo, para facilitar a compreensão do conceito de regras de associação e *itemsets*, são mostradas todas as combinações possíveis na Figura 3.4.

$$I = \{ \{bermuda\}, \{calça\}, \{camiseta\}, \{sandália\}, \{tênis\}, \{bermuda, calça\}, \{bermuda, camiseta\}, \{bermuda, sandália\}, \{bermuda, tênis\}, \{calça, camiseta\}, \{calça, sandália\}, \{calça, tênis\}, \{camiseta, sandália\}, \{camiseta, tênis\}, \{sandália, tênis\}, \{bermuda, calça, camiseta\}, \{bermuda, calça, sandália\}, \{bermuda, calça, tênis\}, \{bermuda, camiseta, sandália\}, \{bermuda, camiseta, tênis\}, \{bermuda, sandália, tênis\}, \{calça, camiseta, sandália\}, \{calça, camiseta, tênis\}, \{calça, sandália, tênis\}, \{camiseta, sandália, tênis\}, \{bermuda, calça, camiseta, sandália\}, \{bermuda, calça, camiseta, tênis\}, \{calça, camiseta, sandália, tênis\}, \{bermuda, calça, camiseta, sandália, tênis\} \};$$

Figura 3.4: Combinações possíveis dos itens comprados.

O suporte é calculado para cada *itemset*, mas somente são considerados os *itemsets* com suporte acima do suporte mínimo, como mostrado na Tabela 3.3. Considerando o suporte mínimo 50%, pode-se observar que existe a regra $\{camiseta\} \Rightarrow \{tênis\}$. O suporte dessa regra é igual a 50% e a confiança é calculada como sendo o suporte de ocorrer o *itemset* $\{camiseta, tênis\}$ dividido pelo suporte de ocorrer somente o *itemset* que forma a *LHS* da regra, ou seja, $\{camiseta\}$. Nesse caso, a confiança é 100%. Também é possível encontrar a regra $\{tênis\} \Rightarrow \{camiseta\}$, com suporte igual a 50%, mas com confiança igual a 66,6%, uma vez que o *itemset* $\{tênis\}$ ocorreu em outras transações que não ocorreu o *itemset* $\{camiseta\}$. As regras compostas somente por um item, como $\{Tênis\} \Rightarrow \text{verdadeiro}$, indicam que o item

foi comprado independentemente da compra de outro item, no caso de tênis comprado em 75% das transações.

Tabela 3.3: *Itemsets* freqüentes e seus respectivos suportes.

Conjunto de atributos	Suporte
{tênis}	75%
{calça}	50%
{camiseta}	50%
{camiseta, tênis}	50%

Existem na literatura diversos algoritmos para extrair regras de associação, entre eles: *Apriori* e *AprioriTid* (Agrawal and Srikant 1994); *Opus* (Webb 1995); *Direct Hasing and Pruning* (DHP) e *Dynamic Set Counting* (DIC) (Park et al. 1995); *Charm* (Zaki and Hsiao 2002); *FP-growth* (Han et al. 2000); *Closet* (Pei et al. 2000). Para auxiliar na compreensão dos algoritmos será apresentado o algoritmo *Apriori*, uma vez que este algoritmo é o mais citado na literatura, trabalhos envolvendo regras de associação e dados georreferenciados também utilizam esse algoritmo e o *Apriori* é o algoritmo utilizado neste trabalho.

3.2.2 O algoritmo *Apriori*

O algoritmo *Apriori*, proposto por Agrawal and Srikant (1994), busca encontrar os itens individuais e seus respectivos suportes para determinar quais são freqüentes, formando o primeiro conjunto de *itemsets*. No passo subsequente, esse conjunto é utilizado para gerar novos *itemsets*, denominados *itemsets* candidatos, acrescentando um item a cada elemento do conjunto. A seleção de *itemsets* freqüentes a partir dos candidatos é realizada por meio do cálculo do suporte para os candidatos, desprezando aqueles com suporte abaixo do mínimo. Esse processo é iterativo uma vez que os *itemsets* freqüentes encontrados darão origem aos novos *itemsets* candidatos até que não seja possível gerar novos *itemsets* freqüentes, pois o suporte de todos os futuros *itemsets* candidatos seriam inferior ao suporte mínimo.

O número k de itens em um *itemset* é denominado tamanho do *itemset* e um *itemset* de tamanho k é representado como um k -*itemset*. Os itens dentro de um *itemset* são mantidos em ordem lexicográfica. Um k -*itemset* é representado por c e o conjunto de k -*itemsets* freqüentes e k -*itemsets* candidatos são representados por L_k e C_k , respectivamente.

O algoritmo para geração de *itemsets* freqüentes é mostrado no Algoritmo 3.1. A entrada do algoritmo é constituída de um banco de dados D composto por conjunto m de itens $A=\{a_1, a_2, \dots, a_m\}$ ordenados lexicograficamente e por conjunto de n transações $T=\{t_1, t_2, \dots, t_n\}$ tal que $t_i \in T$ e $t_i \subseteq A$. Também é necessário informar o suporte mínimo *sup-min*.

Ao iniciar o processamento, o conjunto de *itemsets* contendo apenas um item é ge-

```

begin
   $L_1 = \{1\text{-itemset freqüente}\};$  // Conjunto de itemsets com 1 item
  for ( $k=2; L_{k-1} \neq \emptyset; k++$ ) do
     $C_k = \text{apriori-gen}(L_{k-1});$  // Novos itemsets candidatos
    forall (transação  $t \in D$ ) do
       $C_t = \text{subset}(C_k, t);$  // Candidatos contidos em  $t$ 
      forall (candidatos  $c \in C_t$ ) do
         $c.\text{contador}++;$ 
      end
    end
     $L_k = \{c \in C_k \mid c.\text{contador} \geq \text{sup-min}\};$ 
  end
  Resposta =  $\bigcup_k L_k;$ 
end

```

rado para ser a origem de novos *itemsets*. Isso é feito contando as ocorrências de itens no banco de dados e calculando o suporte para cada item encontrado, gerando o conjunto de 1-*itemset* freqüente, L_1 , utilizado na primeira iteração para encontrar os *itemsets* candidatos C_2 .

O passo seguinte consiste de duas fases, executadas até não ser possível gerar novos *itemsets* devido ao suporte mínimo. Na primeira fase, os *itemsets* freqüentes L_{k-1} são usados para gerar o conjunto de *itemsets* candidatos C_k usando a função **Apriori-gen** descrita no Algoritmo 3.2. Na segunda fase, o banco de dados é percorrido e para cada transação t_i , $t_i \in T$, é verificado se a transação contém *itemsets* candidatos. Se satisfeita a condição, o suporte do *itemset* candidato é incrementado. Isso é feito gerando um subconjunto C_t que possui somente k -*itemsets* da transação e verificando se esses k -*itemsets* estão presentes também em C_k . A tarefa de gerar o conjunto C_t é realizada pela função **Subset**. Todos os suportes para os *itemsets* candidatos são calculados e atualiza-se o conjunto de *itemsets* freqüentes L_k desprezando aqueles candidatos com suporte menor que o mínimo, pois os *itemsets* freqüentes L_{k-1} que não obtiverem o valor superior ao suporte mínimo no nível $k-1$, também não vão possuir suporte maior que o mínimo no nível k .

A função **Apriori-gen**, mostrada no Algoritmo 3.2, tem como argumento L_{k-1} , o conjunto de todos os $(k-1)$ -*itemsets* freqüentes e retorna um conjunto de todos os k -*itemsets* candidatos. Primeiro, no passo de associação, relaciona-se L_{k-1} com L_{k-1} para ser possível encontrar novos *itemsets*. Nessa etapa já é possível eliminar alguns *itemsets* realizando uma filtragem no conjunto C_k . A eliminação de alguns elementos é válida, pois ao se combinar os $(k-1)$ -*itemsets* freqüentes pode-se gerar *itemsets* com combinações de itens que não estão representados nas transações do banco de dados ou que não possuem suporte mínimo. Essa situação pode ser verificada olhando os possíveis subconjuntos s constituído de $(k-1)$ -

```

begin
  insert into  $C_k$ 
  select  $p.item_1, p.item_2, \dots, p.item_{k-1}, q.item_{k-1}$ 
  from  $L_{k-1} p, L_{k-1} q$ 
  where  $p.item_1 = q.item_1, \dots, p.item_{k-2} = q.item_{k-2}, p.item_{k-1} < q.item_{k-1}$ ;
  /* É realizada a etapa de poda, em que todo itemset  $c \in C_k$  é
     removido se algum  $(k-1)$ -subconjunto de  $c$  não pertencer a  $L_{k-1}$  */
  forall (itemset  $candidatoc \in C_k$ ) do
    forall ( $(k-1)$  - subconjuntos  $s$  do itemset  $candidatoc$ ) do
      if ( $s \notin L_{k-1}$ ) then
        delete  $c$  de  $C_k$ 
      end
    ;
  end
end
end

```

itemsets formados a partir de um *itemset* candidato c , $c \in C_k$, e verificando se os *itemsets* de s estão entre os elementos de L_{k-1} . Assim, são eliminados todos os *itemsets* candidatos $c \in C_k$ tal que algum $(k-1)$ -subconjunto de c não esteja em L_{k-1} .

A função **Subset** do algoritmo *Apriori* retorna os k -*itemsets* candidatos que estão contidos em uma dada transação t_i . Os *itemsets* candidatos C_k são armazenados em uma *hash-tree* em que cada nó da árvore pode conter uma lista de *itemsets* ou uma *hash-table* (nó folha e nó intermediário, respectivamente). Partindo do nó raiz, a função encontra todos os *itemsets* candidatos presentes na transação t_i . Se um nó folha é atingido e o *itemset* encontrado está contido na transação t_i , uma referência é adicionada ao conjunto de resposta. Se um nó intermediário é atingido a partir de um item a_j , pesquisa-se em cada item após a_j em t_i . Isso é possível porque os itens estão em ordem lexicográfica. No nó raiz, todos os itens a_j em t_i são pesquisados.

Por fim, o algoritmo para geração de regras proposto por Agrawal et al. (1996) tem como entrada o conjunto de *itemsets* freqüente L_K e a medida confiança mínima *conf-min*. Para gerar as regras, considerando todos os *itemsets* freqüentes l , $l \in L_K$, são encontrados subconjuntos a não vazios de l . Para cada subconjunto a , a saída é uma regra na forma $a \Rightarrow (l - a)$ tal que a proporção $\text{suporte}(l)/\text{suporte}(a)$ seja no mínimo igual a confiança mínima. Esse algoritmo é apresentado no Algoritmo 3.3.

Os conceitos vistos podem ser estendidos para serem utilizados com dados espaciais. Na próxima seção são discutidos alguns trabalhos que já utilizaram regras de associação com dados espaciais.

```

begin
  forall (k-itemsets frequente  $l_k$ ,  $k \geq 2$ ) do
    | call genrules ( $l_k$ ,  $l_k$ );
  end
end

/* genrules gera todas as regras válidas  $\tilde{a} \Rightarrow (l_k \tilde{a})$ , para todo  $\tilde{a} \subset a_m$  */
Procedure genrules (  $l_k$ ,  $a_m$ )
begin
   $A = \{(m-1)\text{-itemset } a_{m-1} | a_{m-1} \subset a_m\}$ ;
  forall ( $a_{m-1} \in A$ ) do
     $conf = \sup(l_k) / \sup(a_{m-1})$ ;
    if ( $conf \geq conf\text{-}min$ ) then
      OUTPUT: regra  $a_{m-1} \Rightarrow (l_k a_{m-1})$ , confiança= $conf$  e suporte= $\sup(l_k)$ ;
      if ( $m - 1 > 1$ ) then
        | call genrules( $l_k$ ,  $a_{m-1}$ );
      end
    end
  end
end
end

```

3.3 Regras de associação espacial

O conceito de regras de associação espacial é semelhante ao de regras de associação. É uma implicação na forma $LHS \Rightarrow RHS$, com suporte e confiança, significando que se um padrão LHS aparece em um dado exemplo, então o padrão RHS também aparecerá, com suporte e confiança. A diferença entre as técnicas está na inclusão de relações espaciais no antecedente ou conseqüente da regra. Uma regra de associação espacial é uma regra que descreve a implicação de uma característica espacial por outra, em um banco de dados espaciais (Koperski and Han 1995). Essa característica pode estar descrita em termos de atributos ou relações espaciais. Considerando que é possível encontrar essas relações como se fossem itens em uma transação, utiliza-se o termo predicado para informar que o item se trata de uma relação. Se a relação é espacial, o predicado também é definido como um predicado espacial. O formato geral de representação é o nome do predicado seguido dos identificadores de objetos relacionados entre parênteses. Por exemplo, $dentro(X, Y)$ e os objetos relacionados X e Y . Pode-se utilizar o termo predicado para se referir a itens que não são relações, por exemplo, atributos descritivos de objetos espaciais, considerando que existe uma relação entre o objeto espacial e o valor que o atributo assume. Assim a relação pode ser representada por um predicado constituído do nome do atributo e entre parênteses o identificador do objeto e o valor assumido. Por exemplo, uma cidade X com uma população

de 100 mil habitantes pode ser representada com o predicado $população(X,100)$.

Uma regra de associação espacial, de acordo com Koperski and Han (1995), é formada por:

$P = \{p_1, \dots, p_m\} \Rightarrow Q = \{q_1, \dots, q_n\}$ (sup% conf%), na qual pelo menos um dos predicados $p_1, \dots, p_m$ ou $q_1, \dots, q_n$ é um predicado espacial.

Como em regras de associação, também é considerado um suporte mínimo e uma confiança mínima. Regras com suporte ou confiança abaixo dos valores mínimos são desprezadas.

A idéia de *itemset* permanece, sendo que um *itemset* deve conter um conjunto de predicados em que pelo menos um dos predicados é espacial. O conceito de *itemset* freqüente pode ser estendido para representar um conjunto de predicados freqüentes, sendo os predicados que possuem suporte acima do suporte mínimo. Um conjunto de predicados é chamado de 1-predicado se o conjunto é composto por um predicado somente, por exemplo, o conjunto $\{igual(x,y)\}$. Um conjunto de k predicados é denominado de k -predicado. Por exemplo, $\{igual(x,y), dentro(x,z)\}$ é um 2-predicado.

Na próxima seção é descrito o primeiro algoritmo proposto para extração de regras de associação espaciais.

3.3.1 Algoritmo para extração de regras de associação espacial

O primeiro algoritmo a tratar de regras de associação espacial foi proposto por Koperski and Han (1995), tendo como base uma abordagem de refinamento progressivo, considerando uma etapa intermediária de extração de relações topológicas. O algoritmo original também considera que os objetos espaciais podem estar associados a hierarquia de conceito, mas para simplificar a compreensão do algoritmo essa característica foi desprezada. Existem cinco etapas a serem executadas pelo algoritmo, a seleção dos objetos no banco de dados, a busca por relações de proximidade, filtragem dos resultados de acordo com o suporte, especialização das relações de proximidade em relações topológicas e por fim, a obtenção das regras de associação espacial. Essas etapas podem ser visualizadas na Figura 3.5.

```
objetos_relevantes = extração_objetos_pertinentes(SDB,RDB);
objetos_proximos = computação_relação_proximidade(objetos_relevantes);
predicados_frequentes = encontra_predicados_frequentes_e_regras(objetos_proximos);
predicados_filtrados = refina_predicados(predicados_frequentes);
encontra_predicados_frequentes_e_regras(predicados_filtrados);
```

Figura 3.5: Etapas para geração de regra de associação espacial. Fonte: (Koperski and Han 1995)

O algoritmo para realizar essas etapas tem como entrada os valores do suporte mínimo e confiança mínima, um banco de dados espaciais constituído de objetos espaciais com dados descritivos (*RDB*) e dados de localização (*SDB*), armazenados em um banco de dados separado. Também é entrada do algoritmo uma consulta, que informa quais os objetos serão utilizados na mineração de dados e, desses objetos, qual é o objeto principal a ser comparado com os outros, quais são os relacionamentos entre esses objetos, a distância entre os objetos a serem relacionados e o objeto principal. Também é possível definir algumas características específicas dos objetos. Um exemplo de consulta é ilustrado na Figura 3.6, extraído de Koperski (1999). Nessa figura, a linha 1 informa qual o objeto de interesse, nesse caso *idades_grandes*. São informadas as relações topológicas entre objetos na linha 2. Na linha 3 informam-se os objetos espaciais a serem utilizados. No final, na linha 4, define-se critérios particulares de seleção dos objetos e também a relação de proximidade entre o objeto de interesse e os outros objetos espaciais.

```

1 Mine spatial association describing cidades_grandes
2 Whith respect to topologia(T.geo, R.geo), R.nome, topologia(T.geo, W.geo),
  W.nome, topologia(T.geo, M.geo), M.nome, topologia(T.geo, B.geo), B.admin_regiao
3 From cidade T, rodovia R, hidrografia W, mina M, fronteira B
4 Where T.tipo = grande and distancia(T.geo, R.geo) < 5KM and
  distancia(T.geo, W.geo) < 5KM and distancia(T.geo, M.geo) < 30KM and
  distancia(T.geo, B.geo) < 80KM and R.tipo = vicinal AND W.tipo
  IN (mar, oceano, lagos_grande, rios_grandes).

```

Figura 3.6: Consulta de entrada do algoritmo de mineração de regras de associação espacial. Adaptado de: (Koperski 1999)

As etapas para extração de regras de associação espacial são detalhada nas próximas seções.

Seleção dos objetos no banco de dados

Na primeira etapa, os dados do banco de dados espaciais que serão utilizados na mineração são recuperados. É necessário informar uma consulta ao banco de dados para selecionar somente os objetos espaciais apropriados para o processamento, construindo novos conjuntos de dados, um conjunto para cada grupo de objeto informado na consulta de entrada e um conjunto de objetos de interesse. O resultado dessa etapa é o banco de dados espaciais *objetos_relevantes*, contendo conjuntos dos objetos espaciais selecionados. Esse processamento está denominado como *extração_objetos_pertinentes*. Também, nessa etapa, os predicados são gerados. Para cada atributo de um objeto espacial é gerado o predicado específico do atributo, com o nome do atributo, o identificador do objeto e o valor que o atributo assume para aquele objeto.

Busca por relações de proximidade

Após reunir os objetos a serem utilizados na mineração de regras de associação espacial, são encontrados todos os objetos que apresentam relação de proximidade. Essa etapa é executada na função `computação_relação_proximidade`, que tem como entrada o banco de dados espaciais encontrado no passo anterior e as distâncias entre o objeto de interesse e os demais objetos. Essa função encontra os objetos próximos do objeto de interesse. Isso pode ser feito considerando qualquer relação topológica como uma relação de proximidade. Assim, a relação de proximidade é mais geral que as relações topológicas. Dessa forma, encontrando só um tipo de relação é possível filtrar os objetos para somente aplicar as outras relações em um número reduzido de objetos. A extração de relações de proximidade pode ser realizada usando métodos de acesso espacial como o *Minimum Bounding Rectangle* (MBR) que agrupa pontos e polígonos nos menores retângulos possíveis e todos os elementos em um mesmo retângulo são elementos próximos. A estrutura de dados R-tree e suas derivações são utilizadas para informar os MBR. Alguns bancos de dados espaciais permitem realizar operações de consultas utilizando relações de distância como condição da consulta (Brinkhoff et al. 1993). O resultado da função `computação_relação_proximidade` é um banco de dados espaciais contendo um conjunto de objetos de interesse próximos aos outros objetos espaciais selecionados, representado por *objetos_proximos*. Além disso, o banco de dados espaciais resultante da função também contém a relação de proximidade, no formato de predicado composto pelo nome da relação de proximidade e os identificadores dos objetos espaciais definidos na consulta de entrada.

Filtragem dos resultados de acordo com o suporte

A próxima etapa é a filtragem dos dados com a finalidade de reduzir o número de dados utilizados para procurar relações topológicas. Para isso, são gerados os k -predicados frequentes e contado o suporte para os mesmos. O procedimento é semelhante ao *Apriori* (Algoritmo 3.1), considerando-se como *itemsets* os predicados espaciais. A função `encontra_predicados_frequentes_e_regras` representa a etapa da filtragem, que tem como entrada o banco de dados espaciais gerado na etapa de busca por relações de proximidade. Além dessa entrada, é necessário informar o valor do suporte mínimo e da confiança mínima. O algoritmo também pode gerar regras em função da relação de proximidade uma vez que utiliza o princípio do algoritmo *Apriori* para selecionar os predicados com suporte acima do mínimo. Além das regras, tem-se o resultado dessa função, *predicados_frequentes*, em um banco de dados espaciais contendo somente os objetos que possuem predicados espaciais contidos no conjunto de k -predicados frequentes.

Especialização das relações de proximidade em relações topológicas

Ao eliminar parte dos dados, torna-se menos custoso computacionalmente a busca por todas as relações topológicas. Assim, a próxima etapa é especializar as relações de proximidade encontrando relações topológicas nos dados que restaram da filtragem. A função `refina_predicados` simboliza o passo em que as relações de proximidade são especializadas em relações topológicas, refinando os resultados obtidos na etapa de filtragem. Para isso, consultas espaciais ao banco de dados de entrada são necessárias. Considerando todas as relações topológicas, para cada relação de proximidade é feita uma nova consulta.

Extração de regras de associação espacial

Por fim, na função `encontra_predicados_frequentes_e_regras`, novamente é empregado o algoritmo *Apriori* para a extração das regras de associação espacial. Os dados encontrados no refinamento são passados para a etapa de mineração para a extração das regras.

Para a melhor compreensão do algoritmo, na próxima seção será apresentado um exemplo de mineração de regras de associação espacial.

3.3.2 Exemplo de mineração de regras de associação espacial

Considere o banco de dados de British Columbia (Koperski 1999). Esse banco de dados espaciais possui tabelas referentes a cidades, rodovias, hidrografia, minas e fronteira, mostradas na Figura 3.7. O atributo *geo* representa informações de localização, latitude e longitude.

cidade(nome, tipo, população, geo)
rodovia(nome, tipo, geo)
hidrografia(nome, tipo, geo)
minas(nome, tipo, geo)
fronteira(tipo, administrador)

Figura 3.7: Descrição das tabelas sobre British Columbia. Fonte: (Koperski 1999)

A consulta de entrada é a apresentada na Figura 3.6. Os dados referente a cláusula *From* da consulta se referem aos objetos do banco de dados espaciais que serão relacionados com o objeto de interesse, nesse caso, definido como grandes cidades. Os dados são selecionados e agrupados de acordo com as relações obtidas na cláusula *Where*. No exemplo, são utilizados dados referentes a grandes cidades, rodovias com estrada vicinal, hidrografia referente a mar, oceano, lagos, rios, qualquer mina e qualquer fronteira. Esse processamento é realizado em `extração_objetos_pertinentes`.

A seguir, a função `computação_relação_proximidade` é executada para encontrar as relações de proximidade. Na Tabela 3.4 são ilustradas algumas relações obtidas. Cada linha da tabela representa uma cidade próxima a outros objetos descritos nas colunas.

Tabela 3.4: Exemplos de relações de proximidade entre objetos. Fonte: (Koperski 1999)

Cidade	Hidrografia	Rodovia	Fronteira
victoria	j_f_strait	highway_1, highway_17	us
saanich	j_f_strait	highway_1, highway_17	us
pentincton	okanagan_lake	highway_97	us
...	...	...	...

Na próxima etapa o suporte é calculado para os predicados encontrados e o suporte mínimo, *sup_min* é utilizado para selecionar os predicados freqüentes. Esses predicados são utilizados como base para encontrar as relações topológicas. Na Tabela 3.5 são apresentadas as relações topológicas encontradas.

Tabela 3.5: Relacionamento espacial detalhado para os predicados freqüentes. Fonte: (Koperski 1999)

Cidade	Hidrografia	Rodovia	Fronteira
victoria	(adjacent_to, j_f_strait)	(intersects, highway_1), (intersects,highway_17)	(close_to,us)
Saanich	(adjacent_to, j_f_strait)	(intersects, highway_1), (intersects,highway_17)	(close_to,us)
Pentincton	(adjacent_to,okanagan_lake)	(intersects,highway_97)	(close_to,us)
...	...	...	...

As regras de associação espacial são extraídas na última etapa. Por exemplo, a regra $is_a(X, cidades_grande), adjacent_to(X, j_f_strait) \Rightarrow intersects(X, highway_1)$ (66%, 100%), representa que 66% das grandes cidades que são adjacentes a *j_f_strait* tem intersecção com *highway_1*.

Esse exemplo foi utilizado para auxiliar na compreensão do Algoritmo proposto por Koperski and Han (1995). Outras abordagens para regras de associação espacial foram propostas como a utilização de programação lógica indutiva (ILP).

3.3.3 Estudos relacionados a regras de associação espacial

A abordagem mostrada anteriormente foi definida considerando a entrada no formato atributo-valor, apesar de definir predicados. O processo de mineração considera os predicados como *itemsets*. Malerba et al. (2001) propuseram um método utilizando programação lógica indutiva para extrair regras de associação espacial. O método possibilita a utilização de conhecimento de fundo em um domínio espacial. Foi implementado em um sistema denominado SPADA (*Spatial Pattern Discovery Algorithm*), o qual opera com um banco de

dados dedutivo (DDB, do inglês *Deductive Relational Databases*). A idéia básica da abordagem ILP é que um banco de dados espaciais pode ser representado por um banco de dados dedutivo uma vez que os objetos espaciais, as propriedades espaciais e as relações espaciais podem ser extraídas de acordo com semânticas pré-definidas.

Li et al. (2003) propõem técnicas de interpolação para mineração de regras de associação espacial para tratar o problema de risco de estiagem, tentando descobrir padrões com parâmetros oceânicos, índices atmosféricos e dados de clima. As regras refletem as influências de parâmetros oceânicos no clima. Como não é possível ter disponível pontos de todas as regiões, é necessária a utilização de interpolação espacial, para encontrar funções que representem melhor a região. A diferença desse trabalho com os anteriores é a definição de regras de associação espacial utilizada, a qual não considera as relações topológicas, sendo classificada como regras de associação espacial por trabalhar com dados espaciais sobre clima e utilizar a localização para encontrar novos dados.

Mennis and Liu (2003) apresentam uma aplicação utilizando mineração de regras de associação e dados geográficos, especificamente, dados sócio-econômico da cidade de Denver, nos Estados Unidos. Para tal aplicação foi utilizado o processamento espacial de um sistema de informação geográfica para realizar a intersecção de várias áreas. O resultado foi transferido para uma tabela utilizada como entrada de um algoritmo de extração de regras de associação, o *Apriori*. Considerando a definição de regras de associação espacial, esse trabalho apesar de ter realizado uma operação de intersecção nos dados, não apresenta na saída das regras uma relação espacial, não atendendo a definição de regras de associação espacial. Porém, essa relação fica subentendida, pois os dados foram obtidos por meio dessa relação. O estudo desses trabalhos originou o projeto de mestrado desenvolvido, apresentado no próximo capítulo.

Extração de regras de associação em dados georreferenciados

Com a grande quantidade de dados coletada e armazenada em sistemas de informações geográficas seria possível encontrar novos padrões de ocorrência de um determinado evento e auxiliar na solução de problemas envolvendo o evento. Um exemplo de descoberta de padrão utilizando localização espacial ocorreu em Londres no século XIX. Um especialista em epidemias marcou em um mapa todas as localizações onde as vítimas da cólera moravam. A localização formou um agrupamento e o centro do agrupamento revelou-se ser uma bomba de água. Com isso, as autoridades governamentais desligaram a bomba e a epidemia de cólera foi controlada (Chawla et al. 1999). Assim, encontrar padrões pode ser muito útil na solução de problemas envolvendo dados espaciais e aplicar técnicas de mineração de dados pode auxiliar a busca por padrões.

O grande desafio em estudar mineração de regras de associação em dados espaciais é encontrar uma boa maneira de organizar as relações topológicas e os dados. Trabalhar com relações implícitas em dados de qualquer natureza é um campo de pesquisa que ainda oferece muitas atividades para serem realizadas. A maioria das abordagens de mineração de dados consideram que os exemplos utilizados já contem toda a informação para a extração do conhecimento, não dando a devida importância a dados que por sua própria definição possuem algum relacionamento a ser utilizado como mais uma informação na geração do conhecimento. Programação lógica indutiva é uma das poucas tecnologias que permitem especificar um conhecimento de fundo, o qual poderia representar esses relacionamentos, mas ainda é necessário um mecanismo para encontrar esses relacionamentos. Somente após

especificar as relações entre os dados é possível utilizá-las como entrada ou conhecimento de fundo. No contexto de dados georreferenciados, pode-se implementar mecanismos que consideram as relações topológicas implícitas nos dados. O problema recai em como organizar as relações para serem genéricas e aproveitadas em diferentes abordagens de mineração de dados. Um mecanismo de pré-processamento de relações topológicas é de fundamental importância para aplicação e pesquisa de mineração de dados em dados georreferenciados.

As abordagens apresentadas do uso de mineração de regras de associação espacial propuseram algoritmos eficientes para a extração de regras. Contudo Koperski (1999) propõe um algoritmo de mineração de regras de associação espacial que não utiliza todos os dados de entrada para encontrar as regras de associação espacial. Em uma etapa inicial são extraídas relações de proximidade. Somente relações com alta incidência no banco de dados são especializadas em relações topológicas para gerar regras de associação espacial. A relação intermediária de proximidade foi utilizada visando melhorar o desempenho do sistema, porém é possível encontrar as relações topológicas diretamente.

Outro ponto observado é que os trabalhos anteriores não apresentaram resultados em função de atributos descritivos, somente regras em função de relações espaciais. Não é possível associar os atributos descritivos de diferentes objetos espaciais ligados por meio de uma relação espacial. Somente é possível encontrar associações entre os atributos de um objeto espacial e a relação existente entre este objeto e outros objetos espaciais.

Por fim, restou a dúvida se não seria possível somente pré-processar as relações topológicas, definindo uma organização dos dados para que qualquer algoritmo de mineração de regras de associação possa ser executado. Assim, seria possível aproveitar as ferramentas de avaliação de regras disponíveis utilizadas para avaliar os resultados. Se isso for praticável, simplificaria a extração de regras em dados georreferenciados e o estudo poderia ser expandido para aplicações de mineração de dados gerais.

4.1 *Objetivo*

O objetivo principal desse trabalho é pré-processar dados georreferenciados de maneira que não seja necessário a alteração dos algoritmos de mineração de regras de associação. Além das relações topológicas, também devem ser considerados os atributos não espaciais. Para isso, deve-se explorar e especificar um formato que permita encontrar regra de associação com características espaciais, sem a obrigatoriedade de conversão dos atributos não espaciais em predicados. Assim, é necessário desenvolver um módulo de pré-processamento de dados espaciais para simplificar a extração de regras de associação em dados georreferenciados.

Para alcançar esse objetivo, a metodologia utilizada é a pesquisa aplicada, visando explorar e manipular conhecimentos para a aplicação prática auxiliando na solução de problemas

de uma determinada área. Neste caso, o módulo de pré-processamento de relações topológicas desenvolvido é aplicado à área específica de planejamento urbano. O intuito é extrair as relações para aplicar um algoritmo de mineração de regras de associação, a fim de descrever fatos ou fenômenos armazenados no banco de dados espaciais. Essa descrição é realizada quantificando as ocorrências das relações topológicas entre os objetos espaciais. Para verificar a eficácia do módulo, um estudo de caso é realizado sobre o sistema de informação geográfica da cidade de Jaboticabal, descrito no Capítulo 5, e os resultados encontrados analisados por um especialista em planejamento urbano. O processo de extração de regras de associação espacial utilizado está descrito na próxima seção.

4.2 Processo de extração de regras de associação em dados georreferenciados

A extração de regras de associação em dados georreferenciados proposta neste trabalho é baseada nas etapas do processo de mineração de dados. Essas etapas foram adaptadas para originar o processo sugerido. Pode-se visualizar essas etapas para dados georreferenciados na Figura 4.1.

No pré-processamento, diversas atividades para compreensão e organização dos dados são necessárias. Os dados espaciais, em geral, estão armazenados em estruturas proprietárias com referência em tabelas do banco de dados espaciais e os atributos descritivos nas tabelas do banco de dados. Assim, a primeira atividade proposta para o pré-processamento, após a compreensão da estrutura de organização dos dados, é desnormalizar as tabelas para os dados expressarem a realidade do usuário. Isto é, trocar os códigos utilizados para diminuir o tamanho do banco de dados pelos significados reais dos atributos. As tabelas desnormalizadas podem ser combinadas para expressarem conhecimento de um determinado evento. Por exemplo, considere uma tabela representando um lote de um bairro e uma tabela com dados de cada domicílio originados de pesquisas de censo. No geral essas características são armazenadas em tabelas diferentes e podem ser combinadas para expressarem as características físicas dos lotes como residências e as características sociais.

Outra atividade proposta para ser executada no pré-processamento é a seleção de atributos interessantes, uma vez que muitos atributos podem não estar relacionados com o objetivo da extração de padrão. No caso de atributos com valores ausentes é necessário eliminar o atributo devido aos poucos valores disponíveis da característica descritiva, com o intuito de garantir um resultado confiável. Também é necessário eliminar os atributos que somente apresentam um valor para todos os objetos espaciais pois quando todos os valores são iguais

Figura 4.1: Adaptação do processo de mineração de dados para mineração de dados georreferenciados

somente será gerado padrões óbvios em relação a essa característica. Por fim, muitas características dos objetos espaciais são descritas por valores numéricos e cada objeto pode apresentar um valor diferente. Nesses casos, é essencial discretizar esses valores em intervalos para facilitar a verificação de padrões.

Além disso, é no pré-processamento dos dados que serão encontradas as relações topológicas. Estas devem ser encontradas verificando qual o relacionamento entre os objetos espaciais de determinado plano de informação com os objetos de outro plano de informação. Como é feita a combinação de todos os objetos é possível demandar um alto tempo de execução mas isso não bloqueia o procedimento. Somente seria inviável se esse procedimento demandasse um alto consumo de memória. Nesse caso os dados devem ser novamente pré-processados e verificado se não é possível utilizar amostras dos dados ao invés de todo o conjunto. Após a extração de relações topológicas, os dados devem ser colocados em um formato que permite a execução de algoritmos de mineração de regras de associação para gerar regras que contenham obrigatoriamente relações topológicas. Esses formatos são apresentados na seção 4.3. A integração entre a etapa de extração de relações topológicas e a aplicação do algoritmos de mineração de regras de associação é feita por meio de arquivo texto no formato de entrada do algoritmo.

Assim, pode-se executar algoritmos tradicionais para gerar as regras. Após, pode ser necessária a aplicação de filtros para eliminar regras desnecessárias ou utilizar medidas de avaliação para auxiliar na busca por regras novas e válidas. As medidas utilizadas neste trabalho para avaliar as regras foram a confiança e o suporte. Uma descrição detalhada do módulo implementado para extrair relações topológicas na etapa de pré-processamento é encontrada na próxima seção.

4.3 Módulo de extração de relações topológicas

Para ser utilizado o módulo de extração de relações topológicas, proposto neste trabalho, é considerado que os dados georreferenciados estão armazenados em banco de dados espaciais, para facilitar o pré-processamento dos mesmos e a reutilização de ferramentas implementadas. A etapa anterior a execução do módulo é a escolha, o pré-processamento dos dados e importação para o banco de dados *Microsoft Access* por meio da ferramenta TerraView. Uma visão geral do módulo de extração de relações topológicas pode ser vista na Figura 4.2.

O módulo de extração de relações topológicas foi projetado considerando que o usuário informa dois planos de informação com os dados a serem minerados e quais atributos desses PIs serão utilizados. As relações trabalhadas são as relações topológicas disjunção, toque (adjacência), cruzamento, sobreposição, inclusão, intersecção, igualdade, cobertura. Também foi utilizada a relação métrica de proximidade. Essas relações são encontradas realizando uma seleção no banco de dados, cruzando todos os dados de um plano de informação, informado pelo usuário, com todos os dados de um segundo plano de informação para cada relação. A relação de região próxima é tratada diferentemente das outras. O usuário precisa informar uma distância para ser calculada a área considerada como região próxima, isto é, a circunferência com o raio igual a distância informada e centro coincidente com o centro do objeto espacial. Após encontrar cada relação, os dados são convertidos para um arquivo com um dos formatos de saída propostos neste trabalho.

Os métodos usados para implementar o módulo e realizar a comunicação com o banco de dados pertencem a biblioteca TerraLib. Essa biblioteca permite o acesso a diversos bancos de dados, porém no módulo foi considerado inicialmente o banco de dados MySQL e posteriormente o banco de dados *Microsoft Access* devido a problemas de compatibilidade entre as versões do MySQL e a biblioteca TerraLib. Também possui métodos de importação de dados para diferentes formatos, sendo o principal para arquivos *Shapefile*, um formato disponível em vários sistemas para o intercâmbio de dados. Porém a tentativa de utilizar os métodos de importação da biblioteca não foi bem sucedida. Ocorreram problemas entre a interligação da biblioteca e a versão do MySQL utilizada. Nessa etapa o banco de dados *Microsoft Access* ainda não havia sido considerado. No final do projeto, quando o módulo já

estava implementado e gerava problemas de alocação de memória, foi substituído o banco de dados MYSQL pelo *Microsoft Access*.

Figura 4.2: Módulo de extração de regras de associação espacial

Com a finalidade de solucionar o problema da importação dos dados, foi utilizado o software TerraView com o objetivo de transformar os dados no formato necessário para que os métodos da biblioteca TerraLib possam acessar e armazenar esses dados no banco de dados. Com isso, foi possível a utilização dos métodos da biblioteca para encontrar as relações topológicas entre os planos de informação.

Ao obter os dados no formato aceito pelo TerraLib, foi possível implementar o algoritmo de extração de relações topológicas para pré-processar os dados espaciais. O algoritmo, proposto neste trabalho, para extração de relações topológicas pode ser observado no Algoritmo 4.1. Para a relação de proximidade, além dos passos descritos no algoritmo, deve-se encontrar objetos em um raio informado pelo usuário, utilizando a consulta espacial específica para essa situação e definida na biblioteca. Como esse algoritmo relaciona todos os a objetos de uma tabela com todos os n objetos de outra, a operação predominante é uma comparação em um *loop* alinhado, comparando a objetos pertencentes a P_1 com n objetos pertencentes a P_2 . Assim, esse algoritmo possui complexidade exponencial, da ordem de a^n .

O algoritmo implementado considera como entrada um conjunto de h relações topológicas $R = \{r_1, ..., r_h\}$ escolhido pelo usuário, entre as possibilidades de relações topológi-

```

begin
  insert into  $O_1$  // Encontrar todos os objetos do primeiro PI
  select  $pi_1.o.geo, pi_1.o.a_1, \dots, pi_1.o.a_q$ 
  from  $pi_1$ 
  where  $pi_1.o.a_1=b_1$  and ... and  $pi_1.o.a_q=b_q$  ;
  forall (objeto  $pi_1.o \in O_1$ ) do
    forall (relação  $r \in R$ ) do
      insert into  $O_2$  // Processar consulta espacial
      select  $pi_2.o.geo, pi_2.o.a_1, \dots, pi_2.o.a_v$ 
      from  $P_2$   $pi_2$ 
      where  $\exists r$  tal que  $r(pi_1.o.geo, pi_2.o.geo) \neq \emptyset$  and
       $pi_2.o.a_1=c_1$  and ... and
       $pi_2.o.a_q=c_v$ ;
      forall (objeto  $u \in O_2$ ) do
        OUTPUT: write  $o$ , write  $u$ , write  $r$  // Grava arquivo
      end
    end
  end
end
end

```

cas disponíveis. Também é entrada um banco de dados D composto por conjuntos de n planos de informação $P=\{p_1, \dots, p_n\}$, em que cada plano de informação é constituído de um conjunto de m objetos espaciais $O=\{o_1, \dots, o_m\}$. Cada objeto espacial é formado por um conjunto de t atributos $A=\{a_1, \dots, a_t, geo\}$. *Geo* representa as informações de geometria e a localização espacial. Esse atributo é uma ligação entre tabelas uma vez que os dados espaciais e os dados descritivos, no geral, são armazenados em tabelas diferentes. Dentre todos os planos de informação pertencentes ao banco de dados, o usuário escolhe dois planos específicos P_1 e P_2 (tal que $P_1 \in D$ e $P_2 \in D$) para serem utilizados. Por fim também é informado quais atributos de cada plano de informação são interessantes para serem utilizados na extração. Esses atributos são representados por um conjunto de atributos $B=\{b_1, \dots, b_q\}$ tal que $B \subset A$ para P_1 e um conjunto de atributos $C=\{c_1, \dots, c_v\}$ tal que $C \subset A$ para P_2 .

No início do processamento do módulo, encontra-se todos os objetos espaciais pertencentes à P_1 , utilizando uma consulta ao banco de dados de maneira que sejam retornados somente os atributos desejados para cada objeto. A seguir, para cada objeto encontrado e para cada relação topológica, é computada uma consulta espacial com resultado formado por todos objetos de P_2 que atendem as condições da relação topológica. Para cada um dos objetos relacionados são retornados os atributos desejados pelo usuário. Esses objetos e a relação topológica são a base para gerar a saída do módulo.

O arquivo de saída depende do formato de entrada dos algoritmos de mineração de regras de associação. O usuário deve informar esse formato para executar o módulo de extração

de relações topológicas. Existem formatos específicos para cada algoritmo de mineração de regras de associação e a idéia inicial do projeto era converter somente para o formato do ambiente DISCOVER ¹(Batista 2003) (Prati 2003). Esse ambiente possui mecanismos já implementados para pré-processamento de dados e conversão para formatos de diferentes algoritmos. Contudo, esse ambiente tem como foco as tarefas de Classificação e Regressão. Regras de associação para dados georreferenciados apresentam particularidades que dificultam o reaproveitamento desse formato para o pré-processamento dos dados e extração das regras de associação. Por exemplo, o formato de entrada prevê que um valor de atributo desconhecido é preenchido com uma marcação especial, no caso, "?". Dessa forma é necessário conhecer previamente todos os valores de atributos e quando um determinado valor não existir informar essa situação. Não é possível omitir esse dado. Só que em regras de associação para dados georreferenciados os atributos descritivos, seus valores e as relações topológicas podem ser muito diversificados, existindo por exemplo, um objeto espacial relacionado com um outro objeto ou com centenas de objetos. Nem sempre é viável estimar todas as relações e todos os objetos participantes da relação com a finalidade de transformar em atributos e valores pré-definidos dos algoritmos de mineração. Assim, um formato mais adequado prevê essa diversidade, como o formato utilizado pelo algoritmo *Apriori*. Nesse formato não é necessário informar previamente quais são os atributos. Essa informação está contida em cada transação do arquivo de entrada do algoritmo e quando um valor não existe, não é colocado na transação.

O estudo de como organizar esses dados para melhor aproveitar a tarefa de Associação foi o maior problema enfrentado. Neste trabalho, três abordagens foram propostas para o arquivo de saída do módulo de pré-processamento, sendo duas considerando lógica proposicional e uma considerando lógica relacional. A seguir são descritas essas abordagens. Para melhor compreendê-las, considere os planos de informação, o mapa de objetos mostrados na Figura 4.3, e a relação de adjacência entre os objetos mapeados.

4.3.1 Regras de associação espacial

Para extrair regras de associação espacial considera-se que a transação é constituída de todos os atributos de um plano de informação e por relações topológicas entre objetos desse plano de informação e objetos espaciais de outro plano de informação. Essa abordagem é muito interessante pois permite encontrar regras com mais de uma relação topológica e com características do plano de informação em função de determinadas relações topológicas. São

¹ Ambiente em desenvolvimento no Laboratório de Inteligência Computacional (LABIC) do Instituto de Ciências Matemáticas e de Computação (ICMC) da USP São Carlos. Tem o objetivo de fornecer um ambiente integrado para apoiar as etapas do processo de descoberta do conhecimento, oferecendo funcionalidades voltadas para aprendizado de máquina e mineração de dados e textos.

$Pi_1:X$				$Pi_2:Y$			
ID	Atributo A	Atributo B	GeoID	ID	Atributo C	Atributo D	GeoID
1	A_1	B	22	1	C_1	D_1	21
2	A_2	B	33	2	C_2	D_2	31
3	A_3	B	44	3	C_3	D_3	41

Figura 4.3: Planos de informação e mapa de objetos

geradas também regras que descrevem os dados sem a informação espacial. Assim, para se obter somente regras de associação espacial, deve-se filtrar as regras eliminando aquelas que não possuem relações topológicas. Essa abordagem só é possível na aplicação do *Apriori*, dentre os formatos de algoritmos estudados, pois esse algoritmo não exige a definição de todos os atributos antes de iniciar o processamento. Em outros formatos é custoso prever o número total dessas relações para colocar em um corpo inicial do arquivo. Além disso, cria-se um elevado número de atributos sendo necessário preencher com algum indicador de atributo não existente para muitos casos.

O formato do algoritmo *Apriori* foi implementado para essa abordagem. Cada linha do arquivo de saída representa uma relação com os atributos na forma *nome_atributo=valor* separados por um espaço em branco, sendo informados os atributos do primeiro plano de informação. Após são colocadas as relações com o identificador do objeto espacial do segundo plano de informação. Todas as relações entre todos os objetos são informadas em sequência na mesma linha. O arquivo gerado tem a extensão **.tab**. Um exemplo do arquivo de saída pode ver visualizado na Figura 4.4

```
X_AtributoA= $A_1$  X_AtributoB=B
X_AtributoA= $A_2$  X_AtributoB=B Adjacente( $Y_1$ ) Adjacente( $Y_3$ )
X_AtributoA= $A_3$  X_AtributoB=B Adjacente( $Y_3$ )
```

Figura 4.4: Exemplo do formato de saída do módulo de extração de relações topológicas para gerar regras de associação espacial

4.3.2 Regras de associação em espaço limitado

A abordagem anterior é interessante porém não relaciona os atributos descritivos dos objetos espaciais pertencentes a relação topológica. Uma maneira de encontrar essa relação é limitar os objetos espaciais de forma que são processados dados de objetos que respeitem a uma determinada condição espacial. Essa condição é representada por uma relação topológica. Cada linha do arquivo de saída representa os atributos descritivos dos dois objetos que participam da relação topológica. Objetos que não atendem a condição não são informados. Objetos que possuem relação com mais de um objeto tem seus atributos repetidos para cada relação. Assim, obtém-se regras em função da relação topológica utilizando os dados descritivos dos dois planos de informação informados. Essas regras representam um conjunto de elementos do banco de dados espaciais restritos a relação topológica.

Para isso, os dados dos planos de informação são gravados considerando todos os atributos do primeiro plano de informação. Em seguida são colocados todos os atributos do segundo plano de informação. Em cada atributo é acrescentada uma identificação no nome para informar a qual plano o atributo pertence. Para cada relacionamento existente são colocados todos os atributos, sendo que muitos dados são repetidos. A idéia dessa abordagem é considerar que cada relacionamento constitui uma transação. O problema é a utilização de apenas uma relação topológica. A interpretação da regra fica um pouco diferenciada das demais, devendo considerar que os dados originais não representam mais todos os objetos espaciais, só uma região limitada pela relação topológica. Essa abordagem, entretanto permite utilizar além do formato para o algoritmo *Apriori*, o formato para o ambiente DISCOVER.

Para o formato do algoritmo *Apriori* é considerado que cada linha do arquivo de saída representa uma relação entre dois objetos espaciais, com os atributos dos dois planos de informação na forma *nome_atributo=valor* separados por um espaço em branco. O arquivo gerado tem a extensão **.tab**. Um exemplo pode ver visualizado na Figura 4.5.

X_AtributoA=A₂ X_AtributoB=B Y_AtributoC=C₁ Y_AtributoD=D₁
X_AtributoA=A₂ X_AtributoB=B Y_AtributoC=C₃ Y_AtributoD=D₃
X_AtributoA=A₃ X_AtributoB=B Y_AtributoC=C₃ Y_AtributoD=D₃

Figura 4.5: Exemplo do formato de saída do módulo de extração de relações topológicas para gerar regra de associação em espaço limitado

No formato do ambiente DISCOVER são necessários dois arquivos, um com a definição dos atributos e outro com os dados. No arquivo de definições, identificado com a extensão **.name**, a primeira linha representa o atributo classe, não utilizado no caso de mineração de

regras de associação. As demais linhas representam os atributos e seus tipos de dados, no formato *nome_atributo: tipo_tributo*. O segundo arquivo tem a extensão **.data**, informando em cada linha os valores de atributos referentes a uma relação, separados por vírgulas. Os atributos não existentes recebem o valor "?". Esse formato é interessante porque o DISCOVER disponibiliza métodos de conversão para vários algoritmos de mineração de dados, podendo ser utilizado como um formato intermediário e não implementar todas as possíveis entradas dos algoritmos de mineração de regras de associação disponíveis.

4.3.3 Regras de associação espacial utilizando lógica relacional

Para encontrar regras de associação espacial considerando atributos descritivos dos dois planos de informação é necessário utilizar lógica relacional pois já existe uma relação espacial envolvida e definida. Assim é possível derivar regras em função dessas relações utilizando algoritmos de mineração de regras de associação relacional, como o WARM (Dehaspe and Toironen 2000). Para permitir a aplicação de lógica relacional, o arquivo de saída precisa apresentar os dados que expressem cláusulas lógicas. Isso pode ser feito considerando cláusulas formadas pelo nome do atributo seguida do identificador do objeto e do valor do atributo para aquele objeto. Para relações topológicas é considerado o nome da relação e os identificadores de cada objeto espacial que faz parte da relação. O problema de utilizar essa abordagem é que os algoritmos, no geral, são mais lentos que os algoritmos proposicionais e requerem um conhecimento de fundo para melhorar o resultado encontrado. O arquivo gerado possui a extensão **.pl** no formato utilizado pelo Prolog, com as relações, entre os objetos espaciais, expressas na forma *nome_relacao(objeto_espacial_pi_1, objeto_espacial_pi_2)* e os atributos *nome_atributo(objeto_espacial_pi, valor_atributo)*. Um exemplo pode ver visualizado na Figura 4.6.

atributoA(x_1, a_1).	atributoA(x_2, a_2).	atributoA(x_3, a_3).
atributoB(x_1, b).	atributoB(x_2, b).	atributoB(x_3, b).
atributoC(y_1, c_1).	atributoC(y_2, c_2).	atributoC(y_3, c_3).
atributoD(y_1, d_1).	atributoD(y_3, d_1).	atributoD(y_3, d_1).
adjacente(x_2, y_1).	adjacente(x_2, y_3).	adjacente(x_3, y_3).

Figura 4.6: Exemplo do formato de saída do módulo de extração de relações topológicas para gerar regras de associação espacial utilizando lógica relacional

O tratamento utilizando lógica relacional não foi abordado neste trabalho, sendo somente implementado um formato de saída para viabilizar a utilização de algoritmos de extração de regras de associação que utilizam lógica relacional. Os outros formatos foram entrada para o algoritmo *Apriori* na etapa de extração de regras de associação com informação espacial.

4.4 Extração de regras de associação com informação espacial

Após extrair as relações topológicas é executado um algoritmo de mineração de regras de associação. O algoritmo é escolhido entre os disponibilizados pela comunidade de mineração de dados e utilizado sem qualquer alteração para dados espaciais. Pode-se encontrar regras de associação espacial e regras de associação com informação espacial mas é necessário adaptar o conceito proposto por Koperski and Han (1995). Isso porque as regras de associação espacial geradas não estão totalmente coerente com a definição, a qual utilizava o conceito de predicados. Assim, é necessário redefinir regras de associação espacial como uma extensão de regra de associação. Também as regras de associação com informação espacial não possuem predicados espaciais em seu corpo e não são informadas as relações topológicas que os dados representam. Contudo, pode se considerar que essa relação está implícita na regra pois essas regras foram originadas por dados selecionados em função de uma relação topológica.

Esse trabalho propõe uma redefinição das regras de associação espacial e a definição de regras de associação com informação espacial. Para redefinir essas regras, considere o conjunto $A = \{a_1, a_2, \dots, a_m\}$ um conjunto de literais chamados itens, D um banco de dados espaciais composto por conjuntos de objetos espaciais $O = \{o_1, \dots, o_m\}$ no qual cada objeto $o_i \in O$ é formado por um conjunto de atributos de A , isto é, $o_i \subseteq A$. Considere também o conjunto de relações topológicas $R = \{r_1, \dots, r_h\}$, tal que para cada relação r , $r \in R$, exista uma função f_r que relacione os objetos espaciais o_i e o_j , isto é, $f_r(o_i, o_j) = \text{verdadeiro}$ se existe a relação topológica r entre o_i e o_j ou $f_r(o_i, o_j) = \text{falso}$ caso contrário. O conjunto F é formado por todas as funções $f_r(o_i, o_j) = \text{verdadeiro}$.

Uma regra de associação espacial pode ser adaptada da definição de regra de associação como:

Seja T o conjunto de transações tal que $T = \{t_1, t_2, \dots, t_n\}$, onde cada transação t_j é composta por um objeto espacial o_i e as funções $f_r(o_i, o_j)$ se $f_r \in F$. Isto é, $T \subset (O \cup F)$. Um conjunto de itens $X \subset (A \cup F)$ é chamado de um *itemset*, ou seja, um *itemset* é um subconjunto formado por atributos dos objetos espaciais e funções que definem relações topológicas. Uma transação t_n contém um *itemset* X se $X \subseteq t_n$. Considerando que LHS e RHS são *itemsets*, uma regra de associação espacial é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$, $LHS \cap RHS = \emptyset$ e $LHS \subset F$ ou $RHS \subset F$. A regra $LHS \Rightarrow RHS$, ocorre no conjunto de transações em D com confiança=*conf* se em *conf*% das transações em D que ocorre LHS , ocorre também RHS . A regra $LHS \Rightarrow RHS$ tem suporte=*sup* se em *sup*% das transações em D ocorre $LHS \cup RHS$.

A regra gerada pode ser interpretada como a relação entre um conjunto de atributos e um conjunto de relações topológicas. Isto é, quando ocorre o conjunto de atributos

tos em um objeto espacial, também existem as relações topológicas dos objetos com essas características descritas pelos atributos e outro objeto espacial. Por exemplo, a regra $\text{TeINTERSECTS}=(\text{APARECIDA}) \text{P2_IPTU_SETOR}=3 \Rightarrow \text{P2_IPTU_SITUACAO}=1$ (11.6, 100.0) pode ser compreendida como todos os imóveis no setor de IPTU 3 e possuem intersecção com *aparecida* apresentam a situação do IPTU referente ao código 1 em 11% dos objetos espaciais.

Já uma regra de associação com informação espacial pode ser definida como:

Seja o conjunto de transações $T=\{t_1, t_2, \dots, t_n\}$, onde cada transação t_j é composta pelo conjunto de atributos de um objeto espacial o_i unido ao conjunto de atributos de outro objeto espacial o_j , $T \subset (O \times O)$, de forma que esses objetos se relacionam em função de uma relação topológica $r, r \in R$ e $\exists f_r(o_i, o_j)$ tal que $f_r(o_i, o_j) \in F$. Um conjunto de itens $X \subset A$ é chamado de um *itemset*, ou seja, um *itemset* é um subconjunto formado por atributos dos objetos espaciais. Uma transação t_n contém um *itemset* X se $X \subseteq t_n$. Considerando que LHS e RHS são *itemsets*, uma regra de associação com informação espacial é uma implicação na forma $LHS \Rightarrow RHS$, em que $LHS \subset A$, $RHS \subset A$, $LHS \cap RHS = \emptyset$. A regra $LHS \Rightarrow RHS$, ocorre no conjunto de transações em D com confiança=*conf* se em *conf*% das transações em D que ocorre LHS , ocorre também RHS . A regra $LHS \Rightarrow RHS$ tem suporte=*sup* se em *sup*% das transações em D ocorre $LHS \cup RHS$.

A regra gerada pode ser entendida como a relação entre os atributos dos planos de informação dado que existe a relação topológica. Assim, em *sup*% dos objetos espaciais relacionados em função da relação topológica, ocorreram determinadas características expressas pelos atributos que aparecem nas regras. Por exemplo, dado somente os objetos lotes que sobrepõe quadras, foi encontrada a regra $\text{P1_Cid_CADIMOB_Parcela_CODLOGRADO}=38 \Rightarrow \text{P2_Quadra_SETOR}=3$ (12.1, 90.7). Essa regra pode ser interpretada como em 12.1% dos lotes que sobrepõe quadras, o código do logradouro do lote é 38 e o setor da quadra é 3. A confiança é de 90.7%, significando que dos lotes com código logradouro igual a 38 e que sobrepõe quadras, 90.7% das quadras sobrepostas estão no setor 3.

Em ambas abordagens das regras de associação encontradas, o algoritmo não necessita alteração. Contudo, as regras geradas nem sempre apresentam um conhecimento novo, útil e interessante. Para avaliar os resultados são necessários métodos de pós-processamento.

4.5 Pós-processamento das regras

Na etapa de pós-processamento, antes de se avaliar o resultado, pode ser necessária a aplicação de filtros para eliminar algumas regras, devido ao grande número de regras

geradas ou para eliminar regras que não estão de acordo com o objetivo da mineração. Esse é o caso das regras de associação espacial, em que são geradas regras com ou sem relações topológicas. Como é necessário ter pelo menos uma relação topológica presente na regra para ser considerada uma regra de associação espacial, foi desenvolvido, neste trabalho, um filtro que elimina as regras sem tal característica. Dessa maneira, o filtro pode ser utilizado para buscar as relações topológicas.

Para tratar o problema do elevado número de regras geradas, também foi desenvolvido um filtro que elimina ou mantém regras com determinado atributo. Outro filtro implementado mantém somente regras que possuem características dos dois planos de informação, eliminando regras com apenas atributos de um plano de informação. O usuário que decide qual filtro utilizar.

Após reunir e filtrar as regras representativas do banco de dados espaciais e do interesse do usuário, é necessário avaliar esses resultados. A avaliação das regras encontradas pode ser realizada por meio de medidas de avaliação. Também pode ser utilizada uma avaliação subjetiva, em que um especialista analisa as regras, auxiliado por medidas de avaliação, para definir se a regra é interessante ou não.

4.6 *Considerações finais*

Nesse capítulo foi apresentado o objetivo do projeto, o módulo de extração de relações topológicas, a metodologia utilizada e as regras possíveis de se encontrar trabalhando com dados espaciais. É importante observar que a opção por utilizar uma biblioteca disponível TerraLib impactou muito no projeto. A própria escolha da linguagem para implementação foi baseada na linguagem da biblioteca pois implementar uma camada de comunicação entre a biblioteca e o módulo iria demandar mais tempo de processamento. Seria possível também não utilizar a biblioteca e reimplementar os métodos, mas inicialmente a idéia de utilizar uma biblioteca *open source* facilitaria a integração com os sistemas de informações geográficas disponíveis no mercado e diminuiria o tempo de implementação do módulo. O emprego da biblioteca facilitou muito o desenvolvimento, pois a implementação de métodos eficientes de extração de relacionamentos iria despendar meses entre pesquisas e implementação, requerendo um tempo maior que o disponível para a conclusão deste trabalho. Além disso, essa reimplementação poderia gerar funcionalidades não eficientes ou dependentes do banco de dados. Por outro lado, algumas escolhas do projeto ficaram limitadas às possibilidades oferecidas pela biblioteca, como o formato de intercâmbio dos dados, o banco de dados, a linguagem de implementação. Também foi necessário sete meses para estudar a biblioteca e a linguagem C++, adequar o ambiente de desenvolvimento, realizar a correta compilação da biblioteca e integrar com o banco de dados. Isso demandou um tempo não previsto no início do projeto.

Para verificar se a metodologia utilizada e o módulo de extração de relações topológicas implementado atendem aos objetivos propostos, no próximo capítulo está descrito o estudo de caso utilizado para a validação da metodologia e do módulo implementado neste trabalho.

Estudo de caso

A viabilidade da proposta apresentada é demonstrada por meio de um estudo de caso envolvendo o sistema de informação geográfica da cidade de Jaboticabal. Nesse Capítulo esse estudo é descrito, mostrando a organização dos dados, os experimentos realizados e os resultados encontrados.

A cidade de Jaboticabal está situada no interior do estado de São Paulo, localizada na latitude $-21,2547^\circ$ e longitude $-48,3222^\circ$ possuindo cerca de 72 mil habitantes, área da unidade territorial de aproximadamente 707 Km^2 . A população residente total é de 67408 habitantes, sendo 63840 na área urbana (IBGE 2005). Essa cidade foi escolhida por já existir um projeto sobre sistema de informação geográfica no departamento de engenharia civil da Universidade Federal de São Carlos e a prefeitura consentiu com a divulgação dos resultados encontrados.

5.1 Sistema de informação geográfica da cidade Jaboticabal

O sistema de informação geográfica da cidade Jaboticabal foi construído a partir de dados pré-existentes na prefeitura e da Superintendência de Água e Esgoto (SAE). Foram reunidos dados de lotes, quadras, setores, área urbana, rural, rodovias, logradouros, corredores urbanos, estabelecimentos comerciais, valores de IPTU e rede de água e esgoto. Além desses dados, para a mineração de regras de associação foram incluídos dados do Instituto Brasileiro de Geografia e Estatística (IBGE) sobre a população de Jaboticabal. No início do projeto estava disponível somente parte desse sistema, uma região escolhida pela prefeitura para ser

a região piloto no projeto desenvolvido pelo departamento de engenharia civil. Contudo, os experimentos realizados com essa região foram invalidados devido a pouca confiabilidade, existindo muitos dados inconsistentes ou não informados. Essa situação foi observada somente depois da realização dos experimentos, uma vez que não foi possível encontrar padrões. Assim, necessitou-se de todos os dados do sistema de informação geográfica de Jaboticabal e todas as etapas de mineração foram novamente realizadas. A vantagem de utilizar todos os dados do sistema é a eliminação da elaboração de amostras que refletem a realidade mapeada. Neste trabalho somente é descrito o estudo de caso com os dados do sistema de informação geográfica completo. Esses dados também apresentavam alguns erros oriundos da integração dos diferentes sistemas, sendo necessário, para a mineração de dados, realizar diversas atividades na etapa de pré-processamento, descritas na seção 5.4.

Para a realização desse estudo de caso, foi necessário conhecer a plataforma em que o sistema de informação geográfica da cidade de Jaboticabal foi desenvolvido. Esse sistema utiliza o software *ARCGIS* para apresentar os mapas e as informações cadastrais. Além desse software, o banco de dados *Microsoft Access* armazena as informações cadastrais e as referências para as informações geográficas. Os dados georreferenciados são armazenados em tabelas específicas reconhecidas somente pelo software *ARCGIS*. Assim, foi necessário a utilização do aplicativo *ARCGIS* para recuperar os dados espaciais e disponibilizá-los para a mineração de dados.

Um problema enfrentado nesse estudo foi a falta de documentação do sistema de informação geográfica da cidade de Jaboticabal. Isso dificultou a compreensão e identificação do foco da mineração de dados, sendo necessário a construção do modelo de banco de dados para identificar possíveis objetos a serem utilizados na mineração. Esse modelo é descrito na próxima seção.

5.2 Modelo do banco de dados

As tabelas do banco de dados de Jaboticabal apresentam na maioria dos casos informações sobre características cadastrais de determinado aspecto da cidade. Além dessas tabelas, existem tabelas de controle do sistema e tabelas para identificação de características espaciais dos objetos. Ao todo são 142 tabelas das quais 30 tabelas são controle do sistema *ARCGIS*, 86 tabelas contêm dados espaciais cadastrais e 26 são tabelas descritivas, podendo existir índices para alguma tabela espacial. Além das tabelas do sistema de informação geográfica foram adicionadas 8 tabelas espaciais fornecidas pelo IBGE, com dados de censo da população de Jaboticabal. As principais tabelas utilizadas são: *alfabetização*, *cadmobiliario*, *chefe_familia*, *cidadao*, *contorno_lote*, *contorno_quadra*, *corredor_comercial*, *dbo_cadimob*, *dbo_iptu2004*, *domicilio*, *ensino_assitencial*, *ensino_especial*, *ensino_fundamental*, *ensino_infantil*, *escolas_estaduais*, *hidrografia*, *rodovia* e *setor*, apresentadas na Tabela 5.1. Nessa tabela pode-

se observar o número de objetos espaciais (tamanho ou número de exemplos), o número de atributos de cada objeto, se é espacial ou descritivo e o tipo de dado geográfico desse objeto.

Tabela 5.1: Principais tabelas do banco de dados de Jaboticabal

Tabela	Número de registros	Atributos	Espacial	Tipo
Alfabetizacao	90	270	sim	Cadastral
Cadmobiliario	4786	11	não	-
Chefe_familia	90	187	sim	Cadastral
Cidadao	90	177	sim	Cadastral
Contorno_lote	25188	9	sim	Cadastral
Contorno_quadra	1411	7	sim	Cadastral
Corredor_comercial	17	3	sim	Cadastral
dbo_Cadimob	26518	23	não	-
dbo_IPTU2004	21725	3	não	-
Domicilio	90	91	sim	Cadastral
Ensino_assistencial	1	4	sim	Cadastral
Ensino_especial	1	4	sim	Cadastral
Ensino_fundamental	5	4	sim	Cadastral
Ensino_infantil	16	4	sim	Cadastral
Escolas_estaduais	11	4	sim	Cadastral
Hidrografia	434	3	sim	Cadastral
Rodovias	26	3	sim	Cadastral
Setor	4	5	sim	Cadastral

A relação entre as tabelas do sistema de informação geográfica pode ser observada no modelo simplificado do banco de dados espaciais da cidade de Jaboticabal, na Figura 5.1. As tabelas espaciais estão separadas em tabelas que foram originadas de dados do IBGE, tabelas referente a lotes em que se encontram escolas e tabelas representando os objetos lotes, quadras, setores, hidrografia, corredor comercial e rodovia. Existem, também, tabelas descritivas com informações a respeito dos lotes, que são associadas a objetos espaciais pela chave **codreduzido**, código formado pelo setor, quadra e lote. Esses lotes podem representar um terreno sem construção, um terreno edificado sendo domicílio ou estabelecimento comercial. Os estabelecimentos comerciais estão representados em **cadmobiliario**. As informações sobre o IPTU em **dbo_ipu2004** e informações como a forma de construção, se o lote tem alguma melhoria como calçadas e muro, se está edificado ou não, detalhes descritivos do lote como pedologia, topografia, categoria do proprietário, se particular, estadual ou federal estão em **dbo_cadimob**. As escolas são representadas pelos lotes que ocupam, em tabelas específicas para cada tipo de escola.

As informações do IBGE estão em função dos setores censitários da cidade. São noventa setores não coincidindo com os setores da tabela **setor** de Jaboticabal. Os atributos estão

codificados em variáveis denominadas Vx e apresentam informações como o número de pessoas alfabetizadas em relação a idade, qual a escolaridade do chefe de família, a renda, os domicílios com tratamento de água, esgoto, quantidade de cômodos, o número de moradores, de cidadãos do sexo feminino, masculino e o número em relação a faixas etárias.

Figura 5.1: Modelo simplificado do banco de dados espaciais da cidade de Jaboticabal

Os principais objetos espaciais utilizados foram rodovia, corredor_comercial, contorno_lote,

hidrografia, setor e contorno_quadras. Na Figura 5.1 pode-se observar a geometria dos objetos pelos atributos *shape*. O atributo *shape* é um índice para a geometria do objeto. Se o objeto tiver a geometria linha, terá o atributo *shape_length*. Se for um polígono, terá o atributo *shape_area*.

Além dessas tabelas, o banco de dados possui outras tabelas referentes a descrição de códigos utilizados e informações não utilizadas neste trabalho. As tabelas selecionadas foram escolhidas no processo de identificação do problema descrito na seção 5.3.

5.3 Identificação do problema

No processo de mineração de dados é necessário definir um foco para a extração de dados, com a finalidade de auxiliar na busca e compreensão dos padrões encontrados. Para isso é necessário estudar o domínio da aplicação e a maioria das vezes, em conjunto com um especialista da área para definir esse foco. Nesse estudo de caso essas estratégias foram utilizadas para selecionar quais tabelas são interessantes para extração de padrões.

O especialista intuiu que as características físicas de lotes e valores de IPTU poderiam apresentar algum comportamento padrão quando próximos de serviços públicos. Em uma consulta ao sistema de informação geográfica, foi verificado que somente dados sobre escolas estavam disponíveis. Não foi possível identificar outros serviços públicos como hospitais, prefeitura e pronto-socorro. Assim, o primeiro aspecto abordado visa definir características semelhantes de lotes e IPTU próximos a escolas. Uma segunda possibilidade proposta pelo especialista é a relação entre os dados da população obtidos do IBGE e as características dos lotes e do IPTU.

Além das relações expressas pelo especialista, também foram pesquisadas relações entre outros objetos espaciais, descritas a seguir.

Lote, IPTU e rodovia: foram selecionados esses aspectos com a finalidade de verificar se existe características físicas e de valores de IPTU de lotes com facilidade de acesso a rodovias.

Lote, IPTU e hidrografia: neste caso o foco é verificar se os lotes apresentam características específicas por estarem próximos de rios ou córregos.

Lote, IPTU e corredor comercial: esse aspecto trata da influência dos valores de IPTU e das características dos lotes por estarem próximos a vias de acesso de regiões comerciais.

Lote, IPTU, setor: o foco é verificar se a divisão física de setores da cidade engloba características específicas dos lotes.

Lote, IPTU e serviços: semelhante a pesquisa sobre escolas, porém buscando padrões de lotes em função da disponibilidade de serviços próximos, como médicos, padarias, supermercados.

Serviços e escolas: esse aspecto trata de encontrar padrões de serviços oferecidos próximos a escolas.

Serviços e Setor IBGE: neste caso é verificado padrões de serviços em função de características da população residente em determinado setor.

Setor IBGE e serviços: o foco desse aspecto é verificar padrões da população na proximidade de determinados serviços.

Setor IBGE e hidrografia: o foco é procurar padrões na população de regiões ribeirinhas.

Setor IBGE e corredor comercial: o foco é procurar padrões na população próximas a centros urbanos.

A Tabela 5.2 apresenta os aspectos procurados, mostrando quais planos de informação oferecem as informações procuradas. Também apresenta a relação que provavelmente possa existir entre os objetos dos planos de informação.

Tabela 5.2: Aspectos procurados na extração de padrões

Plano de Informação 1	Plano de Informação 2	Relação
Lote e IPTU	Escolas	Proximidade
Lote e IPTU	Rodovias	Proximidade
Lote e IPTU	Hidrografia	Proximidade
Lote e IPTU	Corredor comercial	Proximidade
Lote e IPTU	Setor IBGE	Inclusão
Lote e IPTU	Setor	Inclusão
Lote e IPTU	Serviços	Proximidade
Serviços	Escolas	Proximidade
Serviços	Setor IBGE	Proximidade
Setor IBGE	Serviços	Inclusão
Setor IBGE	Hidrografia	Proximidade
Setor IBGE	Corredor comercial	Proximidade

Contudo, no sistema de informação geográfica não existem alguns planos de informação desejados. Esses podem ser criados a partir dos dados do banco de dados espaciais. As tabelas que contém os dados necessários para a criação dos planos de informação são apresentadas na Tabela 5.3.

Tabela 5.3: Relação entre os aspectos procurados e as tabelas do banco de dados de Jaboticabal

Aspecto	Tabela(s)
Setor IBGE	Alfabetizacao, Chefe_familia, Cidadao, Domicilio
Serviços	Cadmobilario
Lote e IPTU	Contorno_lote, dbo_Cadimob, dbo_IPTU2004
Corredor comercial	Corredor_comercial
Escolas	Ensino_assistencial, Ensino_especial, Ensino_fundamental, Ensino_infantil, Escolas_estaduais
Hidrografia	Hidrografia
Rodovias	Rodovias
Setor	Setor

Analisando esses aspectos e as tabelas com os dados necessários, pode-se observar que o banco de dados da cidade de Jaboticabal não contém as tabelas em um formato adequado para o processamento. Assim, atividades de pré-processamento foram aplicadas para adequar as tabelas aos objetivos propostos.

5.4 Pré-processamento

As atividades de pré-processamento da metodologia proposta neste trabalho foram aplicadas aos dados de Jaboticabal. É importante observar que essas atividades foram executadas mais de uma vez, considerando o processo iterativo. Ao se executar e avaliar o conhecimento obtido, em alguns casos foi necessário realizar novamente algumas atividades da etapa de pré-processamento ou executar o algoritmo de mineração de regras de associação com uma nova configuração. O estudo de caso apresentado mostra apenas os resultados da última interação do processo de mineração. Na seção 5.4.1 são apresentadas as atividades de pré-processamento com exceção da atividade de extração de relações topológicas, apresentada em separado na seção 5.4.2.

5.4.1 Preparação dos dados para mineração de regras de associação

Diversas atividades de pré-processamento foram realizadas em diferentes tabelas para viabilizar a mineração dos dados. A primeira tabela alterada foi `dbo_Cadimob`. Essa tabela estava normalizada para diminuir o volume dos dados. Assim, foram realizadas consultas SQL para alterar campos que possuíam índices para valores descritivos desses campos. Essa desnormalização foi aplicada nos campos uso do terreno, benfeitoria, topografia, categoria proprietário e pedologia. Além disso, também foram eliminados campos que apresentavam valores

iguais, encontrados por meio de consultas ao banco de dados. Também foram removidos os campos que não estavam informados ou haviam poucos registros como o **código do condomínio**, o **complemento**, a **unidade** e **sub-unidade**. O campo **área do terreno** foi transformado em intervalos, pois cada terreno possui um valor diferente sendo difícil encontrar um padrão. Para definir os intervalos, foi realizado um estudo gráfico do número de ocorrências nos intervalos, buscando um gráfico com uma curva normal. Isso porque, segundo o especialista, não é esperado que existam muitos terrenos com um tamanho pequeno ou grande demais, estando a maioria próximos dos valores utilizados na divisão de um loteamento, entre 150 e 430 metros quadrados. Os intervalos considerados foram menor que 150, 150-220, 220-300, 300-430, 430-760, 760-1000 e maior de 1000. A transformação em intervalos foi feita por consultas SQL.

Após essas tarefas, a tabela **dbo_IPTU** foi alterada. As informações de IPTU são importantes e podem trazer conhecimento novo, mas a estrutura de tabelas com esses dados dificulta a obtenção dessas informações. Os dados estão armazenados nas tabelas **IPTU2004isentos**, **dbo_IPTU2004**, **IPTU2004devedor** e **dbo_tipoisenção**. As atividades necessárias para melhorar o acesso às informações foram a criação de um novo campo informando quem é isento e a discretização de valores numéricos. Os dados da tabela **IPTU2004isentos** foram filtrados para representar somente o ano de 2004, uma vez que existiam informações de 2005 mas os valores da tabela de IPTU eram de 2004. A tabela **dbo_IPTU2004** foi duplicada em uma tabela temporária para adicionar um campo denominado **isento**. Esse campo foi preenchido de acordo com a existência ou não do **codreduzido** na tabela de isento. Se existisse o **codreduzido** nessa tabela, o valor colocado é **não**. Se não existisse, o valor é **sim**. Após a criação desse novo campo, somente a tabela temporária foi considerada, as outras referentes ao IPTU foram descartadas. Para criar a tabela final de IPTU, foi preciso discretizar os valores dos campos **soma tributo**, que representam a soma total do valor de IPTU pago e **soma área**, sobre a área construída do imóvel. Esses campos apresentam valores numéricos muito específicos para cada lote. Com a finalidade de tornar o resultado da mineração mais interessante, esses valores foram transformados em intervalos. A criação dos intervalos considerou, para a soma do tributo, as faixas de valores 0, 0 a 250, de 250 a 500, de 500-1000, de 1000 a 2500, 2500-5000, 5000-7500, 7500-10000 e maiores que 10000. Para área construída foram considerados os intervalos 0, 0-70, 70-140, 140-200, 200-350 e maior que 350. Para encontrar esses valores, também foi realizado um estudo de diferentes intervalos possíveis e selecionado os intervalos que apresentaram um gráfico próximo a um gráfico normal. Além disso, a área construída apresentava valores não preenchidos, considerado que não possuem área construída. Para confirmar essa informação, foi verificado que os lotes que não estavam preenchidos apresentavam uso do terreno sem edificação, na tabela **dbo_Cadimob**. Pode-se observar essas alterações na Figura 5.2.

Por fim as tabelas temporárias de **dbo_Cadimob** e **dbo_IPTU** foram unificadas em uma

tabela por meio do campo `codreduzido`, pois apresentavam um relacionamento um para um. A tabela resultante `Cadimob-IPTU2004_Dis` foi exportada no formato *Shapefile*. O software *ARCGIS* foi utilizado para realizar uma consulta espacial unindo a tabela com dados georreferenciados `Contorno_lote` e a tabela `Cadimob-IPTU2004_Dis`, por meio do campo `codreduzido`. O formato *Shapefile* é aceito pelo software TerraView para importar os dados. Assim foi possível realizar a importação dos dados de Lote e IPTU, gerando o plano de informação `LoteIPTU` para o formato que o módulo de extração de relações topológicas pode manipular.

Figura 5.2: Modelo das tabelas de IPTU e lote

Além dessa alteração, as tabelas sobre escolas também foram unificadas utilizando uma consulta SQL de união. Essa união foi possível uma vez que as tabelas possuíam o mesmo número e tipo de campos. A tabela resultante da união pode ser observada na Figura 5.3.

Figura 5.3: Modelo das tabelas de lotes de escolas

Foi criado um novo campo, o tipo de escola para identificar a categoria da escola representada. Também foi verificado que as tabelas apresentavam somente escolas estaduais ou

municipais. Essa verificação foi possível porque as escolas particulares estavam cadastradas na tabela **Cadmobilario** e as escolas estaduais e municipais não. A tabela **escola** também foi exportada no formato *Shapefile* e importada pelo software TerraView criando o plano de informação Escolas.

Após a criação do plano de informação Escolas, foi tratado o aspecto referente aos serviços oferecidos, ou seja, locais com estabelecimentos comerciais. Contudo essa informação não estava georreferenciada, somente possuía um campo de ligação com a tabela **dbo_Cadimob** e foi necessário georreferenciar utilizando **Contorno_lote**. A primeira atividade para criar a tabela serviços foi a eliminação dos campos **endereço** e **identificação de proprietário**, pertencentes a tabela **Cadmobilario**. A seguir, realizar a consulta SQL para selecionar os lotes de **dbo_Cadimob** que eram estabelecimentos comerciais, ou seja, existiam na tabela **Cadmobilario**. Nesses lotes foi inserido o campo informando a descrição do serviço oferecido. Esse conjunto de informações associado a tabela espacial **Contorno_lote** constitui a tabela **Serviços**. Como as tabelas geradas anteriormente, também foi exportada pelo software *ARCGIS* e importada pelo TerraView, gerando o plano de informação Serviços. As tabelas envolvidas nesse processo podem ser visualizadas em na Figura 5.4.

Figura 5.4: Modelo das tabelas de lotes que disponibilizam serviços

A próxima tarefa executada para pré-processar os dados foi tratar as tabelas do IBGE para originar o plano de informação SetorIBGE. Esse tratamento consistiu em eliminar muitos atributos devido ao grande número de atributos existentes. A seleção de atributos foi realizada junto ao especialista de planejamento urbano. Esse especialista filtrou as informações mantendo somente os dados que apresentavam totais genéricos e retirando informações muito específicas. Cada atributo estava representado por um nome de variável, V_x . Esses

nomes foram alterados para nomes com o significado real, a fim de facilitar a compreensão na análise das regras. Além disso, cada tabela do IBGE representa 90 setores, sendo possível unificar as tabelas utilizando o campo **object_id** para tal atividade. Esse campo referenciava um setor do IBGE. Com a unificação, as tabelas foram transformadas em uma só tabela com todos os atributos de alfabetização, cidadão, domicílio e chefe de família, referente a cada setor do IBGE. Assim, as tabelas de Alfabetização, Chefe_família, Cidadão e Domicílio foram alteradas conforme mostrado na Figura 5.5.

Figura 5.5: Modelo das tabelas do IBGE

Por fim, as tabelas **Corredor_comercial**, **Hidrografia**, **Setor** e **Rodovias** foram exportadas pelo software *ARCGIS* para constituir os planos de informação **CorredorComercial**, **Hidrografia**, **Setor** e **Rodovias**, respectivamente. Essas tabelas não sofreram alterações por serem dados

georreferenciados e sem informações descritivas. O modelo do banco de dados final utilizado como base para buscar relações topológicas pode ser observado na Figura 5.6.

Figura 5.6: Modelo com tabelas pré-processadas finais

Com todos os dados no formato disponível pelo software TerraView, mesmo formato acessado pela biblioteca TerraLib, foi possível executar o módulo de extração de relações topológicas para encontrar as relações topológicas utilizadas na mineração dos dados.

5.4.2 *Extração de relações topológicas*

O módulo de extração de relações topológicas foi executado para cada aspecto proposto na etapa de identificação do problema. A execução foi realizada considerando as abordagens de regras de associação espacial e regra de associação com informação espacial. A fim de encontrar os relacionamentos necessários para se extrair regras de associação espacial foram utilizadas todas as relações topológicas e a relação de proximidade considerando uma distância de 300 metros. Todos os atributos dos plano de informação também foram utilizados. Isso porque foram eliminados os atributos não relevantes na atividade anterior da etapa de pré-processamento.

Em regra de associação com informação espacial, somente uma relação topológica é considerada. A escolha de qual relação utilizar foi baseada nos resultados do experimento para extrair regras de associação espacial. Todas as relações encontradas para dois planos de informação aparecem no arquivo de saída do experimento sobre regras de associação espacial. Dessa forma, somente as relações topológicas que estão nesse arquivo existem entre os planos de informação utilizados. Assim, utiliza-se somente as relações que com certeza existem na busca por regra de associação com informação espacial. Para cada relação existente foi executado um experimento diferente na abordagem regra de associação com informação espacial.

Os resultados obtidos podem ser observados na Tabela 5.4. A primeira coluna da tabela é um identificador do experimento, utilizado nas Tabelas 5.4, 5.5 e 5.6 para representar o experimento. Os campos PI identificam quais planos de informação foram considerados no experimento. A coluna **Relação** mostra a relação topológica procurada. O valor **Todas** se refere as relações de igualdade, intersecção, toque, cruzamento, inclusão (dentro ou contido), sobreposição, cobertura (cobre e coberto por) e a relação de proximidade considerando a distância de 300 metros. A coluna **Formato** se refere a abordagem e por consequência ao formato de saída adotado no experimento. O formato **RE** é utilizado para representar o formato de geração de regras de associação espacial e **RR** para regra de associação com informação espacial. A coluna **Tempo** é o tempo total de processamento para encontrar as relações e gerar o arquivo de saída no formato adequado. A coluna **Arquivo** é o tamanho do arquivo gerado e **Memória** a memória máxima consumida pelo módulo durante a execução.

Nessa tabela pode-se observar que o tempo para encontrar as relações topológicas foi muito variado, mas está relacionado diretamente com o tamanho dos planos de informação. Além disso, pode-se observar que o uso da memória é alto, por volta de 32Mb mas não é um

Tabela 5.4: Experimentos de extração de relações topológicas

	PI	PI	Relação	Formato	Tempo	Arquivo	Memória
1	LotelPTU	Escolas	Todas	RE	01:03:47	14159	32347
2	LotelPTU	Escolas	Proximidade	RR	00:35:18	10895	30198
3	LotelPTU	Rodovias	Todas	RE	01:03:28	13906	31870
4	LotelPTU	Rodovias	Proximidade	RR	00:31:26	1066	31571
5	LotelPTU	Hidrografia	Todas	RE	01:22:10	14273	31832
6	LotelPTU	Hidrografia	Proximidade	RR	00:50:11	65835	31794
7	LotelPTU	CorredorComercial	Todas	RE	01:07:21	13226	31724
8	LotelPTU	CorredorComercial	Proximidade	RR	00:33:01	3905	31800
9	LotelPTU	SetorIBGE	Todas	RE	02:20:55	14285	32188
10	LotelPTU	SetorIBGE	Inclusão	RR	00:41:04	48820	31796
11	LotelPTU	SetorIBGE	Sobreposição	RR	00:49:05	59538	31796
12	LotelPTU	Setor	Todas	RE	00:49:46	13845	31636
13	LotelPTU	Setor	Inclusão	RR	00:25:59	14798	31476
14	LotelPTU	Setor	Sobreposição	RR	00:18:25	106	31648
15	LotelPTU	Serviços	Todas	RE	05:39:30	65446	32348
16	LotelPTU	Serviços	Proximidade	RR	04:50:13	763712	32520
17	Serviços	Escolas	Todas	RE	00:04:26	719	25472
18	Serviços	Escolas	Proximidade	RR	00:01:29	1056	24676
19	Serviços	SetorIBGE	Todas	RE	00:07:55	801	24994
20	SetorIBGE	Serviços	Todas	RE	00:02:49	594	25616
21	SetorIBGE	Serviços	Inclusão	RR	00:00:54	4483	25164
22	SetorIBGE	Serviços	Sobreposição	RR	00:00:26	1258	24636
23	SetorIBGE	Hidrografia	Todas	RE	00:00:27	157	15828
24	SetorIBGE	CorredorComercial	Todas	RE	00:00:12	142	16036

fator que torna inviável a busca por relações topológicas. Esses dados serão avaliados em detalhes na seção 5.7.

5.5 Extração de padrões

Os arquivos gerados no pré-processamento foram utilizados como entrada para a execução do algoritmo Apriori¹ para encontrar regras de associação. A configuração considerada para o algoritmo em todos os experimentos foi suporte mínimo de 10% e a confiança mínima de 80%. Esse algoritmo está implementado na linguagem C e para utilizá-lo foi desenvolvida uma interface gráfica com a finalidade de auxiliar na configuração dos parâmetros utilizados.

Na Tabela 5.5 estão sintetizados os experimentos executados em função dos campos **Arquivo** e **Regras**, respectivamente, o tamanho do arquivo de regras gerado e o número de regras. Pode-se observar que o número de regras geradas foi elevado na maioria dos casos, dificultando a interpretação das mesmas. No experimento 20 não foi possível obter o número

¹Implementação do algoritmo *Apriori* por Christian Borgelt, disponível em <http://fuzzy.cs.uni-magdeburg.de/borgelt/apriori.html>

de regras devido a grande quantidade de regras geradas. Para viabilizar a análise desse elevado número de regras, na etapa de pós-processamento foi executado um filtro de regras. Essa etapa está descrita a seguir.

Tabela 5.5: Experimentos de extração de regras de associação

	PI	PI	Formato	Arquivo	Regras
1	LotelPTU	Escolas	RE	5485	36794
2	LotelPTU	Escolas	RR	10200	67883
3	LotelPTU	Rodovias	RE	5269	35205
4	LotelPTU	Rodovias	RR	21083	141538
5	LotelPTU	Hidrografia	RE	5269	35205
6	LotelPTU	Hidrografia	RR	14212	94467
7	LotelPTU	CorredorComercial	RE	5269	35205
8	LotelPTU	CorredorComercial	RR	6595	43731
9	LotelPTU	SetorIBGE	RE	5269	35205
10	LotelPTU	SetorIBGE	RR	242397	1607239
11	LotelPTU	SetorIBGE	RR	234527	1557035
12	LotelPTU	Setor	RE	7853	54873
13	LotelPTU	Setor	RR	11928	81035
14	LotelPTU	Setor	RR	11467	78725
15	LotelPTU	Serviços	RE	447418	2347551
16	LotelPTU	Serviços	RR	5805	38627
17	Serviços	Escolas	RE	1	10
18	Serviços	Escolas	RR	5	52
19	Serviços	SetorIBGE	RE	1	4
20	SetorIBGE	Serviços	RE	10797210	-
21	SetorIBGE	Serviços	RR	28039	195891
22	SetorIBGE	Serviços	RR	26813	182884
23	SetorIBGE	Hidrografia	RE	8277	59486
24	SetorIBGE	CorredorComercial	RE	7491	53289

5.6 Pós-processamento

Após gerar as regras foi necessário avaliar se essas regras traziam algum conhecimento interessante a respeito dos dados, de acordo com técnicas de avaliação citadas na seção 3.1.3. As regras foram filtradas e o especialista em planejamento urbano auxiliou na avaliação e seleção de quais regras são interessantes. Na seção 5.6.1 está descrito como os filtros foram aplicados e como o especialista selecionou as regras. Na seção 5.6.2 são apresentadas algumas regras selecionadas.

5.6.1 Seleção de regras

No pós-processamento o filtro implementado foi executado para eliminar regras. Quando se tratava de arquivos com regras de associação espacial o filtro somente mantinha regras que possuíam alguma característica espacial. No caso de regra de associação com informação espacial o filtro mantinha as regras que possuíam atributos dos dois plano de informação. A utilização do filtro não é obrigatória, mas devido ao número elevado de regras facilitou a obtenção dos padrões. Uma síntese dos resultados dos experimentos utilizando filtros pode ser observada na Tabela 5.6. O filtro elimina uma grande quantidade de regras mas o número final ainda foi elevado. Foi interessante observar que em alguns casos todas as regras foram eliminadas, mostrando que não existe uma relação espacial entre as características desejadas. As regras que foram eliminadas mostravam relações entre as características descritivas de objetos de um mesmo plano de informação ou regras de associação que seriam encontradas aplicando o algoritmo nos dados descritivos apenas.

Tabela 5.6: Arquivos e regras de associação após aplicação do filtro

	PI	PI	Formato	Arquivo	Regras
1	LotelPTU	Escolas	RE	216	1589
2	LotelPTU	Escolas	RR	3346	22312
3	LotelPTU	Rodovias	RE	0	0
4	LotelPTU	Rodovias	RR	4162	28089
5	LotelPTU	Hidrografia	RE	0	0
6	LotelPTU	Hidrografia	RR	4096	27353
7	LotelPTU	CorredorComercial	RE	0	0
8	LotelPTU	CorredorComercial	RR	0	0
9	LotelPTU	SetorIBGE	RE	0	0
10	LotelPTU	SetorIBGE	RR	45717	299460
11	LotelPTU	SetorIBGE	RR	36703	241470
12	LotelPTU	Setor	RE	2585	19668
13	LotelPTU	Setor	RR	3184	21068
14	LotelPTU	Setor	RR	3154	21109
15	LotelPTU	Serviços	RE	442147	2312346
16	LotelPTU	Serviços	RR	0	0
17	Serviços	Escolas	RE	1	6
18	Serviços	Escolas	RR	4	38
19	Serviços	SetorIBGE	RE	0	0
20	SetorIBGE	Serviços	RE	5415797	-
21	SetorIBGE	Serviços	RR	7642	52733
22	SetorIBGE	Serviços	RR	7176	49000
23	SetorIBGE	Hidrografia	RE	787	6197
24	SetorIBGE	CorredorComercial	RE	0	0

As regras encontradas foram analisadas por um especialista que utilizou as medidas de suporte e confiança para seleção das regras. O especialista considerou válidas as regras com

confiança superior a 98% e com os maiores suportes, indicando a maior ocorrência do padrão nos dados. Porém não foi possível definir um valor geral devido a variação dessa medida entre os experimentos. Algumas regras encontradas são apresentadas na próxima seção.

5.6.2 Regras avaliadas pelo especialista

Diferentes regras foram encontradas, mas a princípio nenhuma mostrou ser inovadora. Esses resultados devem ser interpretados de acordo com o aspecto procurado e propostos na etapa de identificação do problema. A seguir são apresentados alguns resultados relevantes encontrados de acordo com o aspecto procurado.

Lote, IPTU e Escolas

Considerando os experimentos que utilizavam Lote, IPTU e Escolas foram realizados experimentos buscando encontrar regras de associação espacial. Como observado nas tabelas anteriores, foi possível gerar regras de associação espacial utilizando o pré-processamento de relações topológicas proposto e o algoritmo de mineração de dados padrão, ou seja, sem realizar alteração no algoritmo. Foram encontradas regras como apresentadas na Figura 5.7.

```
BUFFER=Estadual ⇒ P1_Lote_IPTU_PEDOLOGI=NORMAL (14.6, 81.1)
BUFFER=Estadual ⇒ P1_Lote_IPTU_IPTUISENTO=nao (14.7, 81.6)
BUFFER=Estadual P1_Lote_IPTU_USOTERRE=EDIFICADO ⇒
P1_Lote_IPTU_BENFEITO=CALÇADA+_MURO (11.6, 94.0)
P1_Lote_IPTU_CATEG_PROP=PARTICULAR BUFFER=Estadual ⇒
P1_Lote_IPTU_IPTUISENTO=nao (14.5, 99.7)
BUFFER=Infantil P1_Lote_IPTU_CATEG_PROP=PARTICULAR
P1_Lote_IPTU_IPTUISENTO=nao ⇒ P1_Lote_IPTU_DISTRITO=1 (17.8, 100.0)
```

Figura 5.7: Regras de associação espacial encontradas utilizando lotes e escolas.

Na primeira regra, $\text{BUFFER=Estadual} \Rightarrow \text{P1_Lote_IPTU_PEDOLOGI=NORMAL}$, apresenta que 14,6% dos lotes são próximos de escolas estaduais e possuem pedologia normal. Além disso, 81,1% dos lotes que são próximos de escolas estaduais também apresentam pedologia normal. A regra $\text{BUFFER=Estadual} \Rightarrow \text{P1_Lote_IPTU_IPTUISENTO=nao}$ informa que 14,7% são lotes próximos a escolas e não são isentos de IPTU, sendo que 81,6% dos lotes próximos a escolas também não são isentos de IPTU.

Além de ser possível encontrar regras de associação espacial para descrever a relação entre lotes, valores de IPTU e escolas, também foram encontradas regra de associação com informação espacial que dada alguma relação topológica, informa características dos lotes relacionadas com características de IPTU e escolas. O segundo experimento comprova que

é possível obter regras definindo como limitante para os dados a serem minerados alguma relação topológica, sendo neste caso lotes próximos as escolas. Algumas regras geradas nesse experimento podem ser observadas na Figura 5.8.

P1_Lote_IPTU_USOTERRE=NENHUM $\Rightarrow$ P1_Lote_IPTU_CATEG_PROP=PART (10.3, 99.7)
P1_Lote_IPTU_QTDEDIF=0 $\Rightarrow$ P1_Lote_IPTU_CATEG_PROP=PART (10.6, 93.8)

Figura 5.8: Regras de associação com informação espacial encontradas utilizando lotes e escolas.

A regra P1_Lote_IPTU_USOTERRE= NENHUM $\Rightarrow$ P1_Lote_IPTU_CATEG_PROP = PART (10.3, 99.7) informa que 10,3% dos lotes próximos a escolas não apresentam construção no terreno e são particulares. Além disso, 99,7% dos lotes próximos a escolas e que não apresentam construção no terreno também são particulares.

Lote, IPTU e IBGE

Ao buscar regras de associação espacial entre lotes e características de setores do IBGE, não foram encontradas regras que apresentassem alguma relações topológicas considerando o suporte mínimo 10% e confiança mínima de 80%. Já considerando lotes que estão inclusos nos setores do IBGE, encontrou-se regras como a exibida na Figura 5.9.

P1_Lote_IPTU_CATEG_PROP=PARTICULAR P1_Lote_IPTU_IPTUVALOR=1000-2500 $\Rightarrow$
P2_IBGE_Sector_DOM_COLETI=0 (10.6, 83.5)

Figura 5.9: Regra de associação com informação espacial encontrada utilizando lotes e IBGE.

Essa regra informa que 10,6% dos lotes contidos em um setor do IBGE sem domicílios coletivos, possuem categoria do proprietário particular e IPTU entre 1000 e 2500 reais. Além disso, 83,5% dos lotes que tem categoria particular e IPTU entre 1000 e 2500, também estão contidos em setores do IBGE sem domicílios coletivos.

Lote, IPTU e Rodovias

Novamente não foram encontradas regras de associação espacial considerando lotes próximos as rodovias. Já considerando regra de associação com informação espacial, pode-se extrair algumas regras, como as apresentadas na Figura 5.10. Contudo, essas regras contém conhecimentos esperados, por exemplo 11,2% de lotes próximos a rodovias estão nas esquinas e pagam IPTU, sendo que 99,5% de lotes próximos a rodovias e estão em esquinas, também pagam IPTU.

P1_Lote_IPTU_SITUACAO=ESQUINA $\Rightarrow$ P1_Lote_IPTU_IPTUISENTO=nao (11.2, 99.5)
P1_Lote_IPTU_VALOR_IPTU=0-500 P1_Lote_IPTU_QTDEDIF=0 $\Rightarrow$
P1_Lote_IPTU_DISTRITO=1 (58.4, 100.0)

Figura 5.10: Regras de associação com informação espacial encontradas utilizando lotes e rodovias.

Lote, IPTU, setor

Regras de associação espacial e regras de associação com informação espacial foram encontradas para descrever a relação entre lote, IPTU e setor. Na Figura 5.11 pode-se observar exemplos de regras de associação espacial encontradas.

TeCONTAINS=2 $\Rightarrow$ P1_Lote_IPTU_CATEG_PROP=PARTICULAR (13.7, 91.6)
TeCONTAINS=2 $\Rightarrow$ P1_Lote_IPTU_IPTUISENTO=nao (13.8, 91.7)
P1_Lote_IPTU_IPTUVALOR=0 TeCONTAINS=3 $\Rightarrow$ P1_Lote_IPTU_IPTUISENTO=sim (11.1, 99.9)

Figura 5.11: Regras de associação espacial encontradas utilizando lotes e setores.

Observa-se que 13,6% dos lotes estão contidos no setor 2 e a categoria do proprietário é particular. 91,6% dos lotes contidos no setor 2 também apresentam categoria de proprietário particular.

Serviços e Escolas

A relação entre serviços e escolas apresentou um pequeno número de regras, mostradas na Figura 5.12. Todas as regras apresentam um conhecimento óbvio, uma vez que no banco de dados espaciais somente existiam dados do distrito 1.

BUFFER=Fundamental $\Rightarrow$ P1_Servicos_lote_DISTRITO=1 (12.2, 100.0)
BUFFER=Fun_Inf $\Rightarrow$ P1_Servicos_lote_DISTRITO=1 (13.5, 100.0)
BUFFER=Estadual $\Rightarrow$ P1_Servicos_lote_DISTRITO=1 (27.0, 100.0)
BUFFER=Infantil $\Rightarrow$ P1_Servicos_lote_DISTRITO=1 (28.9, 100.0)
BUFFER=Fun_Inf $\Rightarrow$ P1_Servicos_lote_DISTRITO=1
P1_Servicos_lote_SETOR=1 (10.5, 100.0)
BUFFER=Infantil P1_Servicos_lote_SETOR=3 $\Rightarrow$
P1_Servicos_lote_DISTRITO=1 (13.2, 100.0)

Figura 5.12: Regras de associação espacial encontradas utilizando serviços e escolas.

Buscar padrões entre serviços oferecidos e características de setores de IBGE apresentou um elevado número de regras, não sendo possível analisar as regras encontradas. Contudo, é interessante observar que encontrar regras de associação espacial sofre influência da ordem que os plano de informação são selecionados no sistema. Buscar regras considerando características comuns de serviços relacionados no espaço com determinados setores do IBGE não apresentou nenhuma regra de associação espacial. Já características de setores do IBGE relacionadas com determinados serviços apresentou um número incontável pelo sistema de regras geradas.

Considerando as regra de associação com informação espacial, o número de regras geradas não foi tão elevado, podendo ser analisado. A Figura 5.13 mostra um exemplo de regra gerada. 10,4% dos setores do IBGE que sobrepõe lotes com serviços, o setor apresenta 10 cidadãos não alfabetizados com idades entre 20 e 49 anos e também o lote está localizado no setor 3 da cidade. Além disso, 90,2% dos setores do IBGE que sobrepõe lotes com serviços e apresentam 10 cidadãos não alfabetizados, também possuem lotes de serviço no setor 3.

$P1_IBGE_Setor_NALF_20_49=10 \Rightarrow P2_Servicos_lote_SETOR=3 (10.4, 90.2)$

Figura 5.13: Regra de associação com informação espacial encontrada utilizando serviços e IBGE.

Lote, IPTU e serviços

A relação existente entre lotes e serviços oferecidos próximos a lotes apresentou um elevado número de regras de associação espacial. Essa característica foi um empecilho para a avaliação das regras, não sendo possível avaliar o conteúdo do arquivo.

Hidrografia

Nos experimentos para buscar padrões próximos aos rios da cidade foram extraídas regras de associação espacial e regras de associação com informação espacial. Porém as regras não apresentavam um conhecimento interessante, segundo o especialista.

Corredor comercial

Não foi possível encontrar qualquer relação entre as características dos lotes e a proximidade de um corredor comercial. Também não foram encontradas regras de associação

espacial no experimento para verificar características comuns de setores do IBGE que possam ter qualquer relação topológica com corredores comerciais.

Esses resultados sugerem que não existam padrões, com suporte mínimo de 10% e confiança mínima de 80%, a serem descobertos considerando objetos espaciais próximos a corredores comerciais da cidade.

5.7 *Avaliação da metodologia e do módulo de pré-processamento*

As regras encontradas comprovam que é possível extrair regras de associação em dados espaciais, utilizando diferentes abordagens. Essas abordagens dependem do problema a ser pesquisado na mineração de dados. Apesar de ser possível encontrar regras com características espaciais, no estudo de caso apresentado as regras geradas na maioria representavam um conhecimento óbvio.

Outro ponto a ser destacado na análise dos resultados é a variação do número de regras geradas. Esse número está diretamente relacionado com o número de atributos de cada plano de informação. Se esse número for muito elevado, como no caso dos valores do IBGE, muitas regras são geradas.

Além das regras encontradas, é necessário avaliar características do processamento realizado. Para isso foram considerados os dados das Tabelas 5.4, 5.5 e 5.6. As características analisadas foram o tempo de processamento, a memória máxima utilizada, o tamanho do arquivo de relações topológicas gerado pelo módulo de pré-processamento e o volume de dados. Esse volume considera o tamanho de cada plano de informação utilizado. Quanto maior o número de objetos, maior o volume de dados. Para a análise dessas características também foram consideradas as abordagens propostas separadamente.

Nos experimentos realizados considerando a abordagem regras de associação espacial, o tempo de processamento apresentou um comportamento bem variado, sendo que nem sempre aumentou de acordo com o volume dos dados. Porém, de uma maneira geral pode-se observar, na Figura 5.14(a) que o gráfico para representar a relação entre o tempo e o volume de dados é semelhante ao de uma função exponencial. Assim, quanto maior o número de objetos espaciais maior será o tempo de processamento.

Já a memória máxima utilizada apresenta um crescimento inicial elevado, mas tende a estabilização como pode ser visualizado na Figura 5.15(a). O formato geral do gráfico para tamanho do arquivo também é semelhante ao gráfico de uma função exponencial, aumentando conforme cresce o tamanho do plano de informação.

O tempo de processamento considerando a abordagem regra de associação com informação espacial foi semelhante ao tempo encontrado na abordagem regras de associação espacial. Na Figura 5.14(b) pode-se observar um comportamento exponencial do relação entre o tempo

e o volume de dados. Esse comportamento também é observado na Figura 5.16(b), sobre tamanho do arquivo. A memória apresenta um comportamento diferente, apresentado na Figura 5.15(b), mostrando que tende a estabilização.

Essa análise mostra que o método empregado não é eficiente em relação ao tempo de processamento, devido a curva exponencial encontrada. Quanto maior o volume de dados maior será o tempo de processamento das informações, em ambas abordagens. Contudo esse resultado não inviabiliza o trabalho. Se não for necessário uma resposta rápida ao usuário, as relações topológicas podem ser extraídas e as regras de associação espacial e regra de associação com informação espacial podem ser encontradas.

Outro problema é o espaço livre em disco para a geração do arquivo de saída. Pode ser necessário possuir gigabytes livres para contemplar o tamanho do arquivo. A memória também é importante mas em nenhum momento ultrapassou 40Mb. Assim, apesar de limitações físicas do ambiente para executar o módulo e extrair regras de associação utilizando dados georreferenciados, é possível encontrar padrões utilizando a metodologia, o módulo e as abordagens de formato de arquivo de saída propostos.

(a)

(b)

Figura 5.14: Gráficos sobre tempo de processamento

(a)

(b)

Figura 5.15: Gráficos sobre uso de memória.

(a)

(b)

Figura 5.16: Gráficos sobre tamanho arquivo gerado.

Conclusão

Neste trabalho foi proposta uma metodologia de extração de padrões em dados georreferenciados e o módulo de pré-processamento de relações topológicas para extração de regras de associação espacial. A metodologia sugerida foi adaptada do processo de extração de conhecimento, adicionando a atividade de extração de relações topológicas no pré-processamento dos dados dados georreferenciados. Além disso, maneiras diferentes de organizar os dados para tratar diferentes enfoques da extração de regras foram propostas, com a finalidade de viabilizar a mineração de regras de associação em dados georreferenciados. Os formatos de organização dos dados propostos foram:

- encontrar regras de associação com atributos descritivos de um determinado plano de informação agregados a relações topológicas dos objetos espaciais desse plano de informação com objetos espaciais de outro plano de informação, denominadas regras de associação espacial;
- encontrar regras de associação que descrevem relações entre os atributos de dois planos de informação utilizando somente objetos que obedecem determinada relação topológica, denominadas regras de associação com informação espacial;
- encontrar regras de associação utilizando programação lógica indutiva, não explorada neste trabalho.

Para validar a metodologia de extração de regras de associação em dados georreferenciados e o módulo de pré-processamento de relações topológicas foi realizado um estudo de

caso com dados reais da cidade de Jaboticabal. Esse estudo mostrou que a metodologia e o módulo são viáveis. As regras foram obtidas, considerando diferentes aspectos. A avaliação dos resultados mostrou que o tempo de processamento tende a ser elevado. Esse tempo, porém, não prejudica o módulo pois o foco é encontrar as regras e não ser um sistema com uma rápida resposta ao usuário.

Outro ponto importante observado na realização do estudo de caso, para a aplicação da metodologia e do módulo, foi a dependência de um especialista orientando o processo. Sem o conhecimento prévio do assunto e definição de objetivos para se encontrar o padrão é muito difícil obter regras úteis.

Além disso, os dados de origem são fundamentais para a busca de padrões. Somente quando utilizado o banco de dados completo da cidade de Jaboticabal foi possível obter um resultado válido no processo de extração de regras. Quando foram utilizados os dados da região piloto nenhum padrão foi obtido, devido a erros contidos nos dados.

O estudo de caso e as regras obtidas nesse estudo comprovaram o objetivo proposto neste trabalho. Os dados georreferenciados podem ser pré-processados para a mineração de regras de associação de maneira que não é necessário alterar algoritmos de mineração, somente adaptar a metodologia de extração de regras para incluir, na etapa de pré-processamento, a extração de relações topológicas. A aplicação de diferentes formatos nos dados pré-processados também confirmou que é possível obter regras com significados diferentes a partir de dados georreferenciados, permitindo associar dados cadastrais de plano de informação diferentes em função de relações topológicas.

6.1 Principais contribuições

Como resultado deste trabalho, além de comprovar o objetivo proposto, obteve-se:

- o módulo de pré-processamento de relações topológicas foi disponibilizado para encontrar relações topológicas em dados georreferenciados e converter essas relações em diferentes formatos de saída;
- metodologia para extração de regras de associação utilizando dados georreferenciados, isto é, pré-processar as relações topológicas antes de aplicar o algoritmo de extração de regras;
- os formatos de organização dos dados georreferenciados para auxiliar o processo de mineração de dados, isto é, as abordagens propostas para encontrar regras de associação espacial e regras de associação com informação espacial;

- o estudo de caso realizado utilizando dados reais, em que as regras encontradas podem beneficiar os usuários no planejamento urbano da cidade Jaboticabal.

6.2 *Trabalhos futuros*

Esse trabalho representou um estudo inicial da extração de regra de associação em dados georreferenciados. Muitos problemas surgiram e novas abordagens para tratar a extração de regras em dados dados georreferenciados podem ser pesquisadas.

Como continuação deste trabalho, é interessante considerar a utilização de algoritmos de extração de regras de associação utilizando lógica relacional. É esperado que as regras geradas sejam mais interessantes, uma vez que relações topológicas tratam de um relacionamento nos dados e a lógica relacional é uma técnica que lida com relacionamento de dados.

Além disso, pode-se realizar experimentos utilizando a saída do módulo para regras de associação com informação espacial em outras abordagens de mineração de dados, como Classificação e *Clustering*. Também pode-se utilizar taxonomias para generalizar regras muito específicas.

Outro ponto é a avaliação das regras. Neste trabalho foi realizada a avaliação das regras considerando somente suporte e confiança. Seria interessante reavaliar as regras em função de outras medidas disponíveis. A partir disso, estudar e desenvolver filtros para auxiliar na busca por regras de acordo com medidas de avaliação objetivas e subjetivas.

Também pode ser verificada a utilização do suporte mínimo para encontrar os *itemsets* nos algoritmos de mineração. Foi observado que o suporte mínimo deve ser baixo para encontrar padrões que possuam característica espacial. Estudar e propor uma nova forma de contar os *itemsets* baseada nas relações topológicas poderia trazer benefícios para a comunidade de pesquisa de regras de associação e dados georreferenciados.

Referências

- Agrawal, R., H. Mannila, R. Srikant, H. Toivonen, and I. Verkamo (1996). Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, pp. 307–317. American Association for Artificial Intelligence.
- Agrawal, R. and R. Srikant (1994). Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo (Eds.), *Proceedings 20th International Conference on Very Large Data Bases, VLDB*, Santiago, Chile, pp. 487–499. Morgan Kaufmann.
- Batista, G. E. A. P. A. (2003). Pré-processamento de dados em aprendizado de máquina supervisionado. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - USP São Carlos.
- Borges, K. and C. Davis (2002). Modelagem de dados geográfico. Documento eletrônico - INPE. <http://www.dpi.inpe.br/gilberto/livro/introd/cap4-modelos.pdf>. Acessado em: 11/08/2005.
- Brinkhoff, T., H.-P. Kriegel, and B. Seeger (1993). Efficient processing of spatial joins using r-trees. In *Proceedings of the 1993 ACM-SIGMOD International Conference on Management of Data*, Washington, D.C., United States, pp. 237–246. ACM Press.
- Cabena, P., P. Hadjinian, R. Stadler, L. Verhees, and A. Zanasi (1997). *Discovering data mining from concept to implementation*, pp. 56. Upper Saddle River, New Jersey: Prentice Hal.
- Câmara, G. and C. Davis (1999). Introdução à ciência da geoinformação. In G. Câmara, C. D. Jr., A. M. V. Monteiro, J. A. Paiva, and J. C. d’Alge (Eds.), *Geoprocessamento: teoria e aplicações*. Livro on-line, Instituto Nacional de Pesquisas Espaciais, <http://www.dpi.inpe.br/gilberto/livro/introd/cap1-introducao.pdf>. Acessado em: 16/10/2003.
- Camara, G. and A. M. V. Monteiro (2002). Conceitos básicos da ciência da

- geoinformação. Documento eletrônico - INPE. Apostila do INPE no CD do Spring. <http://www.dpi.inpe.br/gilberto/livro/introd/cap2-conceitos.pdf>. Acessado em 11/08/2005.
- Chawla, S., S. Shekhar, W. Wu, and U. Ozesmi (1999). Modeling spatial dependencies for mining geospatial data: An introduction. In H. Miller and J. Han (Eds.), *Geographic data mining and Knowledge Discovery (GKD)*. http://www.siam.org/meetings/SDM01/pdf/sdm01_28.pdf. Acessado em: 09/04/2004.
- Clementini, E., P. D. Felice, and P. van Oosterom (1993). A small set of formal topological relationships suitable for end-user interaction. In *Proceedings of the Third International Symposium on Advances in Spatial Databases*, pp. 277–295. Springer-Verlag. Lecture Notes in Artificial Intelligence.
- Dehaspe, L. and H. Toironen (2000). Discovery of relational association rules. New York, NY, USA, pp. 189–208. Springer-Verlag New York, Inc.
- Egenhofer, M. J. and J. R. Herring (1990). Categorizing binary topological relations between regions, lines, and points in geographic databases. Technical Report 94, National Center for Geographic Information & Analysis. <http://www.spatial.maine.edu/max/9intReport.pdf>. Acessado em: 02/12/2003.
- Ester, M., H. Kriegel, and J. Sander (2001). Algorithms and applications for spatial data mining. In H. Miller and J. Han (Eds.), *Geographic Data Mining and Knowledge Discovery*. Taylor and Francis. <http://www.dbs.informatik.uni-muenchen.de/Publikationen/Papers/Chapter7.revised.pdf>. Acessado em: 19/08/2003.
- Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (1996). From data mining to knowledge discovery: Na overview. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (Eds.), *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA, pp. 6–22. American Association for Artificial Intelligence.
- Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth (1996). The KDD process for extracting useful knowledge from volumes of data. In *Communication of the ACM*, Volume 29, pp. 27–34.
- Han, J., J. Pei, and Y. Yin (2000). Mining frequent patterns without candidate generation. In W. Chen, J. Naughton, and P. A. Bernstein (Eds.), *2000 ACM SIGMOD International Conference on Management of Data*, pp. 1–12. ACM.
- Horst, P. S. (1999). Avaliação do conhecimento adquirido por algoritmos de aprendizado de máquina utilizando exemplos. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação - USP São Carlos.
- IBGE (2005). Informações sobre cidade. <http://www.ibge.gov.br/cidadesat/default.php>. Acessado em 18/10/2005.

- Koperski, K. and J. Han (1995). Discovery of spatial association rules in geographic information databases. In *Proceedings of the 4th International Symposium Advances in Spatial Databases*, pp. 47–66. Springer-Verlag.
- Koperski, K. (1999). *A progressive refinement approach to spatial data mining*. Ph. D. thesis, Simon Fraser University.
- Lee, H. D. (2005). Seleção de atributos importantes para a extração de conhecimento de bases de dados. Tese de Doutorado, Instituto de Ciências Matemáticas e de Computação - USP São Carlos.
- Li, D., J. Deogun, and S. Harms (2003). Interpolation techniques for geo-spatial association rule mining. In *LNAI*, Volume 2639. Springer-Verlag.
- Maguire, D. J., M. F. Goodchild, and D. W. Rhind (1991). *Geographical information systems, principles and applications*. Longman Scientific Technical.
- Malerba, D., F. A. Lisi, and S. Harms (2001). An ILP method for spatial association rule mining. In *Proceedings Workshop Multi-relational Data Mining MRDM*, pp. 18–29. <http://citeseer.nj.nec.com/488801.html>. Acessado em: 12/07/2003.
- Mennis, J. and J. W. Liu (2003). Mining association rules in spatio-temporal data. In *Proceedings of the 7th International Conference on GeoComputation*, United Kingdom. http://www.geocomputation.org/2003/Papers/Mennis_Paper.pdf. Acessado em: 5/12/2003.
- Monard, M. C. and J. A. Baranauskas (2003). Conceitos sobre aprendizado de máquina. In S. O. Rezende (Ed.), *Sistemas Inteligentes: Fundamentos e Aplicações*, Barueri, SP, pp. 96–97. Manole.
- Ortiz, G. (2005). Qué son los sistemas de información geográfica. <http://recursos.gabrielortiz.com/index.asp?Info=012>. Acessado em: 08/08/2005.
- Park, J. S., M.-S. Chen, and P. S. Yu (1995, 22–25). An effective hash based algorithm for mining association rules. In M. J. Carey and D. A. Schneider (Eds.), *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, San Jose, California, pp. 175–186.
- Pei, J., J. Han, and R. Mao (2000). Closet: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp. 21–30.
- Prati, R. C. (2003). O framework de integração do sistema discover. Dissertação de Mestrado, Instituto de Ciências Matemáticas e de Computação - USP São Carlos.
- Rezende, S. O., J. B. Pugliesi, E. A. Melanda, and M. F. Paula (2003). Mineração de dados. In S. O. Rezende (Ed.), *Sistemas Inteligentes: Fundamentos e Aplicações*, Barueri, SP, pp. 307–335. Manole.

- Ribeiro, G. and G. Câmara (2002). Arquitetura de sistemas de informação geográfica. Documento eletrônico - INPE. Apostila do INPE no CD do Spring. <http://www.dpi.inpe.br/gilberto/livro/introd/cap3-arquitetura.pdf>. Acessado em 11/08/2005.
- Samet, H. (1990). *The design and analysis of spatial data structures*. AddisonWesley.
- Schneider, M. (1997). Spatial data types for database system: Finite resolution geometry for geographic information system. Lecture Notes in Artificial Intelligence.
- Spring (2005a). Documentação on-line do sistema de informação geográfica spring. <http://www.dpi.inpe.br/spring/portugues/index.html>. Acessado em 11/08/2005.
- Spring (2005b). Help da ferramenta spring.
- Terralib (2005). Documentação on-line do terralib. <http://www.terralib.org>. Acessado em 11/08/2005.
- TerraView (2005). Documentação on-line do visualizador de dados georreferenciados terraview. <http://www.dpi.inpe.br/terraview/index.php>. Acessado em 11/08/2005.
- Webb, G. I. (1995). OPUS: An efficient admissible algorithm for unordered search. *Journal of Artificial Intelligence Research* 3, 431–465. <http://citeseer.nj.nec.com/35589.html>.
- Weiss, S. M. and N. Indurkha (1998). *Predictive data mining: a practical guide*. San Francisco, CA: Morgan Kaufmann.
- Zaki, M. J. and C.-J. Hsiao (2002). CHARM: An efficient algorithm for closed itemset mining. Acessado em: 23/10/2003.