UNIVERSIDADE FEDERAL RURAL DO SEMIÁRIDO
UNIVERSIDADE DO ESTADO DO RIO GRANDE DO NORTE
MESTRADO EM CIÊNCIA DA COMPUTAÇÃO

ARTUR LUIZ TORRES DE OLIVEIRA

TMIGRAF'S - UM MODELO DE MINERAÇÃO DE TRAJETÓRIAS UTILIZANDO GRAFOS DIRECIONADOS.

MOSSORÓ – RN

2012

Artur Luiz Torres de Oliveira

TMIGRAF's - Um Modelo de Mineração de Trajetórias Utilizando Grafos Direcionados.

Dissertação de Mestrado submetida ao Programa de Pós-graduação em Ciência da Computação da Universidade do Estado do Rio Grande do Norte e Universidade Federal Rural do Semi-Árido como parte dos requisitos necessários para obtenção do grau de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Angélica Félix de Castro
Co-Orientador: Prof. Dr. Francisco Chagas de Lima Júnior

MOSSORÓ, RN
2012

Artur Luiz Torres de Oliveira

TMIGRAF's - Um Modelo de Mineração de Trajetórias Utilizando Grafos Direcionados.

Dissertação apresentada ao Programa de Pós-graduação em Ciência da Computação para obtenção do grau de Mestre em Ciência da Computação Aprovada em Março/ 2012

Profª. Angélica Félix de Castro, D.Sc
Orientadora
Universidade Federal Rural do Semi-Arido - UFERSA

Profº. Francisco Chagas de Lima Júnior, D.Sc
Co-Orientador
Universidade do Estado do Rio Grande do Norte - UERN

Profº. Marcelino Pereira dos Santos Silva, D.Sc
Membro
Universidade do Estado do Rio Grande do Norte - UERN

Profº. Marco Antonio de Oliveira Domingues, D.Sc
Membro
Instituto Federal de Pernambuco - IFPE

MOSSORÓ, RN, 2012

Dedicatória

Dedico esta dissertação primeiramente a DEUS todo poderoso que sempre está comigo e me ajuda na minha caminhada e às duas pessoas mais importantes da minha vida, minha mãe Ana Lucia Torres de Oliveira e ao meu pai Luiz Carlos Carvalho de Oliveira pela educação, carinho e dedicação como conduziram meus passos até os dias de hoje.

Agradecimentos

Agradecer primeiramente a **DEUS**, pelo dom da vida, pelo sol de todos os dias, o ar que respiro, minha família, meus amigos, meu emprego, enfim minha vida como um todo.

Aos meus pais, **Luiz Carlos Carvalho de Oliveira** e **Ana Lucia Torres de Oliveira** por tudo de bom que me proporcionaram desde o primeiro suspiro de vida até os dias atuais e a quem eu dedico este trabalho.

À minha querida amiga, coordenadora, professora e orientadora **Profª Dra Angélica Felix de Castro** pelas orientações, paciência e intermináveis conversas na sala da coordenação. Seus ensinamentos e motivação foram fundamentais para a conclusão de mais essa etapa da minha vida.

A todos os Professores do MCC. Por cada aula ministrada, cada cobrança, cada minuto que dispensaram a minha pessoa na busca incessante no objetivo de transmitir o saber.

As Minhas irmãs **Polyana Torres de Oliveira** e **Cynthia Torres de Oliveira** pelo apoio a cada passo dado e pelo companheirismo que sempre me foi dispensado.

Aos meus amigos e colegas da UFERSA/UERN, em especial os da turma de 2009, 2010 e 2011, Pelas inúmeras batalhas vencidas juntos, por cada noite de sono, por cada projeto em equipe, por cada minuto que me foi reservado com a mais sincera atenção.

Ao IFPI - Instituto Federal do Piauí - Por ter me concedido a liberação e financiado os meus estudos. Espero poder retribuir a confiança que me foi depositada nesses dois anos com muita dedicação ao meu oficio de educar.

À UFERSA/UERN - Instituições que, através dos seus quadros docentes e administrativos, foram responsáveis pela minha formação acadêmica. Com certeza, em uma oportunidade de retorno vou poder chama-las de casa.

A Instituição **Clube de Regatas Vasco da Gama**, pelas alegrias que me deu durante esses dois anos de mestrado. Momentos de descontração às quartas-feiras e finais de semana.

E aqueles que contribuíram de alguma forma para realização deste trabalho. **Meu muito Obrigado!**

Resumo

O presente trabalho apresenta uma proposta de modelo de mineração de objetos espaço-temporais utilizando técnicas de KDD (*Knowledge-Discovery in Databases*) associados ao poder de visualização de padrões dos grafos direcionados. A maioria das pesquisas acerca de descoberta de conhecimento em objetos espaciais utiliza as técnicas de clusterização em busca de ROI´s (Regiões de interesse). Essas adaptações de técnicas a novos tipos de dados geralmente levam em consideração apenas uma dimensão do problema, o espaço. O modelo TMIGRAF´s, por sua vez, ataca tanto a dimensão espaço como a dimensão tempo, proporcionando uma análise mais completa das características do problema. Outra vantagem do modelo e a aplicação de técnicas de mineração de dados de aprendizagem não supervisionada de associação. Devido às características das trajetórias, a utilização das técnicas de associação e mais especificamente, de padrões sequenciais frequentes proporcionam ao modelo um poder de análise mais rebuscado descobrindo padrões que podem ser muito úteis na tomada de decisões estratégicas. Uma base de dados foi submetida ao modelo visando a demonstração da eficiência do mesmo e padrões semanticamente relevantes são apresentados.

Palavras-chave: Mineração de Dados, Objetos Espaço-Temporais, Teoria dos Grafos, Padrões Sequenciais Frequentes, Regras de Associação.

Abstract

This work proposes a mining model objects spatio-temporal using techniques KDD (knowledge-discovery indatabases) associated with the power to see patterns of directed graphs. Most research on knowledge discovery in spatial objects uses clustering techniques to search for ROI's (Regions of Interest). These adaptations of techniques to new data types usually take into account only one dimension of the problem, the space. The model's TMIGRAF, in turn, attacks both the size space and time dimension, providing a more complete analysis of the characteristics of the problem. Another advantage of the model and application of data mining techniques for unsupervised learning of association. Because the characteristics of the trajectories, using the techniques of association and more specifically, frequent sequential patterns give the model a more refined analysis can discover patterns that can be very useful in making strategic decisions. A data base was submitted to the model in order to demonstrate the efficiency of it and semantically relevant standards are presented.

keywords: Data Mining, spatio-temporal Objects, Graph Theory, Frequent Sequential Patterns, Association Rules.

Lista de Símbolos e Abreviaturas

Abreviatura	Descrição	Pag
KDD	Knowledge Discovery in Databases	02
$WI-FI$	Marca registrada da Wi-Fi Alliance	02
$GPRS$	General Packet Radio Services	02
GPS	Global Positioning System	02
$SGBD$	Sistema de Gerenciamento de Baco de Dados	06
$RFID$	Radio-Frequency IDentification	09
$GeoPKDD$	Geographic Privacy-aware Knowledge Discovery and Delivery	17
ROI	Regions of Interest	17
$TRACTS$	Trajectory Classification using Time Series	18
$P2P$	Redes Peer-to-peer	19
$MANETS$	Mobile ad-hoc Network	19
PL/SQL	Procedural Language/Structured Query Language	25
IBM	International Business Machines	28
GSP	Generalized Sequential Pattern	34
$SPMF$	Sequential Pattern Mining Framework	36
$JUNG$	Java Universal Network/Graph Framework	36
$WEKA$	Waikato Environment for Knowledge Analysis	36
API	Application Programming Interface	36
$SAMU$	Serviço de Atendimento Móvel de Urgência	47

Lista de Figuras

2.1 Ponto. Fonte: Autoria própria 7
2.2 Linha. Fonte: Autoria própria 7
2.3 Polígono. Fonte: Autoria própria 7
2.4 Superfície. Fonte: Autoria própria 7
2.5 Dados Contínuos. Fonte: (BOGORNY, 2010) 7
2.6 Relações Topológicas. Fonte: Autoria própria 8
2.7 Representação espacial de uma trajetória. Fonte: (GIANNOTTI; TRASARTI, 2000). 10
2.8 Etapas do Processo do KDD. Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SHYTH, 1996). 11
2.9 Áreas em que a Mineração de dados se apropria dos recursos. Fonte: (HAN; KAMBER, 2001). 12
2.10 Esta figura é um desenho do grafo cujos vértices são t , u, v , w, x, y , z e cujas arestas são vw, uv , xw, xu, yz e xy . Fonte: (FEOFILOFF; KOHAYAKAWA; WAKABAYASHI, 2007). 14
2.11 Digrafo. Fonte: (CUBAS, 2008). 14
2.12 Matriz de Adjacência do Grafo. Fonte: (CUBAS, 2008). 14
2.13 Matriz de Adjacência do Digrafo. Fonte: (CUBAS, 2008). 15
2.14 SubGrafo Gerador. Fonte: (CUBAS, 2008). 15
2.15 SubGrafo Induzido. Fonte: (CUBAS, 2008). 16
2.16 Isomorfismo de Grafos. Fonte: (CARVALHO, 2009). 16
2.17 Função Bijetora. Fonte: Autoria própria. 17
2.18 Geographic Privacy-aware Knowl-edge Discovery and Delivery. Fonte: (GIANNOTTI; TRASARTI, 2000). 18

3.1 Modelo TMIGRAF's. Fonte: Autoria própria 21
3.2 Representação do conjunto de pontos das trajetórias. Fonte: (BOGORNY, 2010) 22
3.3 Representação dos dados de uma trajetória. Fonte: Autoria própria 24
3.4 Exemplo dos dados submetidos ao processo de Transformação. Fonte: Autoria própria . . . 25
3.5 Representação dos dados de um grafo. Fonte: Autoria própria 25
3.6 Mapeamento dos Itens da Base de Produtos de um supermercado Fonte: (AMO, 2003). . . . 27
3.7 Base de Dados Transacional com os ItemSet´s. Fonte: (AMO, 2003) 27
3.8 Frequências de ocorrências dos ItemSet´s. Fonte: (AMO, 2003) 28
3.9 Pseudo Código do APRIORI. Fonte: (ARBEX; SABOREDO; MIRANDA, 2004) 29

3.10 Função APRIORI_Gen. Fonte: (AMO, 2003) 30
3.11 Base de Dados de Sequência de compras de clientes. Fonte: (AMO, 2003) 32
3.12 Base de Dados de Sequência com informações de Tempo. Fonte: Autoria Própria. 33
3.13 Resultado visual dos padrões sequenciais analisados . Fonte: Autoria Própria. 34
3.14 Iteração 1 do GSP - suporte mínimo de 50%. Fonte: (DEVêZA, 2011). 36
3.15 Iteração 2 do GSP. Fonte: (DEVêZA, 2011). 36
3.16 Aplicação de manipulação de arestas em JUNG. Fonte: (BERNSTEIN, 2009). 38
3.17 Aplicação de análise de tráfego na malha viária. Fonte : (SANTOS, 2010). 38

4.1 Schema conceitual de trajetórias. Fonte : (ALVARES et al., 2007) 40
4.2 Parte do Schema conceitual das características geográficas. Fonte : (ALVARES et al., 2007) . 41
4.3 Locais escolhidos para montar a base de deslocamento de cidades. Fonte: Autoria própria . 41
4.4 Base de dados das trajetórias. Fonte: Autoria própria 42
4.5 Base de dados em formato de grafos. Fonte: Autoria própria 42
4.6 Bases de dados em formato de sequências temporais e sequencias simples. Fonte: Autoria própria 43
4.7 Itens fequentes gerados pelo algoritmo APRIORI aplicado às trajetórias de participantes. Fonte: Autoria própria 44
4.8 Regras de associação geradas pelo algoritmo APRIORI aplicado às trajetórias de participantes. Fonte: Autoria própria 44
4.9 Padrões sequenciais frequentes com restrição de tempo. Fonte: Autoria própria 45
4.10 Padrões sequencial com restrição de tempo L2 minerado em formato de grafo. Fonte: Autoria própria 46
4.11 Padrões sequencial com restrição de tempo L3 minerado em formato de grafo. Fonte: Autoria própria 46

Sumário

1 Introdução **1**

1.1 Introdução 1

1.2 Contextualização 2

1.3 Motivação 3

1.4 Objetivo Geral 4

1.5 Objetivos específicos 4

1.6 Estrutura da Dissertação 4

2 Fundamentação Teórica **6**

2.1 Dados Geográficos 6

2.2 Relacionamentos Espaciais 8

2.3 Objetos Espaço-Temporais 9

2.4 Representação das trajetórias 9

2.5 Descoberta de Conhecimento em Bancos de Dados 10

2.5.1 Métodos de Mineração de Dados 11

2.6 Teoria dos Grafos 13

2.6.1 Representação dos Grafos 14

2.6.2 Subgrafos 15

2.6.3 Isomorfismo de Grafos 16

2.7 Trabalhos Relacionados 17

3 Projeto e Implementação do TMIGRAF's **20**

3.1 Introdução 20

3.2 Modelo TMIGRAF's 21

3.2.1 Pré-Processamento 22

3.2.2 Processo de Transformação 24

3.2.3 Processo de Mineração de Sequências 26

3.3 Ferramentas 37

3.3.1 SPMF - *Sequential Pattern Mining Framework* 37

3.3.2 JUNG - *Java Universal Network/Graph Framework* 37

4 Estudo de Caso **39**

4.1 Base de dados do estudo de caso 39

4.2 Participantes de Conferência 39

4.2.1 Modelo Conceitual Utilizado 39

4.2.2 Dados Utilizados 41

4.2.3 Aplicação do modelo TMIGRAFs 42

4.2.4 Resultados 43

5 Considerações finais e Trabalhos Futuros **48**

Referências Bibliográficas **50**

Capítulo 1

Introdução

Este capítulo introdutório descreve os principais fatores para realização desse trabalho. Ele descreve os objetivos, a justificativa, a motivação, a metodologia aplicada e toda a estrutura da dissertação.

1.1 Introdução

Com a enorme quantidade de dados armazenados nas bases de dados, torna-se cada vez mais importante o desenvolvimento de métodos e ferramentas que nos propiciem uma análise e extração de conhecimento sobre os mesmos. Este desenvolvimento se justifica pelas preciosas informações que podemos extrair dessas bases e transformá-las em diferencial nos seus respectivos domínios, sejam eles acadêmico, empresarial, comercial, dentre outros. Segundo Carvalho (2005), mineração de dados pode ser entendida como o uso de técnicas automáticas para a exploração de grandes quantidades de dados de forma a descobrir novos padrões e relações que devido ao volume de dados, não seriam facilmente descobertos pelo ser humano. Além do crescimento dessas bases, novas aplicações veem surgindo e com elas novos tipos de dados.

É visível como o número de aplicações que manipulam informações georreferenciadas[1] vem evoluindo suas tecnologias de coleta e tratamento dos dados espaciais. Essa evolução tem ocasionado um aumento da quantidade de dados espaciais, fazendo com que seu volume cresça muito além da capacidade humana de processá-los para extrair informações relevantes e conhecimentos úteis que possam apoiar a tomada de decisão. Dessa forma, as técnicas de mineração de dados surgem como alternativa de descoberta de conhecimento também para essas bases.

Neste contexto de dados espaciais surgem as trajetórias. Trajetórias são objetos espaço-temporais. Segundo Güing, Almeida e Ding (2006) e Wolfson et al. (1998), as trajetórias tem sido consideradas como os caminhos seguidos por um objeto em movimento no espaço e no tempo. Cada ponto nesse caminho

[1] Georreferenciar um dado é o processo de atribuir coordenadas a esse dado do sistema de referência que se deseja utilizar.

representa uma posição no espaço em um determinado instante de tempo.

Segundo Kang, Kim e Li (1997), dentre as técnicas de mineração de dados existentes, a mais difundida para a detecção de padrões em grandes bancos de dados espaciais é a análise de agrupamentos. Esta técnica consiste em organizar um conjunto de objetos em grupos, sendo que objetos do mesmo grupo são similares uns aos outros, ao contrário de objetos de grupos diferentes. Porém, devido à incorporação cada vez maior de semântica aos dados coletados, outras técnicas começam a despontar como boas alternativas de descoberta de conhecimento em bases de dados espaciais.

O presente trabalho se propõe a fornecer, um modelo de mineração de dados espaço-temporais utilizando-se de algumas técnicas de *Knowledge Discovery in Database* (KDD), vantagens dos algoritmos de mineração de dados não supervisionado de associação e conceitos importantes da teoria dos grafos. O estudo segue sob forma de pesquisa bibliográfica, da proposta e implementação do modelo TMIGRAF's e da aplicação do modelo a alguns conjuntos de dados para descoberta de novos conhecimentos sobre essas bases.

1.2 Contextualização

Hoje em dia, cada vez mais os dispositivos móveis estão presentes em nossas vidas. O uso desses dispositivos juntamente com as tecnologias Wireless[2], tais como Bluetooth, Wi-FI[3] e GPRS[4] vem proporcionando uma verdadeira revolução no cotidiano e no comportamento da humanidade.

A presença constante desses dispositivos associados a sensores de posicionamento tipo GPS [5] possibilitam a geração de uma gigantesca massa de dados que por sua vez podem ser armazenada e posteriormente analisada. Vários domínios de aplicações podem se beneficiar dessa realidade, como deslocamento de pessoas, de veículos, migração de animais, percurso de furacões, análise de criminalidade, entre outros.

Segundo Monreale et al. (2009), a difusão dessas tecnologias onipresentes no mundo, garante que haverá um aumento significativo na disponibilidade de grandes quantidades de dados sobre as trajetórias individuais e com precisão crescente em termos de localização e características associadas. Esses dados podem servir como um grande nicho de pesquisa para comunidade acadêmica e civil como um todo.

Apesar dessa nova realidade, os dados de trajetórias ainda são pouco conhecidos. Segundo Guo, Liu e Jin (2010), é difícil visualizar e extrair padrões significativos a partir de grandes massas de dados de trajetória. Um dos principais desafios é o de caracterizar, comparar e generalizar as trajetórias para encontrar padrões e tendências.

A análise dessas bases de dados de trajetórias permitirá que se observe tendências, que se detecte pontos com maior incidência dessas trajetórias, gargalos de movimentos em pontos críticos e pontos

[2]Tecnologia de transmissão de dados sem fio.

[3]Tecnologia utilizada por dispositivos para acessso a rede sem fio locais

[4]Tecnologia de aumento de taxa de transferência em redes de telefonia móvel

[5]Sistema de Posicionamento Global.

ociosos; comportamentos de deslocamentos frequentes, sequencias de deslocamento padrão, entre outras ações.

1.3 Motivação

Hoje em dia, é cada vez mais preocupante o problema de mobilidade nas cidades e entre elas. Estruturas governamentais e grandes blocos econômicos se esforçam em reuniões e gastos de milhares de dólares na busca de soluções para o problema de infra-estrutura de deslocamento dentro dos seus cotidianos ou mesmo fora das suas jurisdições. O modelo internacional de industrialização e a globalização econômica estabeleceram um novo patamar de exigências para deslocamento de indivíduos de diferentes cidades do mundo. Atrelado ao fato que cada vez mais pessoas precisam se deslocar rapidamente, o trânsito das grandes metrópoles se torna cada vez mais violento e um planejamento para a diminuição desses prejuízos tanto de capital financeiro como pessoal se torna fundamental. Segundo o SIM - Sistema de Informação de Mortalidade ligado ao Ministério da Saude, na década 1998/2008, registrou-se um total de 38.273 mortes nos diversos tipos de acidentes de trânsito. Esse número pode ser considerado muito elevado, superior até ao número de mortes em muitos dos conflitos armados com duração semelhante. Esse dado coloca o Brasil em 10^{o} lugar entre os 100 países analisados no relatório do estudo divulgado em 24 de fevereiro de 2011. O gerenciamento dessa mobilidade em massa de pessoas e produtos torna-se de vital importância para planejadores e tomadores de decisões.

O problema de gerenciamento de mobilidade não se resume apenas à seara dos deslocamentos de cargas e pessoas. Outros domínios se encaixam nesse contexto. É visível um aumento substancial de esforços de governos do mundo todo com a preservação do planeta. Nesse sentido, uma análise dos deslocamentos de espécies animais, principalmente nas suas épocas de migração, nos fornecem padrões de comportamento importantes que podem servir de subsídios para tomada de ações estratégicas no sentido de preservar determinada espécie.

Ainda no domínio da natureza, mais voltado à segurança, é cada vez mais frequente a ocorrência de desastres naturais que destroem a vida de milhares de pessoas mundo afora. Dados acerca do deslocamento de furacões, Tsunames, Inundações de rios, maremotos, entre outros podem fazer toda a diferença para autoridades competentes em tomada de decisões de construção de estruturas de prevenção ou mesmo de evacuações rápidas de grandes aglomerados populacionais, podendo resultar assim no salvamento de milhares de vidas.

Portanto a motivação da presente dissertação surge do interesse de investigar mais detalhadamente o crescente número de bases de dados de deslocamentos (Trajetórias) dos mais diferentes domínios de aplicação, para tentar entender como esses objetos se comportam e a partir desse conhecimento, proporcionar subsídios inteligentes para tomada de decisões de gestores dentro das suas demandas.

1.4 Objetivo Geral

O estudo proposto por essa dissertação pretende provocar o interesse em pesquisadores envolvidos em tarefas de modelagem, desenvolvimento, gerência, utilização e análise de projetos de descoberta de conhecimento em bancos de dados espaciais. Assim, o resultado deste trabalho terá sua validade para todos os profissionais envolvidos, de alguma forma, em projetos de mineração de dados espaciais.

Os objetivos gerais do trabalho são:

- Propor um modelo de descoberta de conhecimento em bases de dados de objetos espaço-temporais;
- Implementar esse modelo;
- Encontrar padrões e conhecimento nos dados de objetos espaço-temporais submetidos a esse modelo.

1.5 Objetivos específicos

A dissertação visa contribuir para o melhor entendimento das trajetórias (objetos espaço-temporais), seu comportamento ao longo do tempo, os pontos de trajetórias mais visitados concomitantemente e as sequencias de trajetos que são comumente utilizados pelos objetos. Para isso, são aplicadas tecnologias de bancos de dados, aprendizagem de máquina e análise de padrões resultantes. Tais tecnologias se propõem a oferecer suporte computacional para o desenvolvimento do modelo, proporcionando uma maior precisão da descoberta do conhecimento que será útil para o entendimento desses objetos e a tomada de decisão embasada por esses padrões.

De acordo com o domínio dos dados, os objetivos específicos são:

- Determinar quais pontos foram mais visitados pelas trajetórias;
- Verificar quais pontos foram mais visitados em um mesmo trajeto e com espaço de tempo similares;
- Encontrar padrões sequenciais frequentes de trajetos presentes nas trajetórias.

1.6 Estrutura da Dissertação

A presente dissertação está organizada em 5 capítulos, incluindo-se este introdutório. O capítulo 2 configura o estado da arte da pesquisa e tem como objetivo apresentar os principais conceitos envolvidos com o tema da dissertação, sob forma de uma revisão bibliográfica.

O capítulo 3 apresenta e caracteriza todo o modelo proposto, relaciona a teoria exposta no capítulo 2 com o que foi aplicado na elaboração do TMIGRAF´s. Nele serão discutidas toda a arquitetura e as tecnologias que foram aplicadas; motivações nas tomadas de decisões de algumas peças chaves da proposta, principais funções, vantagens e aplicabilidade.

No capítulo 4 o modelo TMIGRAF´s é submetido a uma base de dados a título de estudo de caso. É neste capítulo que se observa a implementação do modelo e sua aplicabilidade em bases de dados de trajetórias. Neste capítulo padrões extraídos dessa base através do TMIGRAF´s são apresentados e discutidos dentro do domínio da aplicação a fim de se entender o comportamento da mesma.

O Capítulo 5 retoma as discussões gerais do trabalho de forma conclusiva, finalizando a dissertação com os resultados e contribuições relevantes, dificuldades encontradas e indicações de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Esse capítulo está incumbido de apresentar o estado da arte da dissertação e demonstra os vários conceitos utilizados na pesquisa. Abordaremos os principais conceitos, histórico e importância dos objetos espaciais, especificamente os objetos espaços-temporais, o processo de KDD, a teoria dos grafos e os sistemas de apoio a decisão, mostrando sua relevância aplicada à mineração de trajetórias.

2.1 Dados Geográficos

Os dados geográficos são dados que descrevem uma localização de um objeto que existe fisicamente. Esses dados representam objetos ou fenômenos que ocorrem na Terra e que, necessariamente, estão associados a uma posição geográfica. Segundo Câmara et al. (2005), dado geográfico é definido como o fato, fenômeno ou objeto, natural ou artificial, que se apresenta inserido num contexto geográfico ou espacial. Ainda segundo Câmara et al. (2005), "tudo que acontece, acontece em algum lugar", logo deve ser mapeado, armazenado e analisado.

Três características principais descrevem os dados geográficos:

- Possuem atributos não espaciais: Esses atributos descrevem qualitativa e quantitativamente uma entidade geográfica. Esses atributos são tratados nos SGBD's convencionais;
- Possuem atributos espaciais: Descrevem a localização espacial e representação do objeto geográfico, considerando-se a geometria e um sistema de coordenadas geográficas. Esses atributos requerem um tipo de dado especial, apenas presentes nos SGBD's espaciais;
- Relações espaciais: Essas relações espaciais são implementadas por operadores espaciais. Esses operadores estão presentes apenas em SGBD's espaciais.

Uma classificação de dados geográficos foi proposta por Güting (1994) e é aceita hoje pela comunidade científica, na qual os dados geográficos podem ser agrupados como objetos ou fenômenos.

Os objetos apresentam uma subdivisão. Eles podem ser:

- 0-Dimensionais: Representados por pontos (Figura 2.1)

Figura 2.1: Ponto. Fonte: Autoria própria

- Uni-Dimensionais: Representados por linhas (Figura 2.2)

Figura 2.2: Linha. Fonte: Autoria própria

- Bi-Dimensionais: Representados por polígonos (Figura 2.3)

Figura 2.3: Polígono. Fonte: Autoria própria

- Tri-Dimensionais: Representados por superfícies (Figura 2.4)

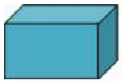


Figura 2.4: Superfície. Fonte: Autoria própria

Já os fenômenos tratam-se de dados contínuos, e, por sua vez podem apresentar diversas representações, por exemplo, as exibidas na figura 2.5.

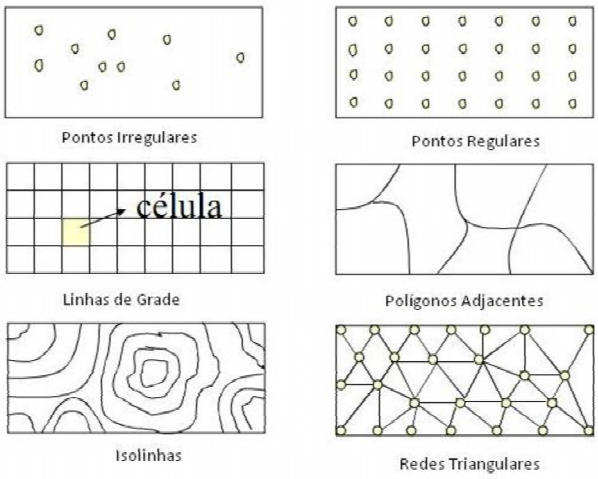


Figura 2.5: Dados Contínuos. Fonte: (BOGORNY, 2010)

2.2 Relacionamentos Espaciais

Devido às suas características peculiares, os dados espaciais apresentam relacionamentos específicos. Segundo EGENHOFER Max J. (1991), os relacionamentos espaciais dividem-se em três grandes categorias:

a) Relações Topológicas => Se referem ao posicionamento dos objetos em relação a outros objetos. Eles servem como referências mútuas e se mantêm invariante ante as transformações como escala e rotação. Nesse tipo de relacionamento podemos destacar alguns operadores (Figura 2.6):

- Disjoint: onde as bordas e interiores não se interceptam;
- Touch: onde as bordas se interceptam, mais o interiores não;
- Overlap: representa a interseção de bordas com interior;
- Equals: significa que dois objetos tem a mesma borda e o mesmo interior;
- Contains: acontece quando o interior e bordas de um objeto estão contidos em outro;
- Inside: é o oposto do Contains. A inside B, implica em B Contains A.

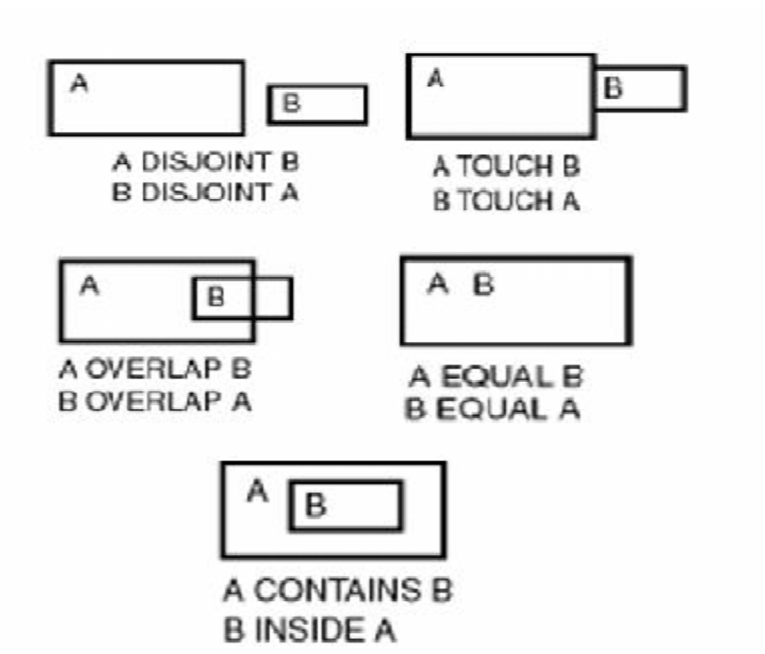


Figura 2.6: Relações Topológicas. Fonte: Autoria própria

b) Relações Métricas => São consideradas em termos de direções e distâncias, sendo que as relações direcionais são aquelas que descrevem a orientação no espaço, por exemplo, "norte" e "sul", e as relações

de distâncias são aquelas que dependem de definições métricas no sentido de se parametrizar o quanto é perto ou longe, por exemplo, "perto de" e "longe de". Tal parametrização dependerá das circunstâncias e das entidades geográficas relacionadas;

c) Relações Geo-Posicionadas (Ordem) => São aquelas que expressam a ordem, total ou parcial, dos objetos espaciais, sendo descritas por preposições como "em frente a", "atrás de", "acima de" e "abaixo de".

2.3 Objetos Espaço-Temporais

De acordo com Yi (2004), objetos espaço-temporais são objetos que possuem pelo menos um atributo espacial e um temporal. Esses objetos são utilizados para modelar muitos elementos do meio físico que apresentam características eminentemente dinâmicas, como seres humanos, animais, veículos, furacões, etc.

Segundo Guo, Liu e Jin (2010), no passado era difícil coletar dados sobre movimentos, porém hoje em dia, essa tarefa se torna facilitada com o advento de dispositivos baseados em localização, tais como, GPS, triangulação de rádio e de celulares, redes wireless, RFID e telemetria.

O aumento do uso dessas tecnologias está levando a geração de um volume muito grande de dados, que por sua vez nos oferecem uma gama de possibilidade de análise de comportamentos individuais e coletivos, dependendo do escopo dos dados, na busca de informações valiosas. Nesse contexto, os dispositivos móveis geram um novo tipo de dados chamado de "trajetórias de objetos em movimento" (BOGORNY, 2010). Esses trajetórias nos possibilitam um novo nicho de possibilidade de análise:

- Como as pessoas se movimentam pelas cidades?
- Existem movimentos com comportamentos típicos? Em uma determinada área? Em um determinado espaço de tempo?
- Existem relações de movimentos entre duas áreas distintas?
- Como se comporta uma determinada espécie de animais? Qual seu padrão de migração?
- Qual o deslocamento de determinado fenômeno natural? Em quanto tempo?

2.4 Representação das trajetórias

Uma trajetória nada mais é que o registro do movimento (YI, 2004). Trata-se do percurso realizado por um objeto, dentro de um sistema de coordenadas espaciais e variando seu tempo no decorrer desse percurso.

Segundo Santos e Alvares (2011), uma trajetória é uma sequência de pontos (id, x,y,t), em que x e y indicam a posição geográfica do objeto móvel identificado por id no tempo t. A figura 2.7 apresenta a representação de uma trajetória.

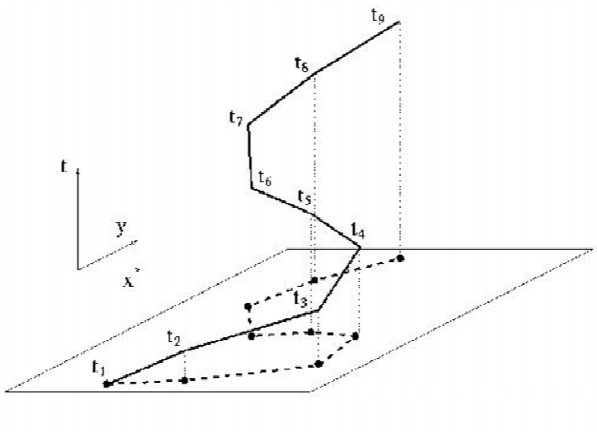


Figura 2.7: Representação espacial de uma trajetória. Fonte: (GIANNOTTI; TRASARTI, 2000).

As trajetórias são representadas por sequências finitas de tempo que fazem referência a locais específicos. Existem várias maneiras dessas sequências serem observadas:

- Baseadas no Tempo - As posições dos objetos são gravadas, levando-se em consideração um espaço de tempo pré definido pela aplicação;
- Baseada na mudança de posição - A observação é feita no momento que a posição do objeto diferir do posicionamento anterior;
- Baseada na localização - Nesse tipo de observação, locais são pré-definidos e a observação é realizada sempre nesses locais, definindo assim o tempo de deslocamento;
- Baseada em eventos - Tanto a posição quanto o tempo do deslocamento do objeto são registrados quando determinado evento for observado.

Geralmente, o processo de observação e registro dos movimentos das trajetórias são acompanhados de certo grau de incerteza, para que possíveis ruídos e imprecisões oriundas de limitações físicas de hardware ou humanas não influenciem na correta aquisição desses dados.

2.5 Descoberta de Conhecimento em Bancos de Dados

Com a globalização e a alta competitividade, a informação tornou-se um dos bens mais valiosos para uma organização. Ter acesso à informação precisa, de maneira rápida e eficiente, é um dos grandes diferenciais que podem levar ao sucesso. Neste aspecto, a Tecnologia da Informação tem evoluído, proporcionando aos tomadores de decisão uma infraestrutura e ferramentas que aliadas às metodologias de coleta, organização, processamento e utilização da informação, tornam-se vitais para a sobrevivência das organizações (COSTA et al., 2009).

Knowledge Discovery in Databases (KDD) é o processo, não trivial, de extração de informações implícitas, previamente desconhecidas e potencialmente úteis, a partir dos dados armazenados em um banco de dados (FAYYAD; PIATETSKY-SHAPIRO; SHYTH, 1996). O Termo "Não trivial" indica a existência de algumas técnicas de busca ou inferência. Por "Previamente desconhecidos" entenda-se que a informação deve ser nova para o sistema e de preferência também para o usuário. E, por último, "potencialmente úteis", deixa claro que esta informação deverá possibilitar ao usuário algum ganho.

O processo de KDD contém uma série de passos, a saber: seleção, pré-processamento e limpeza, transformação, mineração de dados (*data mining*) e interpretação/avaliação. Como se pode ver, o processo compreende, na verdade, todo o ciclo que o dado percorre até tornar-se conhecimento. O processo é interativo, pois o usuário pode intervir e controlar o curso das atividades. Também é iterativo, por ser uma sequência finita de operações onde o resultado de cada uma é dependente dos resultados das que a precedem (PRASS, 2004). Tal processo pode ser visto na figura 2.8:

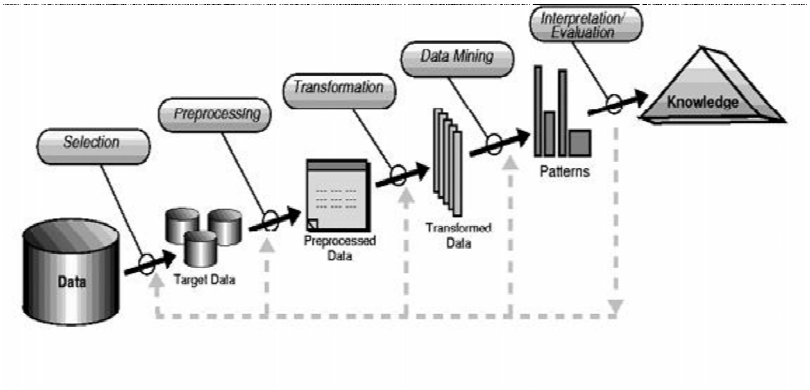


Figura 2.8: Etapas do Processo do KDD. Fonte: (FAYYAD; PIATETSKY-SHAPIRO; SHYTH, 1996).

Dentre as etapas, a mineração de dados se destaca, pois é a fase responsável pela transformação de dados em informações. A mineração de dados está relacionada com a descoberta de novos fatos, regularidades, restrições, padrões e relacionamentos e não apenas consultas complexas e elaboradas com a finalidade de confirmar uma hipótese em função dos relacionamentos existentes.

2.5.1 Métodos de Mineração de Dados

O processo de mineração de dados engloba várias áreas do conhecimento. Segundo Han e Kamber (2001), trata-se de uma área interdisciplinar que inclui uma série de disciplinas, entre elas sistemas de bancos de dados, estatística, aprendizagem de máquina, visualização e sistemas de informação, como podemos ver na figura 2.9.

Para um melhor entendimento do processo de mineração de dados, será apresentada uma classificação dos métodos mais comuns utilizados por essa etapa do KDD e as técnicas que englobam os mesmos. Esses métodos atualmente são aplicados em várias áreas do conhecimento, facilitando o entendimento do

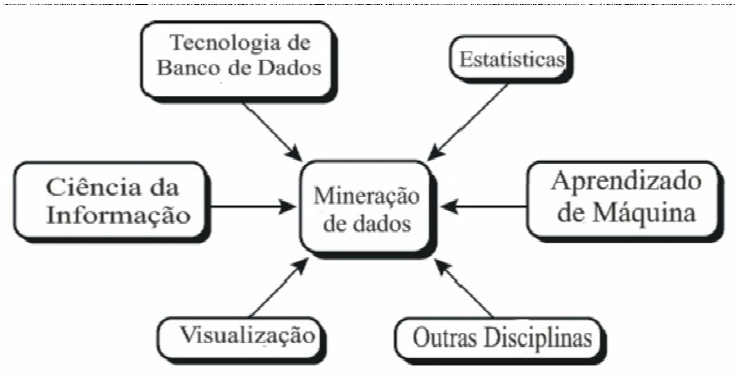


Figura 2.9: Áreas em que a Mineração de dados se apropria dos recursos. Fonte: (HAN; KAMBER, 2001).

comportamento de grandes volumes de dados e abrindo assim um leque de novas opções dentro dos seus respectivos escopos.

2.5.1.1 Aprendizagem Supervisionada

De acordo com Silva (2004), esta categoria de algoritmos possui esta denominação porque a aprendizagem do modelo é supervisionada, ou seja, é fornecida uma classe à qual cada amostra no treinamento pertence. O desafio da aprendizagem supervisionada é capacitar o sistema a atuar de acordo com o padrão observado nos exemplos de entradas e saídas.

Matematicamente, isso significa que o algoritmo deve ser capaz de construir uma fórmula capaz de produzir resultados semelhantes aos valores de entradas e saídas observados nos exemplos de treinamento.

Entre as classes de algoritmos de Aprendizagem Supervisionada destacam-se:

- Algoritmos de classificação e predição: Segundo Amo (2004), classificação é o processo de buscar modelos (funções) que descrevam e distingam classes ou conceitos, com o propósito de utilizar os modelos para predizer ou explicar o contexto. Geralmente o modelo baseia-se em dados de amostragem ou de treinamento. No caso da predição, o objetivo é inferir valor no conjunto de dados. Normalmente os classificadores exibem como resultados do seu processo, regras e árvores.

- Seleção de Atributos: A seleção de atributos envolve conjuntos de algoritmos que determinam os atributos mais e menos relevantes a serem considerados em um processo de mineração de dados. Segundo Oliveira, Dutra e Rennó (2005), a seleção de atributos é um problema de otimização e busca pelo menor subconjunto com a melhor acurácia no processo de classificação.

- Indução de Regras: Como citado em Oliveira e Lima (2011), nesta técnica vários algoritmos e índices são colocados para executar esse processo, onde a maioria desses processos são feitos pela máquina e só uma parte insignificante é feita pelo usuário. Na indução de regras, essas regras são apresentadas aos usuários como uma lista chamada de "não encomendada".

2.5.1.2 Aprendizagem Não-Supervisionada

Segundo Alvarez e Luque (2003), são considerados algoritmos de aprendizagem não supervisionada, quando não existe um agente externo indicando a resposta desejada para os padrões de entrada. Este tipo de aprendizado também é conhecido como aprendizado auto-supervisionado ou de auto-organização por que não requer saída desejada e/ou não precisa usar supervisores para seu treinamento.

Silva (2004) coloca que, além de serem auto-supervisionados, também são descritivos, pois descrevem de forma concisa os dados disponíveis, fornecendo características das propriedades gerais dos dados minerados. Entre as classes de algoritmos destacam-se:

- Associação: Segundo Zhao e Bhowmick (2003) é uma das mais importantes e pesquisadas técnicas de mineração de dados. Foi introduzida pela primeira vez por Agrawal, Imielinski e Swami (1993) em 1993. As regras de associação consistem em padrões do tipo X => Y, onde X e Y são conjuntos de valores antecedentes e consequentes. Seu objetivo principal é descobrir combinações de itens ou valores de atributos que ocorrem com frequência significativa em uma base de dados.

- Clustering: Nessa técnica os dados são agrupados conforme sua classificação e grupo a que pertencem. Esses grupos são construídos com base na semelhança que há entre os elementos dos próprios grupos. Conforme descrito por Oliveira e Lima (2011), a formação do cluster se baseia em alguma medida de similaridade, logo, os padrões pertencentes a um dado cluster devem ser mais semelhantes entre si, do que em relação a outros clusters.

2.6 Teoria dos Grafos

Segundo a teoria dos conjuntos, qualquer conjunto V, possui um conjunto V^2 de todos os pares não-ordenados de elementos de V. Se V tem n elementos, então V^2 tem $\frac{n(n-1)}{2}$ elementos. Os elementos de V^2 serão identificados como subconjunto de V que possui cardinalidade 2.

Dentro desse contexto, segundo Feofiloff, Kohayakawa e Wakabayashi (2007), um grafo é um par (V,A), em que V é um conjunto arbitrário e A é um subconjunto de V^2. Os elementos de V são chamados de vértices e os elementos de A são chamados de arestas.

Geralmente os grafos são nomeados (Rotulados). Por definição, dado o Grafo G, teremos: G=(V(G), E(G),ψ)

Onde:

V(G) => Corresponde ao conjunto não vazio de vértices;

E(G) => Corresponde ao conjundo disjunto de V(G), chamado de Aresta;

ψ => Função que associa cada aresta de G um par de vértices de G

A Figura 2.9 exibe a representação visual de um grafo.

Ainda dentro das definições iniciais de grafos, um dígrafo, ou grafo dirigido ou direcioado, (*digraph = directed graph*) é um conceito que consiste em dois conjuntos: um conjunto de objetos conhecidos como

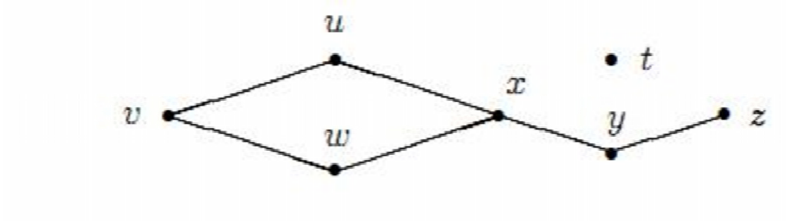


Figura 2.10: Esta figura é um desenho do grafo cujos vértices são t , u, v , w, x, y , z e cujas arestas são vw, uv , xw, xu, yz e xy . Fonte: (FEOFILOFF; KOHAYAKAWA; WAKABAYASHI, 2007).

vértices e outro conjunto de objetos conhecidos como arcos (BANG-JENSEN; GUTIN, 2009). Cada arco é um par ordenado de vértices. O primeiro vértice do par é a ponta inicial do arco e o segundo é a ponta final. A Figura 2.11 exibe a representação visual de um grafo direcionado.

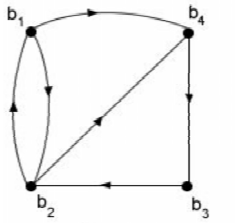


Figura 2.11: Digrafo. Fonte: (CUBAS, 2008).

2.6.1 Representação dos Grafos

Existem várias maneiras de representação de grafos. Apesar de normalmente representarmos de maneira gráfica, convém lembrar que os computadores não entendem desenhos.

Segundo Boaventura e Jurkiewicz (2009), a mais intuitiva representação consiste em dizer que: para cada vértice quais outros vértices estão ligados a ele, ou seja, que os são adjacentes. O conceito de adjacência se torna mais importante quando tomamos os rótulos desses vértices e associamos a uma matriz quadrada, formando assim uma matriz de adjacência.

Nas figuras 2.12 e 2.13 apresentamos como se representam as matrizes de adjacência:

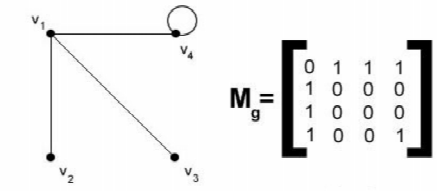


Figura 2.12: Matriz de Adjacência do Grafo. Fonte: (CUBAS, 2008).

Para determinar cada elemento basta adicionar 1 (um), quando houver conexão entre i e j (nesta ordem), ou 0 (zero), quando não houver, à uma matriz nXn (n = número de vértices do grafo/digrafo).

OBS: Para i=j ser 1 deve haver um loop no vértice senão é sempre igual a 0.

No caso da matriz de um digrafo a representação seria:

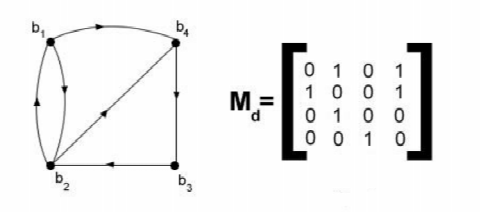


Figura 2.13: Matriz de Adjacência do Digrafo. Fonte: (CUBAS, 2008).

2.6.2 Subgrafos

De acordo com Boaventura e Jurkiewicz (2009), o conceito de um subgrafo se refere a um subconjunto de um grafo G, de tal sorte que o conjunto dos vértices de H V(H) está contido em V(G) e o conjunto das arestas de H E(H) está contidas em E(G).

O conceito de subgrafo envolve basicamente os vértices e as arestas que o compõem. Desta forma há mais de um tipo de subgrafo.

Um subgrafo H é dito gerador de G quando H é um subgrafo de G e V(H) = V(G). A restrição para um subgrafo ser gerador de algum grafo é que ele deve possuir os mesmos vértices do grafo original. A Figura 2.14 mostra um exemplo de Subgrafo gerador.

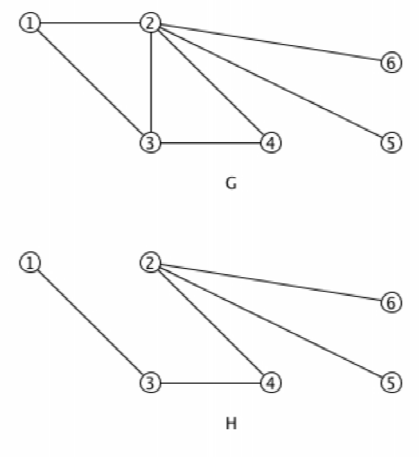


Figura 2.14: SubGrafo Gerador. Fonte: (CUBAS, 2008).

Note que o grafo H possui todos os vértices de G, entretanto não possui as mesmas arestas.

Um subgrafo induzido H é um subgrafo de um grafo G, no qual a partir de um conjunto de vértices ou arestas pertencentes ao grafo original G, é possível se obter sua indução, denotada por G[K], onde K é o conjunto de arestas ou vértices que servirá para a indução.

Os subgrafos induzidos são obtidos a partir de G, pela remoção de vértices no conjunto V(G) e suas arestas incidentes, no caso da indução por vértices, ou analogamente pelo subconjunto de arestas e de vértices extremos dessas arestas (Figura 2.15).

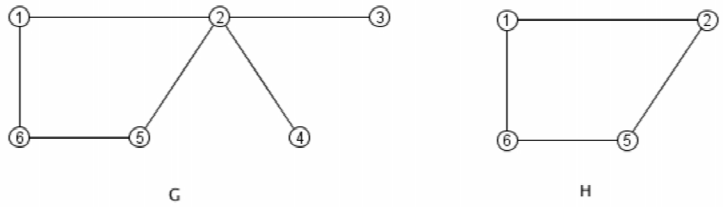


Figura 2.15: SubGrafo Induzido. Fonte: (CUBAS, 2008).

2.6.3 Isomorfismo de Grafos

Segundo Feofiloff, Kohayakawa e Wakabayashi (2007), isomorfismo entre dois grafos G e H é uma bijeção f de V (G) em V (H) tal que dois vértices v e w são adjacentes em G se e somente se f(v) e f(w) são adjacentes em H.

Notadamente o conceito de isomorfismo deriva do conceito de equivalência estrutural. Mesmo os grafos não possuindo os mesmos rótulos, se for possível que se faça um mapeamento desses vértices e arestas e a função bijetora for respeitada, estaremos diante de um clássico caso de isomorfismo de grafos.

No figura 2.16, retirada de Carvalho (2009), nota-se que os dois grafos, aparentemente não são isomorfos. Se os grafos não possuem os mesmos rótulos é possível estabelecer um relacionamento.

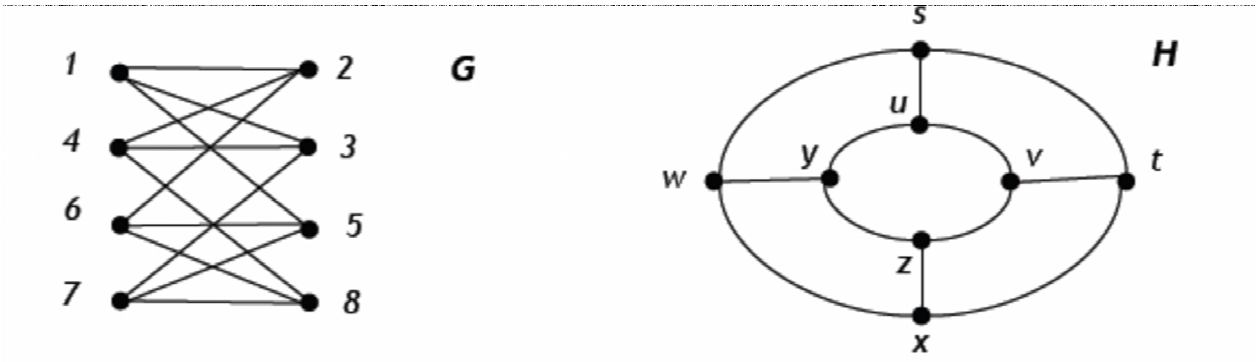


Figura 2.16: Isomorfismo de Grafos. Fonte: (CARVALHO, 2009).

Através da função bijetora f:

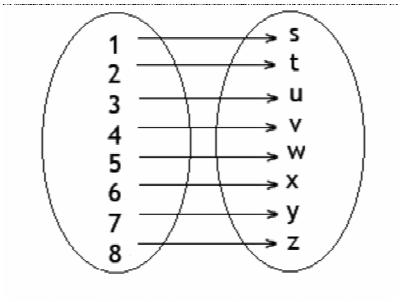


Figura 2.17: Função Bijetora. Fonte: Autoria própria.

Se analisarmos a função de bijeção na figura 2.17, nota-se que, no grafo G, o nó 1 é vizinho dos nós 2, 3 e 5 e não a outros. Da mesma forma, no grafo H, observamos que s=f(1) é vizinho a t=f(2), u=f(3) e w=f(5) e não mais a outros. Dessas relações observadas podemos concluir que função de bijeção entre a vizinhança dos grafos G e H é verdadeira, logo, os grafos são isomorfos.

Segundo Read e Corneil (1977), retirado de Carvalho (2009), não existem algoritmos polinomiais para identificar se dois grafos são isomórficos. Uma das formas de verificar se ocorre o isomorfismo entre dois grafos é provar exatamente o contrário através do conceito de invariante[1].

2.7 Trabalhos Relacionados

Centros de pesquisas do mundo todo vêem demandando esforços no sentido de aprofundar as pesquisas acerca dos objetos espaço-temporais. Tais esforços tendem a buscar novas técnicas e metodologias que viabilizem o armazenamento, extração e análise automática de conhecimento desses objetos. Apesar de ser um tema relativamente novo, já existem alguns trabalhos que norteiam as pesquisas na área.

O artigo "*Mobility, Data Mining and Privacy: The GeoPKDD Paradigm*" foi um dos trabalhos significativos da revisão bibliográfica. Os autores Giannotti e Trasarti (2000) apresentaram de forma clara e concisa a metodologia utilizada, fato que contribuiu bastante para o entendimento e clareza de detalhes para a realização deste trabalho.

O artigo propõe uma arquitetura de descoberta de conhecimento útil sobre o comportamento do movimento humano a partir de dados de mobilidade, auxiliando na melhoria da tomada de decisões gerenciais.

Um outro aspecto interessante é que em sua arquitetura intitulada GeoPKDD, eles procuram preservar a privacidade das pessoas sob observação.

O Método GeoPKDD apresentado na figura 2.18, define duas tarefas básicas na aplicação dos seus processos de análise de dados de trajetórias: no primeiro ele define o formato dos padrões espaço-temporais e modelos a serem extraídos de dados de trajetórias e o segundo implementa algoritmos eficientes para extrair tais modelos. Nesse sentido, o framework se baseia em duas premissas básicas: as ROI´s, que são regiões de interesse pelo espaço de dados e o tempo de viagem típico de objetos em

[1] Propriedade que é preservada pelo isomorfismo, como, o número de nós e o grau.

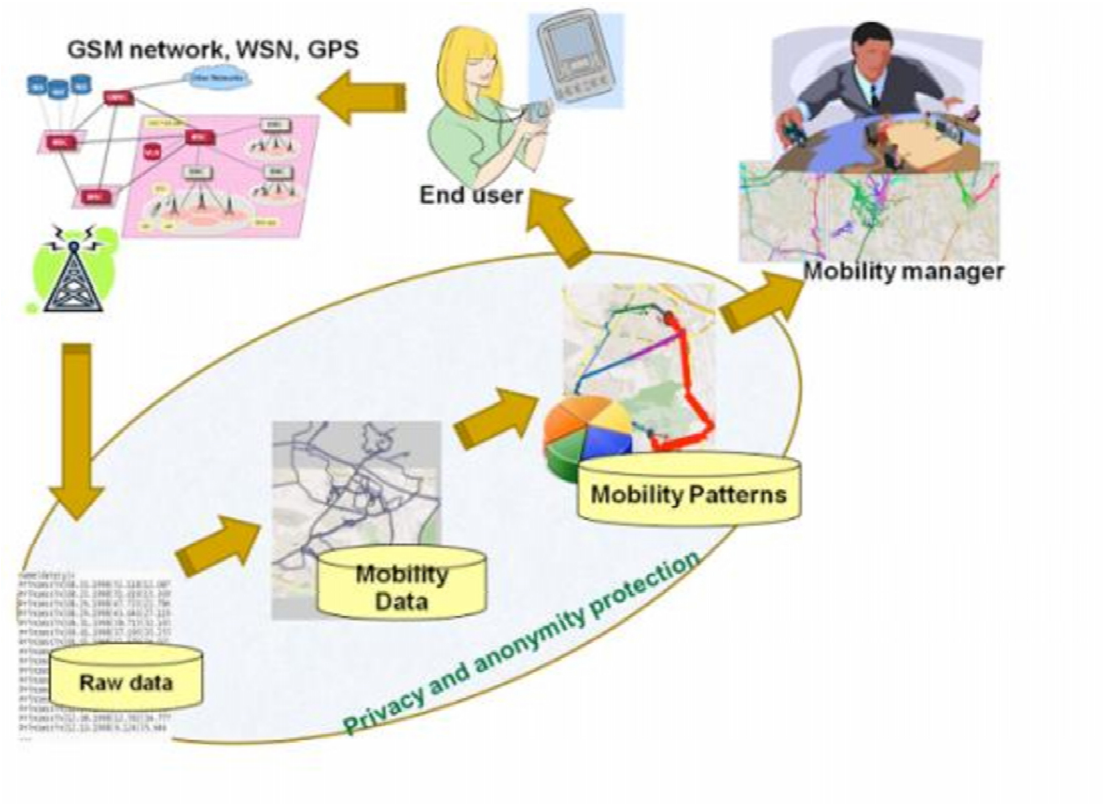


Figura 2.18: Geographic Privacy-aware Knowl-edge Discovery and Delivery. Fonte: (GIANNOTTI; TRASARTI, 2000).

movimento de região para região analisada. Dessa forma temos uma definição do padrão de trajetórias como sendo o par ordenado (S,A), onde S é a sequência de regiões analisadas e o A o tempo para essas transições.

Os resultados apresentam várias análises visuais, indicando por exemplo as células mais densas, as regiões de maior interesse e a aplicação de algoritmos de clusterização para descoberta de sequencias frequentes juntamente com seus tempos de transição. Além desses resultados, ao final do processo de geração e análise de padrões é aplicado um algoritmo de generalização de trajetórias para que o anonimato seja preservado sem que a utilidade analítica seja perdida.

O Artigo "TRACTS: Um método para a classificação de trajetórias de objetos móveis usando séries temporais" (SANTOS; ALVARES, 2011), apresentou uma nova metodologia de classificação de trajetórias utilizando aproximação por séries temporais.

Segundo Santos e Alvares (2011), o método consiste em transformar cada trajetória em um conjuto de séries temporais, para que destas sejam extraídas as características que serão submetidas a um processo de geração do modelo de classificação. Ainda, segundo o autor, os trabalhos similares apresentam soluções específicas para cada estudo de caso, enquanto o método TRACTS propõe um método mais geral, aplicável a vários domínios de conjunto de dados.

Como resultado, Santos e Alvares (2011) conclui que o método, apesar de apresentar uma grande independência do domínio de dados e resultados superiores, se comparados a outros métodos elaborados para o mesmo fim, apresenta dificuldades de realizar classificação com dados caóticos.

Além do escopo de mineração de objetos espaço-temporais, outra parte envolvida no processo de mineração deve ser levada em consideração que é a parte de mineração de grafos. A dissertação de mestrado "Enriquecimento Semântico de Padrões minerados em Grafos Utilizando Ontologias" de França (2011) propõe a aplicação de técnicas e algoritmos na descoberta de padrões frequentes dentro dos grafos. Além da descoberta de padrões, um outro aspecto forte do trabalho é o enriquecimento semântico desses padrões através de uma ontologia de contexto.

Como resultado, França (2011) apresenta uma aplicação prática com utilização de bases de dados de proteínas, mostrando assim um avanço na interpretação dos padrões minerados, onde informações extras acerca desses padrões são exibidas de forma mais amigável ao usuário.

Durante o levantamento da pesquisa bibliográfica foram encontrados na literatura outros trabalhos significativos na área de extração de conhecimento em objeto espaço-temporais e mineração de grafos, alem dos já discutidos, dentre eles (CHAKRABARTI, 2005), que aborda técnicas de mineração de grafos para propor um modelo de detecção de *outlair*[2] em redes P2P, (HäRRI; BONNET; FILALI, 2007), que discute os conceitos que envolvem os grafos cinéticos em redes moveis(MANET´s), (ONG et al., 2010), que trata de um framework para descoberta de conhecimento em base de dados de trajetórias e (ANDRIENKO et al., 2009), que aborda técnicas e algoritmos de clusterização, adaptados a base de dados de trajetórias.

[2]Significa fora dos padrões "normais".

Capítulo 3

Projeto e Implementação do TMIGRAF's

Esse capítulo descreve a metodologia proposta na pesquisa, proveniente da implementação de uma estrutura de mineração de objetos espaço-temporais baseada em regras de associação e sequências frequentes com visualização em grafos. Além das discussões do modelo proposto, o capítulo prossegue com análise minuciosa de todas as etapas da implementação e tecnologias aplicadas.

3.1 Introdução

A maioria dos trabalhos desenvolvidos no âmbito da análise de objetos espaço-temporais foca em apenas uma das dimensões do problema, ou seja, o espaço ou o tempo. Esforços na busca da compreensão do comportamento desses objetos vem sendo desenvolvidos, porém muitas vezes, as propostas de solução passam por adaptações de trabalhos já consolidados no estado da arte não se adequando totalmente à nova realidade, deixando assim algumas lacunas a serem exploradas.

Algumas técnicas de mineração de dados tradicionais já foram aplicadas às trajetórias e apresentaram soluções interessantes. Foi utilizado o conceito tradicional de classificação (TAN; STEINBACH; KUMAR, 2006), onde é aprendida uma função alvo F, que mapeia um conjunto de atributos X para um dos rótulos de classes predefinidos Y (Aplicado às trajetórias, a função F mapeia cada trajetória a um rótulo de classe Y predefinido). Esse rótulo de classe Y determina o objetivo do processo de classificação que pode identificar desde diferentes tipos de objetos (LEE et al., 2008) até os meios de transportes utilizados pelos mesmos (ZHENG et al., 2008).

Outras técnicas utilizadas podem ser as de clusterização, com os métodos de análise de similaridade (PANAGIOTAKIS; PELEKIS; KOPANAKIS, 2009) ou de análise de regiões no espaço (LEE et al., 2008).

O modelo TMIGRAF´s busca uma técnica de análise de trajetórias baseada no conceito bem fundamentado de regras de associação juntamente com o poder de visualização dos grafos. A idéia é utilizar as técnicas de análise e mineração de dados frequentes aplicado às trajetórias, mapeando os pontos importantes nas trajetórias, como vértices de um grafo e os pesos das arestas dos mesmos representando o

tempo de deslocamento dos objetos entre esses pontos, de forma que todas as características dos objetos espaço-temporais sejam levados em consideração no processo de descoberta de conhecimento.

3.2 Modelo TMIGRAF's

Como dito acima, os trabalhos relacionados à análise e mineração de objetos espaço-temporais que envolvem trajetórias não contemplam a utilização de técnicas de regras associação aplicadas juntamente com a teoria dos grafos. A arquitetura mostrada na figura 3.1 demonstra os passos necessários para a aplicação desta metodologia.

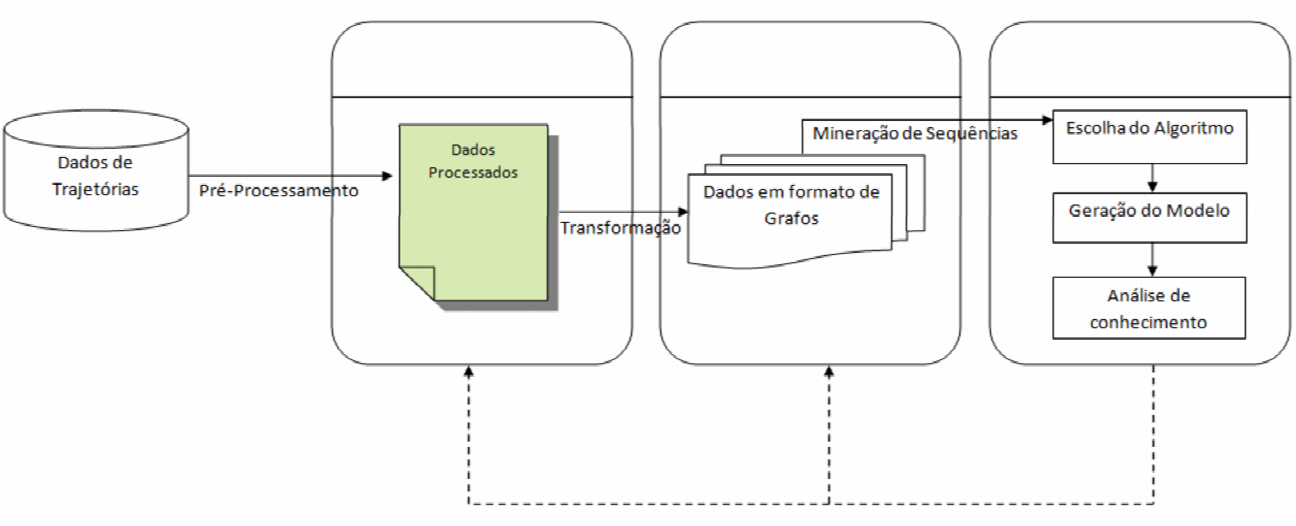


Figura 3.1: Modelo TMIGRAF's. Fonte: Autoria própria

O primeiro desafio do modelo é encontrar bases de trajetórias padronizadas. Segundo (BOGORNY, 2010), os dados de objetos espaço-temporais são representados por conjuntos de pontos, localizados no espaço e no tempo, agrupados pelas trajetórias que eles representam, no formato T=(X1,Y1,T1), (X2,Y2,T2), ..., (Xn,Yn,Tn), como apresentado na figura 3.2. De posse desses dados, o passo de pré-processamento os conduz ao primeiro estágio do modelo que é a retirada de ruído dos dados. Após essa etapa um processo crítico é o de transformação, que é a fase onde os dados de trajetórias serão modelados em formato de grafos. Já em formato de grafos, esses dados são submetidos à etapa mais importante do processo que é a de mineração de sequências. Para isso, eles são submetidos a mineradores específicos, para que, de posse dos padrões resultantes desse processo, possa se interpretar o conhecimento aferido sobre as trajetórias que deram início ao processo.

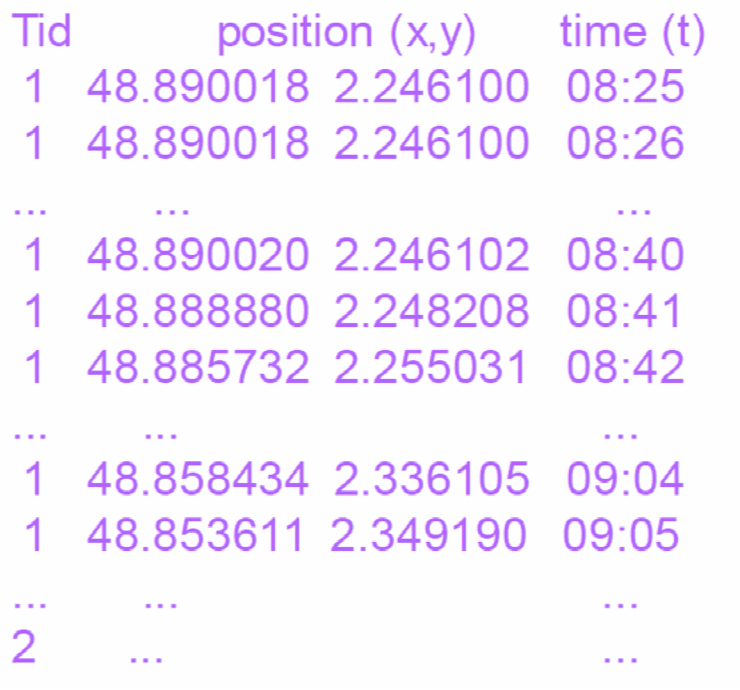


Figura 3.2: Representação do conjunto de pontos das trajetórias. Fonte: (BOGORNY, 2010)

3.2.1 Pré-Processamento

O Pré-Processamento de dados do TMIGRAF´s segue os moldes do Pré-Processamento do KDD convencional. Segundo Han e Kamber (2001) citado por Silva (2004), rotinas de limpeza de dados tentam suprir valores ausentes, reduzir discrepâncias de valores ruidosos e corrigir inconsistências. Para valores ausentes, algumas técnicas aplicáveis são:

1. Ignorar a tupla
2. Suprir valores ausentes:
   (a) manualmente;
   (b) através de uma constante global;
   (c) utilizando a média do atributo;
   (d) utilizando a média do atributo para todas as instâncias da mesma classe;
   (e) com o valor mais provável (regressão, inferência, etc.).

As técnicas 2b, 2c, 2d e 2e podem "viciar" os dados. Dentre elas, a técnica 2e é uma estratégia interessante, pois em comparação com outros métodos utiliza um maior número de informações dos dados disponíveis.

Ruído pode ser definido como um exemplo (ou conjunto de exemplos) que aparentemente é inconsistente com o restante dos exemplos encontrados em um conjunto de dados, pois não segue o mesmo padrão dos demais (BARNETT; LEWIS, 1994). Vale ressalvar que exemplos considerados ruído podem, em alguns casos, representar dados corretos, que seguem padrões diferenciados dos demais exemplos presentes no conjunto de dados (JHON, 1995). A eliminação de ruídos pode ser realizada através de:

1. Interpolação => Função que, a partir de uma amostra de dados, consegue prever valores corretos;
2. Agrupamento => Valores de dados agrupados tendem a corrigir ruídos existentes;
3. Inspeção humana e computacional combinadas => Verificação manual humana ou através de algoritmos de verificação/validação;
4. Regressão => Modelo estatístico de inferência de dados.

Para o processo de limpeza de dados do TMIGRAF´s, adotou-se a classificação de problemas de Rahm e Do (2000) e o estudo de Oliveira, Rodrigues e Henriques (2004) que definiu o problema de ruídos e inconsistências de dados e propõe algumas soluções. Esse estudo foi adaptado às características das bases de dados espaciais. Segundo ele, os problemas podem ser de duas naturezas:

1. Em termos absolutos, ou seja, do valor do atributo na tabela. Essa classificação se subdivide em:
   (a) Atributos
   (b) Tuplas
   (c) Tabelas
2. Em termos relativos, que se referem aos relacionamentos nas tabelas.

Ainda dentro do agrupamento feito de problemas supracitados, dentro do grupo de problemas de atributos, segue uma subdivisão em termos de problemas ao nível do próprio atributo, ao nível de tuplas e problemas ao nível de tabelas.

Um exemplo da aplicação do Pré-Processamento é demonstrado na figura 3.3. Os dados[1] apresentados representam:

- TID - O identificador da trajetória;
- MID -O identificador de um movimento realizado pela trajetória;
- Stop1 - Ponto de saída do movimento;
- Stop2 - Ponto de chegada do movimento;

[1] Existem vários formatos válidos de dados de trajetórias

- IntTime - Data/Hora de inicio do deslocamento.

No registro do movimento 1 o campo Stop2 está vazio. Nesse caso nota-se um problema de falta de dados. No registro 3, o campo stop1 está com o valor 10. Observa-se que um valor numérico está presente em um campo alfanumérico, caracterizando-se como um problema de inconsistência de tipos de dados.

Um outro problema frequente em bases de objetos espaço-temporais, notadamente em bases de trajetórias é a ausência de sequência do processo. Como já vimos anteriormente, uma trajetória é uma sequência de pontos e o tempo de deslocamento entre os mesmos agrupados em uma trajetória. No exemplo do registro 9, um movimento tem o valor de chegada no atributo Stop2 igual a "*Hotel*", e no registro 10, o atributo Stop1 apresenta o valor "*beach*", demonstrando uma inconsistência semântica de dados.

	TID	MID	STOP1	STOP2	INTTIME
1	1	1	University		04/10/2011 - 06:42:13
2	1	2	Beach	Shopping Center	04/10/2011 - 06:42:25
3	1	3	10	Tourist Place	04/10/2011 - 06:42:33
4	1	4	Tourist Place	Conference Center	04/10/2011 - 06:42:39
5	1	5	Conference Center	football stadium	04/10/2011 - 06:42:41
6	1	6	football stadium	Beach	04/10/2011 - 06:42:55
7	1	7	Beach	Shopping Center	04/10/2011 - 06:43:07
8	1	8	Shopping Center	Conference Center	04/10/2011 - 06:43:15
9	1	9	Conference Center	Hotel	04/10/2011 - 06:43:17
10	1	10	Beach	Conference Center	04/10/2011 - 06:43:21
11	1	11	Conference Center	Shopping Center	04/10/2011 - 06:43:23
12	1	12	Shopping Center	University	04/10/2011 - 06:43:31
13	1	13	University	Tourist Place	04/10/2011 - 06:43:41
14	1	14	Tourist Place	Beach	04/10/2011 - 06:43:47

Figura 3.3: Representação dos dados de uma trajetória. Fonte: Autoria própria

Pelas próprias características das trajetórias, a opção de ignorar um registro poderia gerar valores totalmente não condizentes com a realidade. Uma opção plausível seria eliminar uma trajetória por completo, porém, a perda de dados deve ser considerada apenas em último caso. Felizmente, nos três casos supracitados podemos optar por suprir os valores dos dados com 100% de aproveitamento, pois os valores de sequência das trajetórias estão presentes na própria base.

Um outro tipo de erro encontrado é a falta de sequência dos movimentos da trajetória. Nesses casos uma análise manual pode ser necessária para resolução do problema, porém, o mais recomendável é o descarte da trajetória por completo. Se o erro for em um dado de peso da aresta, ou seja, no tempo do deslocamento, opta-se por preencher o campo com um valor médio do atributo. Essa técnica resolve o problema, porém como foi citado acima, dependendo do número de erros, os dados podem ficar viciados e comprometer os resultados do processo como um todo.

3.2.2 Processo de Transformação

O processo de transformação dos dados é uma parte crítica no modelo TMIGRAF´s. Esse processo consiste na aplicação de algoritmos de mapeamento de dados visando a mudança de formato das trajetórias em grafos direcionados e formatos de sequências. Os formatos de entrada e de saída dos dados devem

ser bem definidos para o algoritmo e o mapeamento dos atributos preciso para que a semântica não seja comprometida.

Na etapa inicial, os dados ainda estão em formato de trajetórias, e depois desse processo de transformação irão passar para o formato de grafos. Como vimos no capítulo anterior, existem várias formas de representação de grafos. O formalismo escolhido pelo TMIGRAF´s baseia-se em uma entrada de dados padrão exigida pela maioria dos mineradores que recebem um grafo como entrada. Ele facilita a construção e visualização dos padrões obtidos em formato de grafos.

O formato consiste de um conjunto de trajetórias transformadas em grafos. Na figura 3.4 a representação dos dados ainda está em formato de trajetória, e na figura 3.5, depois de submetido ao processo de transformação, já está em formato de grafo com a devida visualização.

Cada grafo inicia-se com a letra "t"seguida de um caractere de comentário "#"(esses comentários podem servir como espaço adicional para informações interessantes, como frequência de padrões, sequências de trajetórias, parâmetros utilizados para os mineradores, etc) . Nas linhas iniciadas com o caractere "v", aparece o identificador do vértice do grafo e o rótulo (todos os grafos que foram trabalhados no modelo são direcionados e rotulados). Na linha que se inicia pelo caractere "u", representamos as arestas por meio dos identificadores dos vértices que estão concatenados por ela, seguido do seu rótulo (o rótulo das arestas representam o peso de cada uma delas em formato de tempo discretizado. Essa informação é muito importante e será utilizado na busca dos padrões a serem minerados). O padrão pode ser visualizado na figura 3.5.

	TID	MID	STOP1	STOP2
▶ 1	38	1	Hotel	Shopping Center
2	38	2	Shopping Center	University

Figura 3.4: Exemplo dos dados submetidos ao processo de Transformação. Fonte: Autoria própria

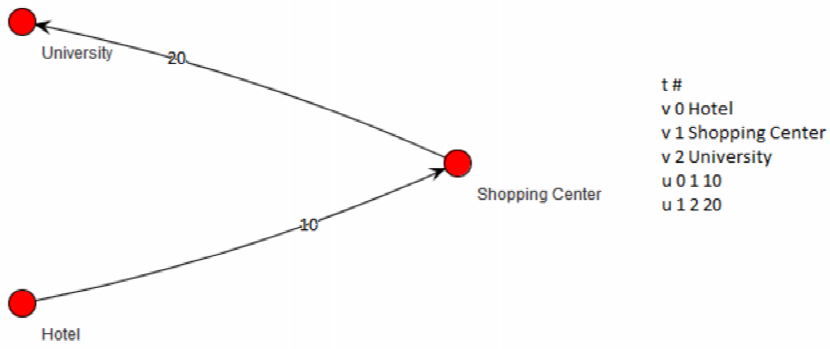


Figura 3.5: Representação dos dados de um grafo. Fonte: Autoria própria

Um algoritmo foi implementado para realizar o processo de transformação dos dados. Esse algoritmo

foi desenvolvido em PL/SQL[2]. A escolha da linguagem se deu pelo fato de ser uma linguagem estruturada e já estar integrada com o próprio SGBD. Dependendo do formato de entrada das trajetórias, algumas adaptações podem se fazer necessárias. A abordagem baseia-se em algoritmos de mapeamento de dados, onde os pontos de partida e chegada das trajetórias foram mapeados para vértices e o tempo de deslocamento foi mapeado para o peso a ser considerado nas arestas. Ao término do processo temos uma base de grafos, cada um deles representando uma trajetória específica.

3.2.3 Processo de Mineração de Sequências

A etapa final do modelo TMIGRAF´s consiste na aplicação de algoritmos de mineração de dados na busca de padrões semanticamente relevantes. Para esta tarefa, optou-se pelos algoritmos de aprendizagem não supervisionada (método descritivo) de associação e de descoberta de padrões sequenciais. Essas opções se deram tendo em vista que nas investigações científicas realizadas para o trabalho, não foram encontrados na literatura trabalhos que abordassem esses paradigmas para descoberta de conhecimento em bases de dados de trajetórias, além das características sequenciais das mesmas. Outro motivo que incentivou a escolha das técnicas de aprendizagem não supervisionadas foi a sua facilidade de percepção à análise e visualização dos resultados.

3.2.3.1 Descoberta de Associações

Segundo Gonçalves (2011), o objetivo desses algoritmos é descobrir combinações de itens ou valores de atributos que ocorrem com frequência significativa em uma base de dados. Esse problema foi introduzido em Agrawal e Srikant (1994).

Para que se possa entender como funciona o processo de mineração de regras de associação, é importante se ter em mente alguns conceitos básicos. O mais primitivo deles é o conceito de item. Um item é um valor atômico de um atributo da base de dados que se quer analisar. Esses itens são agrupados nas transações dos bancos de dados. Essas transações, que são registros na base de dados, são conjuntos de itens, que por sua vez são chamados de itemset´s. Um itemset com k elementos é chamado de k-itemset.

Segundo exemplo retirado de Amo (2003), imagine um supermercado onde os itens são as mercadorias armazenadas na base de dados como mostrado na figura 3.6:

[2]PL/SQL (acrônimo para a expressão inglesa *Procedural Language/Structured Query Language*) é uma extensão da linguagem padrão SQL para o SGBD Oracle da Oracle Corporation.

Artigo (item)	número que o representa
Pão	1
Leite	2
Açúcar	3
Papel Higiênico	4
Manteiga	5
Fralda	6
Cerveja	7
Refrigerante	8
Iogurte	9
Suco	10

Figura 3.6: Mapeamento dos Itens da Base de Produtos de um supermercado Fonte: (AMO, 2003).

Esses itens são mapeados para números para facilitar o processamento das regras de associação. De posse do mapeamento, cada compra realizada pelo cliente é armazenada na base transacional em forma de itemset´s, como visto na figura 3.7.

TID	Itens comprados
101	{1,3,5}
102	{2,1,3,7,5}
103	{4,9,2,1}
104	{5,2,1,3,9}
105	{1,8,6,4,3,5}
106	{9,2,8}

Figura 3.7: Base de Dados Transacional com os ItemSet´s. Fonte: (AMO, 2003)

Esses itemset´s podem ser de qualquer tamanho. Na base da figura 3.7, por exemplo, a transação 105 possui um itemset de k = 6, ou seja um 6-itemset.

De posse dessa base de dados podemos procurar as regras de associação dependendo dos parâmetros que desejamos aplicar. Pode-se considerar, por exemplo, que um itemset frequente é aquele que aparece mais de 40% das vezes na base transacional. Caso deseje-se uma análise mais criteriosa, para que seja considerado frequente, as ocorrências de itemset´s na base de dados devem passar dos 60%. Esse número de ocorrência na base de dados é chamado de suporte e é passado por parâmetro para o algoritmo de análise da base de dados. Na figura 3.8, temos alguns itemset´s que foram encontrados na base e seus respectivos suportes. A definição de quais serão considerados frequentes vai depender justamente do valor do suporte de entrada.

Itemset	Suporte
{1,3}	0,6666
{2,3}	0,3333
{1,2,7}	0,16666
{2,9}	0,5

Figura 3.8: Frequências de ocorrências dos ItemSet´s. Fonte: (AMO, 2003)

Os algoritmos que descobrem regras de associação devem gerá-las de forma a atender parâmetros que são informados pelo usuário: suporte e confiança (AGRAWAL; SRIKANT, 1994). Suporte é o percentual de incidência da regra no conjunto de transações e confiança indica a validade da regra (GONçALVES, 2008).

Formalmente temos:

Itens => $I = \{i_1, i_2, i_3, ..., i_n\}$

D => Banco de dados transacional composto por itemset´s

Itemset´s => Sub-conjunto não vazio de I

X = Número de transações que suportam (A U B);

Y = Número de transações da base;

Z = número de transações que suportam A;

Suporte =>

$$Sup(A => B) = \frac{X}{Y}$$

Confiança =>

$$Conf(A => B) = \frac{X}{Z}$$

Logo, dizemos que uma transação T, suporta um itemset I, se, somente se $I \subseteq T$.

De posse desses conceitos podemos partir para uma definição formal de regra de associação como sendo uma expressão da forma A => B, onde A e B são itemsets com um suporte predefinido (AMO, 2004).

Nem sempre o grau de confiança é um bom parâmetro para se analisar determinada regra. Para que ele seja relevante, geralmente vem associado ao grau de suporte. Segundo Amo (2003), uma regra de associação r é dita interessante se conf(r)> α e sup(r) > β , onde α e β são respectivamente um grau mínimo de confiança e um grau mínimo de suporte especificados pelo usuário.

Por exemplo, na base de dados da figura 3.7, temos a confiança do item {Fralda} => {Manteiga}, ou seja {6}=> {5} de 100%. Porém um suporte de apenas 16%. A análise apenas do grau de confiança isoladamente poderia nos levar a um escolha de um padrão frequente equivocado.

As regras de associações podem ser ainda (AMO, 2004):

- Multidimensionais => Em que atributos de diferentes tipos aparecem na regra.

  Ex. (Sexo = 'M') ^(20<Idade<30) => {cerveja}

  Esta regra indica que homens com idade entre 20 e 30 anos compram cerveja.

- Híbridas => Em que atributos de mesma dimensão aparecem várias vezes na regra:

  Ex. (Sexo = 'M') ∩ (20<Idade<30) ∩ {café} => {leite}

  No caso acima verifica-se que homens com idade entre 20 e 30 anos, que compram café, também adquirem leite.

- Negativas => Auferindo que o consumidor que compra os produtos A e B, não compra o produto C, por exemplo.

- Associações => podem obedecer uma hierarquia de generalizações. Por exemplo, o conceito "Meio de Transporte" representando avião, carro, navio, etc.

3.2.3.2 Algoritmo APRIORI

O algoritmo APRIORI é considerado um clássico na extração de Regras de Associação. Ele foi proposto pela equipe de pesquisa "*The Quest Data Mining System*" da IBM "*Almaden Research Center*", em 1994. Foi com as pesquisas acerca do APRIORI que se deu origem ao Software *Intelligent Miner*.

Este algoritmo realiza buscas em profundidade recursivas em um banco de dados à procura dos conjuntos frequentes (conjuntos que satisfazem um suporte mínimo estabelecido). A técnica é a busca por itens de tamanho K para geração de conjuntos de tamanho K + 1. Segundo Nong (2003), o primeiro passo do algoritmo é encontrar os conjuntos de itens frequentes com k = 1 item. Este conjunto é denominado de L1. O conjunto de L1 é usado para gerar o de L2, que representa os conjuntos de itens frequentes com k = 2 itens, e assim sucessivamente, até que nenhum conjunto de itens frequentes possa ser gerados. A figura 3.9 apresenta um pseudo código ilustrando os passos do APRIORI.

```
L1 = {large 1-itemsets};
for (k=2; Lk-1 ≠ ∅; k++) do begin
   Ck = apriori_gen(Lk-1);  // Novos candidatos
   forall transactions t ∈ D do begin
      Ct = subset(Ck, t);  // Candidatos contidos em t
      forall candidates c ∈ Ct do
         c.count++;
   end
   Lk = {c ∈ Ck | c.count ≥ minsup};
end
Answer = ∪k Lk;
```

Figura 3.9: Pseudo Código do APRIORI. Fonte: (ARBEX; SABOREDO; MIRANDA, 2004)

O APRIORI é composto de três fases principais, são elas:

- Geração dos conjuntos candidatos;

- Poda dos conjuntos e candidatos;

- Contagem do suporte.

As duas primeiras ocorrem na memória principal, enquanto a terceira na memória secundária, fazendo-se necessário uma consulta constante ao banco de dados. Visando a otimização do processo, as duas primeiras fases obedecem uma propriedade fundamental do algoritmo que é a antimonotonia da relação de inclusão de itemset´s.

A propriedade da antimonotonia usa como principio que cada subconjunto de um conjunto de itens frequentes também deve ser frequente. Essa regra é utilizada para diminuir o número de comparações entre candidatos nas transações. Seguindo a regra da inversão, todos os candidatos gerados que possuírem algum subconjunto que não é frequente serão automaticamente excluídos do processo, ou seja, podados.

Fase de Geração de candidatos:

Como o próprio nome diz, essa fase é a de geração de itemset´s candidatos a serem frequentes. Esse processo segue a propriedade da antimonotonia e faz parte da geração de candidatos de tamanho k a partir de cada item frequente L (k-1). O conjunto de candidatos a itemset´s e montado concatenando-se pares de itemset´s de tamanho k-1 que tenham elementos de k-2 em comum. Esse processo garante a obtenção de um itemset com, pelo menos dois subitemset´s de tamanho k-1 frequentes. O script mostrado na figura 3.10 apresenta a função APRIORI_GEN, responsável pela construção dos candidatos.

$$\textbf{insert into } C'_k$$
$$\textbf{select } p.\text{item}_1,\ p.\text{item}_2,\ \ldots,\ p.\text{item}_{k-2}, p.\text{item}_{k-1},\ q.\text{item}_{k-1}$$
$$\textbf{from } L_{k-1}\ p,\ L_{k-1}\ q$$
$$\textbf{where} p.\text{item}_1 = q.\text{item}_1,\ p.\text{item}_2 = q.\text{item}_2,\ \ldots, p.\text{item}_{k-2} = q.\text{item}_{k-2},\ p.\text{item}_{k-1} < q.\text{item}_{k-1};$$

Figura 3.10: Função APRIORI_Gen. Fonte: (AMO, 2003)

Fase de Poda de candidatos:

Essa é uma fase simples do APRIORI. Novamente é levada em consideração a propriedade da anti-monotonia. Sabe-se que se os candidatos possuírem pelo menos um subconjunto de itens (Subitemsets) de k-1 que não forem frequentes, ele automaticamente será descartado, pois, não tem a mínima possibilidade de ser frequente. Partindo dessa premissa calcula-se os conjuntos de candidatos frequentes a serem submetidos ao suporte pré-estabelecido.

Fase de Cálculo do Suporte

De posse dos candidatos a frequente, faz-se uma passagem na base de dados comparando o grau de suporte estabelecido pelo parâmetro de entrada do algoritmo com o número de incidência dos intemset´s na base de dados. Essa comparação é feita para cada transação; quais os candidatos são suportados e para cada resposta afirmativa é incrementada uma unidade no contador. Caso esse número seja igual ou maior que o suporte ele é considerado um padrão.

3.2.3.3 Descoberta de Padrões Sequenciais

Sequências são conjuntos de itens que representam várias ocorrências em uma ordem cronológica de tempo. Segundo Amo (2003), uma sequência ou padrão sequencial de tamanho k (ou k-sequência) é uma coleção ordenada de itemsets <I1;I2;...; In >. Esse problema foi introduzido por Agrawal e Srikant (1995) e atualmente vem sido bastante utilizado e difundido no meio acadêmico.

Para um correto entendimento do arcabouço teórico que envolve mineração sequencial, alguns conceitos tornam-se fundamentais:

Padrão Sequencial => Seja S uma sequência <S1,S2,...,Sn>, onde Si é um itemset no padrão <A1, A2, ..., An>. A sequência P <P1,P2,...,Pn> é considerada um padrão sequencial se S contém P. Entenda-se: S contém P, onde P é uma subsequência de S, ou seja, S1 contém P1, S2 contém P2 e assim sucessivamente. Um exemplo claro é a sequência <{Carro},{Pneu,CD Player}>. Esse padrão sequencial me informa que "Cliente compra carro, tempos depois compra pneu e CD Player de carro".

Um outro exemplo de padrão sequencial: seja S uma sequência de dados, tal que S = <{a,b}, {a,c}, {b,c,d}, {a,d,e}>. Dessa sequência podemos inferir que:

- O padrão sequencial<{a},{b},{a,d}> está contido em S;
- O padrão sequencial<{c},{e}> está contido em S;
- O padrão sequencial<{c,e},{d}> não está contido em S;
- O padrão sequencial<{a,d},{b}> não está contido em S.

Um outro conceito importante é o de suporte de padrão sequencial. Similar ao suporte calculado nas regras de associação, um padrão sequencial S é suportado por uma sequência T se $S \subseteq T$. Essa definição vem da porcentagem de sequências que aparecem na amostra de dados, ou seja:

X = Número de Sequências da amostra que contém S;

Y = Número total de Sequências da Amostra S

Suporte =>

$$Sup(S) = \frac{X}{Y}$$

Nesse sentido, um padrão sequencial S é dito frequente em relação à amostra de dados D em um nível mínimo de suporte α se, Sup(S) $\geq \alpha$.

Como exemplo, apresentamos na figura 3.11,uma base de produtos de uma loja de departamentos, retirado de Amo (2003).

IdCl	Sequências de Itemsets
1	< {TV, ferro-elétrico}, {sapato, lençol}, {biscoito, açúcar} >
2	< {sapato, **aparelho-de-som, TV**}, {lençol, **Vídeo**}, {**DVDPlayer**, fax} >
3	< {**aparelho-de-som, TV**, ventilador}, {**Vídeo**, fitas-de-vídeo}, {**DVDPlayer**, liquidificador}>
4	< {iogurte, suco}, {telefone}, {TV, Vídeo} >

Figura 3.11: Base de Dados de Sequência de compras de clientes. Fonte: (AMO, 2003)

Alguns padrões foram identificados e seus respectivos suportes foram calculados:

- <{Sapato}> => 50%
- <{Sapato},{Açúcar}> => 25%
- <{TV}> => 100%
- <{TV, Aparelho-de-Som}, {Vídeo},{DVD-Player}> => 50%

Para que um padrão sequencial S seja considerado frequente em relação a uma amostra de dados qualquer, ele deve atender ao limiar mínimo estabelecido para o suporte. Caso o padrão mínimo estabelecido no nosso exemplo seja 75%, apenas o padrão sequencial <{TV}> seria dito frequente. Caso esse valor seja ajustado para 50%, além do padrão <{TV}>, teríamos ainda as sequencias, <{Sapato}> e <{TV, Aparelho-de-Som}, {Vídeo},{DVD-Player}> fazendo parte dos padrões sequenciais resultantes. Notem que as sequências são calculadas em relação ao itemset. Ou seja, apesar de um item poder vir em uma sequência invertida, o importante é a sequencia de itemset´s.

Mineração de sequências com restrições

A mineração de padrões sequenciais frequentes cada vez mais se configura como uma importante ferramenta de extração de conhecimento em bases de dados sequenciais. Isto se dá pelo poder de adaptação dos dados de vários domínios de aplicação ao padrão sequencial. Essa flexibilidade porém, gera novas demandas aos mineradores que não estão previstas nos algoritmos. Para resolver essas demandas algumas restrições foram incorporadas às aplicações originais.

Segundo Amo (2003), restrições são condições impostas pelos usuários, que os padrões sequenciais devem satisfazer a fim de serem minerados. As restrições podem ser classificadas em duas categorias:

- Restrições de Geração => Aquelas que são implementadas na fase de geração de itemset´s candidatos a serem frequentes dos algoritmos. A grande vantagem delas é que diminuem o espaço de busca de padrões;

- Restrições de Validação => São aquelas que só podem ser observadas após a geração dos padrões, na fase de cálculo do suporte. Otimizam a fase de geração, porém tem um grande custo computacional de validação.

Para o escopo do nosso trabalho vamos abordar a restrição de validação MIN-MAX. Essa restrição foi introduzida por Srikant e Agrawal (1996) e é importante para identificar o intervalo de tempo entre ocorrências dos padrões sequenciais encontrados. Por exemplo, o padrão sequencial <x;y> com um intervalo de tempo de 1 dia é completamente diferente de um padrão <x;y> com um intervalo de tempo de 1 ano. O intervalo de tempo entre as sequências é muito grande e deve ser desconsiderado na avaliação do minerador. Da mesma forma, se um itemset qualquer for verificado e logo em seguida, em um curto intervalo de tempo, o outro também ocorrer, este deve ser desconsiderado para contagem do suporte. Portanto, impõem-se dois limites, um limite mínimo *m* e um limite máximo *M*.

Formalmente temos:

Um par de inteiros => *(m,M)*

Padrão Sequencial => S = <S1,S2,...,Sn>

X = Número de Sequências da amostra que contém S com restrição(m,M);

Y = Número total de Sequências da Amostra S

Suporte =>

$$Sup(S) = \frac{X}{Y}$$

Por exemplo, considere a base de dados de clientes da figura 3.12 composta de duas colunas. A coluna "idCli"representa o identificador das clientes. A coluna "Sequências de Itemsets"representa uma sequência de produtos adquiridos por esses clientes e um número associado que indica a unidade de tempo entre as compras realizadas. Levando-se em consideração o padrão sequencial S = <a,b,c>, com m = 2, M = 6, teremos:

IdCli	Sequências de Itemsets
1	<(a,1),(b,3),(c,7)>
2	<(a,1),(b,2) ,(c,4)>
3	<(a,1),(b,2) ,(c,8)>

Figura 3.12: Base de Dados de Sequência com informações de Tempo. Fonte: Autoria Própria.

Os resultados da mineração podem ser observados na figura 3.13. A primeira sequência observada atende ao padrão sequencial <a,b,c> com as restrições m=2 e M=6. As outras duas sequência observadas atende ao padrão sequencial <a,b,c>,porém saem do escopo da mineração porque não obedecem as restrições de tempo estipuladas na entrada do algoritmo.

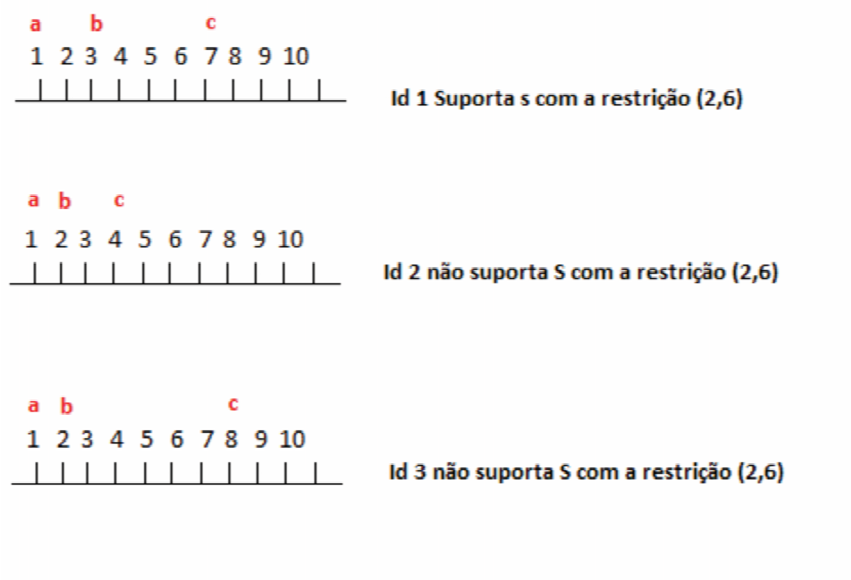


Figura 3.13: Resultado visual dos padrões sequenciais analisados . Fonte: Autoria Própria.

3.2.3.4 AprioriAll

O AprioriAll é um algoritmo da família Apriori e foi proposto por Agrawal e Srikant (1995). Foi concebido da necessidade de descoberta de padrões sequenciais frequentes, valendo-se da estrutura do APRIORI.

Por ser da família do Apriori, também obedece a propriedade da antimonotonicidade descrita no sub tópico 3.2.3.2. O objetivo neste algoritmo é encontrar sequências de itens que ocorrem com uma determinada frequência na base de dados. Para atingir o objetivo, a base de dados precisa de algum indicador de sequência, que, nem sempre é a ordem de registro de transações. O AprioriAll identifica sequências de itens que podem ser representadas como <X, Y, Z> onde X, Y e Z são itemsets frequentes que ocorrem segundo a ordem apresentada.

A base de dados deverá passar por uma transformação inicial que traduz as sequências de itens em sequências de registros. Para realizar este procedimento, as transações passam a ser representadas em uma mesma tupla, por exemplo: <{20, 30}, {10}, {4, 12} >. Este registro indica que os itens foram registrados nessa sequência em grupos de itemsets. Reescrevendo a sequência, representando os conjuntos de itens como itemsets, temos: < A, B, C >, onde A = {20, 30}, B = {10} e C = {4, 12}.

O Aprioriall é composto de três fases. Na primeira fase são gerados os itens frequentes. Nessa fase utiliza-se o Apriori tradicional para encontrar os itens frequentes e após esse processo eles são mapeados para números. Esse mapeamento é realizado visando uma otimização do processo.

A segunda fase é a fase de transformação. Nessa fase do processo, aplica-se a propriedade da antimonotonidade aos itens frequentes encontrados. Como o objetivo é calcular as sequências frequentes e, segundo a propriedade, todo itemset de uma sequência frequente deve ser frequente, esse processo facilita o cálculo do suporte padrão. Basicamente, como o mapeamento dos itemset´s já foi feito na primeira etapa, o algoritmo apenas verifica se o número que representa o itemset está presente no padrão sequencial.

Na etapa final do processo, que é a das sequências, aplicamos o limiar do suporte estabelecido e calculamos as sequências frequentes.

3.2.3.5 PrefixSpan

O algoritmo PrefixSpan foi proposto por Pei et al. (2004) e difere das técnicas utilizadas pelos algoritmos da família Apriori. O PrefixSpan trabalha com projeções consecutivas da própria base de dados, a fim de se obter padrões sequenciais diretamente da base, sem passar pela etapa de geração e poda de candidatos. O custo da geração dessas projeções é alto, porém ele reduz os esforços de geração de candidatos e substancialmente o tamanho da base de dados.

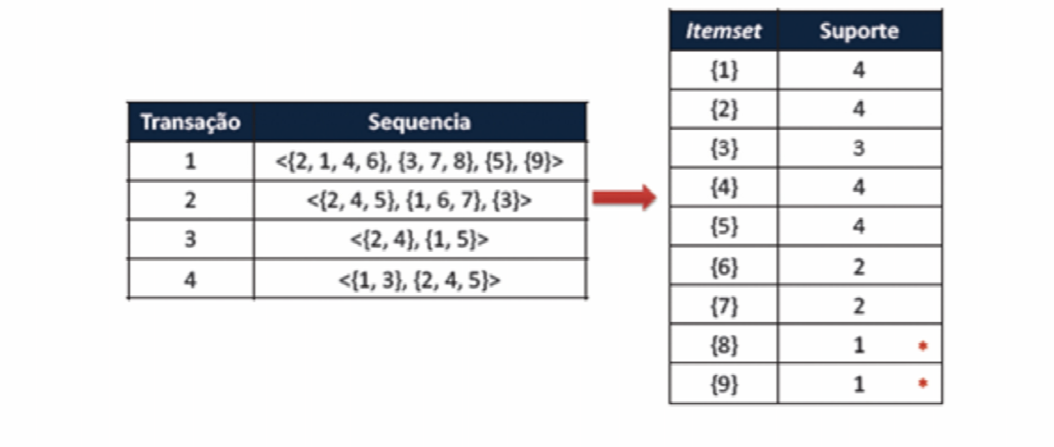
3.2.3.6 Algoritmo *Generalized Sequential Pattern* (GSP)

O algoritmo GSP é um algoritmo similar aos da família do APRIORI. Foi introduzido por Srikant e Agrawal (1996). Sua maior vantagem em relação ao AprioriAll são as otimizações na fase de geração e poda de candidatos.

No algoritmo AprioriAll, em cada iteração k, os conjuntos Lk e Ck (Itemsets frequentes e itemsets candidatos) são constituídos de sequências de k itemsets.

No algoritmo GSP, em cada iteração k, os conjuntos Lk e Ck (Itemsets freqüentes e itemsets candidatos) são constituídos de sequências de k itens. Ou seja, os itemsets frequentes <{A}> e <{B}> dão origem, no AprioriAll, ao candidato <{A}, {B}>. Já no algoritmo GSP, os mesmos dão origem à dois candidatos: <{A},{ B}> e <{A, B}>. Ao invés de darem origem a um candidato que possui dois itemsets, dá origem a dois candidatos que possuem dois itens, estejam eles em itemsets distintos ou não.

O GSP, a exemplo do AprioriAll resume-se a gerar candidatos, calcular o seu suporte e depois proceder a poda dos não conformes. No exemplo abaixo, tirado de Devêza (2011), vemos a sequência de geração do GSP. Na figura 3.14 é ilustrada a fase de geração k = 1 com um suporte de 50%. Tanto a primeira geração de candidatos quanto o calculo do suporte do GSP são similares ao AprioriAll.



Transação	Sequencia
1	<{2, 1, 4, 6}, {3, 7, 8}, {5}, {9}>
2	<{2, 4, 5}, {1, 6, 7}, {3}>
3	<{2, 4}, {1, 5}>
4	<{1, 3}, {2, 4, 5}>

Itemset	Suporte	
{1}	4	
{2}	4	
{3}	3	
{4}	4	
{5}	4	
{6}	2	
{7}	2	
{8}	1	*
{9}	1	*

Figura 3.14: Iteração 1 do GSP - suporte mínimo de 50%. Fonte: (DEVêZA, 2011).

Na segunda iteração apresentada na figura 3.15, cada itemset k1 é combinado com ele e com os outros itemset´s, gerando itemsets não somente de k2 elementos, como também sequências de k2 itemsets.

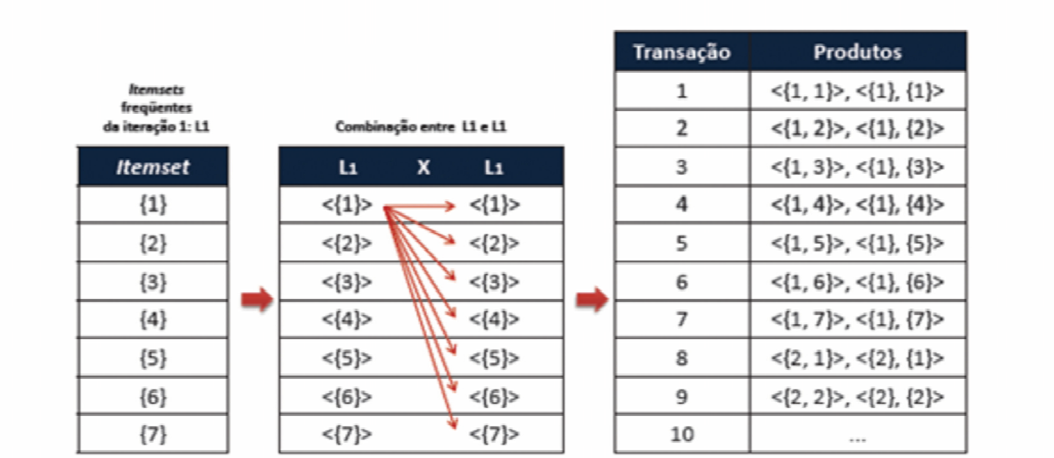


Figura 3.15: Iteração 2 do GSP. Fonte: (DEVêZA, 2011).

Após a geração de candidatos, o algoritmo GSP realiza o cálculo do suporte e a poda. O cálculo do suporte é realizado através da construção de uma árvore hash[3]. Os nós da árvore hash são preenchidos com os itemsets candidatos e após esse procedimento, a base de dados é novamente percorrida para obtenção das frequências dos candidatos presentes nesses nós para a definição da poda ou não dos mesmos.

[3]Uma árvore-hash é uma árvore onde as folhas armazenam conjuntos de itemsets, e os nós intermediários (inclusive a raiz) armazenam tabelas hash contendo pares do tipo (número, ponteiro) (AMO, 2003).

3.3 Ferramentas

Nessa sessão discutiremos duas ferramentas que foram utilizadas como auxilio na implementação do modelo TMIGRAF´s: o *Sequential Pattern Mining Framework* (SPMF) e o *Java Universal Network/Graph Framework* (JUNG).

3.3.1 SPMF - *Sequential Pattern Mining Framework*

Existe um grande número de bibliotecas e algoritmos isolados que nos auxiliam no processo de mineração de dados, por exemplo: Weka (WITTEN; FRANK, 2008), Rapid Miner (TEAM, 2008), Vis-Stamp (GUO, 2009), SPMF (FOURNIER-VIGER, 2008), dentre outras. Devido à facilidade de utilização, clareza da documentação e poder de adaptação ao escopo do trabalho, SPMF foi a escolhida.

SPMF é uma biblioteca escrita em Java e de código aberto. O software foi projetado e implementado por Philippe Fournier-Viger e sua primeira versão foi lançada em 07/12/2008. Atualmente, SPMF está sob a licença *Creative Commons Attribution-NonCommercial-ShareAlike 2.5 Canada License*.

Originalmente, o SPMF foi concebido para trabalhar apenas com mineração de padrões sequenciais frequentes, porém com as suas recentes atualizações já contemplam algoritmos de mineração de regras de associação, mineração de regras sequenciais e mineração de itemset´s frequentes. A versão atual da biblioteca é a V0.77 que inclui 43 algoritmos e foi lançada em 21 de outubro de 2011. Como visto nos tópicos anteriores, os algoritmos presentes na API[4] da biblioteca e que foram utilizados na pesquisa foram o APRIORI e o GSP modificado.

3.3.2 JUNG - *Java Universal Network/Graph Framework*

O JUNG é uma biblioteca de software desenvolvida na linguagem JAVA que fornece recursos de modelagem, análise e visualização de dados em formato de grafos ou redes. Trata-se de um framework de código aberto agregado junto ao projeto SourceForce. Sua arquitetura é projetada para suportar uma variedade de representação de entidades e suas relações tais como: grafos direcionados e não direcionados, grafos multi-modais, grafos com bordas paralelas e hiper-grafos (TEAM, 2012). Possui flexibilidade de manipulação de inclusão e exclusão de vértices e arestas, manipulação de rótulos e criação de aplicações de análises de propriedades dos grafos, como graus de vértices, isomorfismo de grafos, representação de matrizes de adjacência e incidência, entre outros.

Alem da manipulação dessas entidades, o JUNG contempla uma vasta coleção de algoritmos de teoria dos grafos, mineração de dados, análise de redes sociais, análise de click´s, cálculos estatísticos, médias de distância de redes e medidas de importância de redes (Centralidade, PageRank, HITS,etc.). A figura 3.16 e 3.17 apresentam alguns exemplos do poder de visualização do JUNG.

[4]Uma Application Programming Interfaces (API) é uma interface bem definida, que permite a um componente de software acessar outro componente de forma transparente através de rotinas de programação (SOUZA et al., 2004).

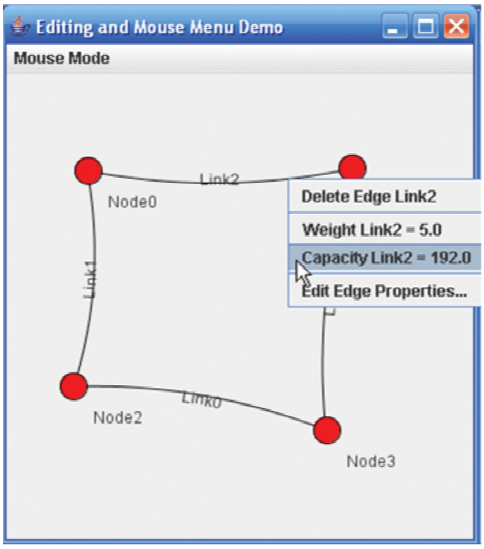


Figura 3.16: Aplicação de manipulação de arestas em JUNG. Fonte: (BERNSTEIN, 2009).

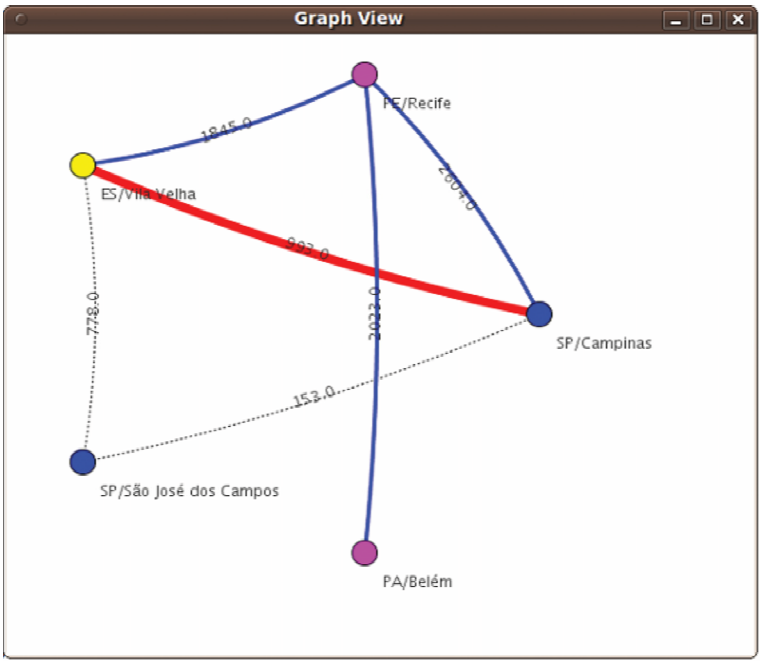


Figura 3.17: Aplicação de análise de tráfego na malha viária. Fonte : (SANTOS, 2010).

Capítulo 4

Estudo de Caso

Este capítulo apresenta a aplicação do modelo TMIGRAF´s a uma base de dados de objetos espaço temporais, fazendo a demonstração da eficácia do modelo e apresentando os padrões encontrados, tanto de regras de associação como de padrões sequenciais frequentes em forma de grafos. Discute os resultados extraídos, apresentando as demonstrações e análises referentes aos conhecimentos e padrões adquiridos.

4.1 Base de dados do estudo de caso

Para comprovar a eficiência do modelo TMIGRAF´s, foi submetido a ele uma base de dados de participantes de uma conferência. Antes de apresentarmos os testes realizados e os resultados obtidos, faz-se necessária uma breve explanação dos conceitos utilizados e como eles foram organizados semanticamente.

4.2 Participantes de Conferência

O conjunto de dados abordado na pesquisa, que serviu de base para todo o processo de amadurecimento do trabalho foi a base de dados dos deslocamentos de um grupo de participantes de uma conferência. Trata-se de uma base de dados sintética. Essa base foi criada a partir de um algoritmo randômico desenvolvido em PL/SQL e que tomou como base o exemplo de dados fornecido por Alvares et al. (2007).

4.2.1 Modelo Conceitual Utilizado

Pela natureza dos dados, dois modelos conceituais foram desenvolvidos. Um para modelar os dados dos movimentos das trajetórias e outro para modelar a parte geográfica dos dados.

A figura 4.1 apresenta o modelo conceitual de trajetórias[1]. Notem que ele é composto de quatro entidades e as suas devidas cardinalidades.

[1]Como trata-se de uma base de dados espaciais, a notação utilizada segue os moldes proposto por Gazola e Filho (2005)

Trajectory - É a entidade que representa um conjunto de trajetórias. O conceito representa uma lista ordenada de paradas S e movimentos M;

Stop - Representa os pontos de paradas das trajetórias. Esses pontos de paradas são representados pelas características espaciais predefinidas em um intervalo de tempo não vazio. Vale salientar que os Stop´s são semanticamente dependentes da aplicação. Um semáforo pode ser considerado um stop em uma aplicação de controle de trânsito e provavelmente não será em uma aplicação de turismo;

Move - É onde são registrados os movimentos das trajetórias. É a parte da trajetória que possui um intervalo de tempo delimitado por dois pontos consecutivos S1 e S2;

SpatialFeatureType - Esta entidade representa locais do mundo real presentes na superfície da terra. É nessa entidade que agruparemos nossas características geográfias.

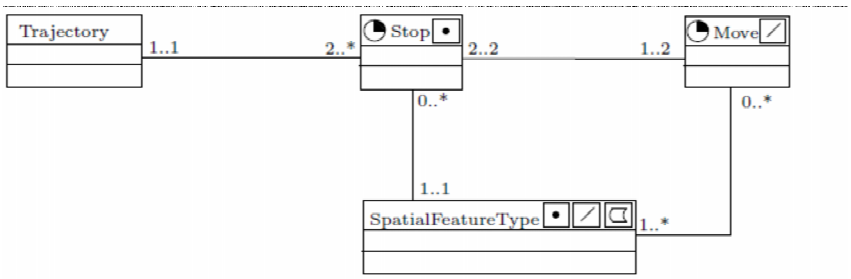


Figura 4.1: Schema conceitual de trajetórias. Fonte : (ALVARES et al., 2007)

A figura 4.2 apresenta o modelo de dados dos locais espaciais em que se baseou o nosso estudo de caso. Essa modelagem foi estendida e adaptada à realidade da aplicação. No modelo original apenas quatro locais foram incorporadas na análise, são elas "*Hotel*", "*ConferenceCenter*", "*TouristicPlace*"e "*Airport*". Além dessas, e, por questões de clareza de apresentação de resultados do trabalho, foram incorporados, "*ShoppingCenter*", "*University*", "*Beach*"e "*FootballStadium*". A questão das cardinalidades segue o padrão estabelecido.

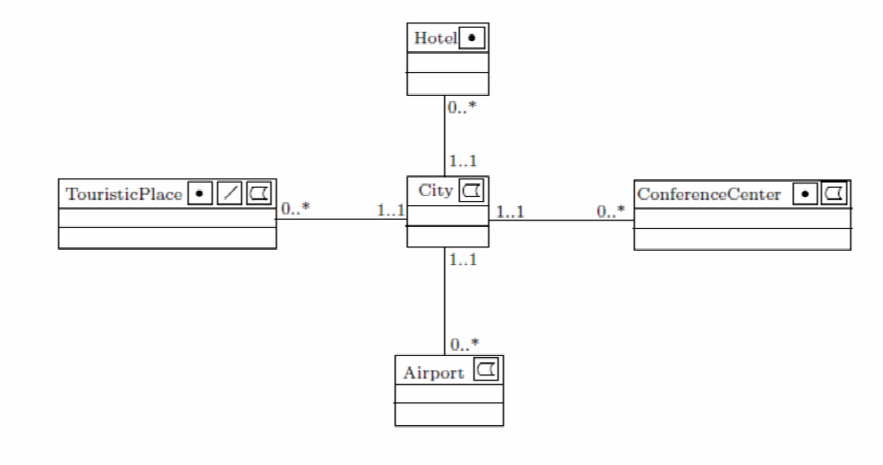


Figura 4.2: Parte do Schema conceitual das características geográficas. Fonte : (ALVARES et al., 2007)

4.2.2 Dados Utilizados

Na figura 4.3 podemos observar os locais utilizados para análise de deslocamento. Foram utilizados locais genéricos encontrados na maioria das cidades e que estivessem presentes na cidade de Mossoró-RN.

	ID	LOC
1	0	Airport
2	1	Conference Center
3	2	Hotel
4	3	Tourist Place
5	4	Shopping Center
6	5	University
7	6	Beach
8	7	football stadium

Figura 4.3: Locais escolhidos para montar a base de deslocamento de cidades. Fonte: Autoria própria

A base é composta por 3628 registros, perfazendo um total de 1000 trajetórias armazenadas como podemos ver no exemplo da figura 4.4. Explicações acerca dos atributos da tabela podem ser encontradas na sessão 3.2.1 do capítulo 3 deste documento.

	MID	TID	STOP1	STOP2	INTTIME
342	2	46	Airport	Hotel	04/10/2011 - 07:25:13
343	3	46	Hotel	Airport	04/10/2011 - 07:25:17
344	4	46	Airport	Tourist Place	04/10/2011 - 07:25:17
345	5	46	Tourist Place	football stadium	04/10/2011 - 07:25:23
346	6	46	football stadium	Conference Center	04/10/2011 - 07:25:37
347	7	46	Conference Center	Hotel	04/10/2011 - 07:25:39
348	8	46	Hotel	University	04/10/2011 - 07:25:43
349	1	47	Beach	Tourist Place	04/10/2011 - 07:25:53
350	2	47	Tourist Place	Hotel	04/10/2011 - 07:25:59
351	3	47	Hotel	Beach	04/10/2011 - 07:26:03
352	4	47	Beach	Airport	04/10/2011 - 07:26:15
353	5	47	Airport	Shopping Center	04/10/2011 - 07:26:15
354	6	47	Shopping Center	Hotel	04/10/2011 - 07:26:23
355	7	47	Hotel	Conference Center	04/10/2011 - 07:26:27
356	8	47	Conference Center	Airport	04/10/2011 - 07:26:29

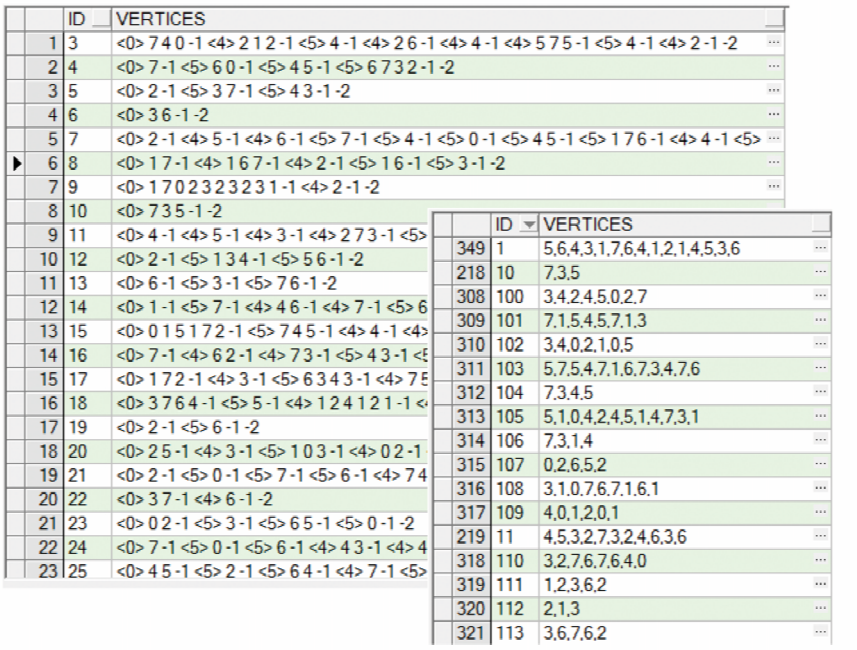
Figura 4.4: Base de dados das trajetórias. Fonte: Autoria própria

4.2.3 Aplicação do modelo TMIGRAFs

Após o processo de limpeza dos dados descrito no subtopico 1.2.1, os mesmos são transfomados em formato de grafos para visualização e no formato de sequências para servir de entrada para os algoritmos APRIORI e GSP modificado, como mostrado nas figuras 4.5 e 4.6 respectivamente.

	ARQ_PROC	LINHA
1	61	t# 1
2	61	v 0 Beach
3	61	v 1 Conference Center
4	61	v 2 Hotel
5	61	v 3 Shopping Center
6	61	v 4 Tourist Place
7	61	v 5 University
8	61	v 6 football stadium
9	61	u 5 0 04/10/2011 - 06:42:13
10	61	u 0 3 04/10/2011 - 06:42:25
11	61	u 3 4 04/10/2011 - 06:42:33
12	61	u 4 1 04/10/2011 - 06:42:39
13	61	u 1 6 04/10/2011 - 06:42:41
14	61	u 6 0 04/10/2011 - 06:42:55
15	61	u 0 3 04/10/2011 - 06:43:07
16	61	u 3 1 04/10/2011 - 06:43:15
17	61	u 1 2 04/10/2011 - 06:43:17
18	61	u 2 1 04/10/2011 - 06:43:21

Figura 4.5: Base de dados em formato de grafos. Fonte: Autoria própria



	ID	VERTICES
1	3	<0> 7 4 0 -1 <4> 2 1 2 -1 <5> 4 -1 <4> 2 6 -1 <4> 4 -1 <4> 5 7 5 -1 <5> 4 -1 <4> 2 -1 -2
2	4	<0> 7 -1 <5> 6 0 -1 <5> 4 5 -1 <5> 6 7 3 2 -1 -2
3	5	<0> 2 -1 <5> 3 7 -1 <5> 4 3 -1 -2
4	6	<0> 3 6 -1 -2
5	7	<0> 2 -1 <4> 5 -1 <4> 6 -1 <5> 7 -1 <5> 4 -1 <5> 0 -1 <5> 4 5 -1 <5> 1 7 6 -1 <4> 4 -1 <5>
6	8	<0> 1 7 -1 <4> 1 6 7 -1 <4> 2 -1 <5> 1 6 -1 <5> 3 -1 -2
7	9	<0> 1 7 0 2 3 2 3 2 3 1 -1 <4> 2 -1 -2
8	10	<0> 7 3 5 -1 -2
9	11	<0> 4 -1 <4> 5 -1 <4> 3 -1 <4> 2 7 3 -1 <5>
10	12	<0> 2 -1 <5> 1 3 4 -1 <5> 5 6 -1 -2
11	13	<0> 6 -1 <5> 3 -1 <5> 7 6 -1 -2
12	14	<0> 1 -1 <5> 7 -1 <4> 4 6 -1 <4> 7 -1 <5> 6
13	15	<0> 0 1 5 1 7 2 -1 <5> 7 4 5 -1 <4> 4 -1 <4>
14	16	<0> 7 -1 <4> 6 2 -1 <4> 7 3 -1 <5> 4 3 -1 <5
15	17	<0> 1 7 2 -1 <4> 3 -1 <5> 6 3 4 3 -1 <4> 7 5
16	18	<0> 3 7 6 4 -1 <5> 5 -1 <4> 1 2 4 1 2 1 -1 <4
17	19	<0> 2 -1 <5> 6 -1 -2
18	20	<0> 2 5 -1 <4> 3 -1 <5> 1 0 3 -1 <4> 0 2 -1
19	21	<0> 2 -1 <5> 0 -1 <5> 7 -1 <5> 6 -1 <4> 7 4
20	22	<0> 3 7 -1 <4> 6 -1 -2
21	23	<0> 0 2 -1 <5> 3 -1 <5> 6 5 -1 <5> 0 -1 -2
22	24	<0> 7 -1 <5> 0 -1 <5> 6 -1 <4> 4 3 -1 <4> 4
23	25	<0> 4 5 -1 <5> 2 -1 <5> 6 4 -1 <4> 7 -1 <5>

	ID	VERTICES
349	1	5,6,4,3,1,7,6,4,1,2,1,4,5,3,6
218	10	7,3,5
308	100	3,4,2,4,5,0,2,7
309	101	7,1,5,4,5,7,1,3
310	102	3,4,0,2,1,0,5
311	103	5,7,5,4,7,1,6,7,3,4,7,6
312	104	7,3,4,5
313	105	5,1,0,4,2,4,5,1,4,7,3,1
314	106	7,3,1,4
315	107	0,2,6,5,2
316	108	3,1,0,7,6,7,1,6,1
317	109	4,0,1,2,0,1
219	11	4,5,3,2,7,3,2,4,6,3,6
318	110	3,2,7,6,7,6,4,0
319	111	1,2,3,6,2
320	112	2,1,3
321	113	3,6,7,6,2

Figura 4.6: Bases de dados em formato de sequências temporais e sequencias simples. Fonte: Autoria própria

4.2.4 Resultados

Foram realizados os testes de mineração com 8 localidades geográficas distribuídas em uma base de dados de 500 trajetórias de tamanhos diferentes, em um total de 3628 registros.

Nos primeiros testes realizados, a base de dados foi submetida ao algoritmo APRIORI para obtenção de itemset´s frequentes e regras de associação. Ao algoritmo foi aplicado um suporte mínimo de 40% e uma confiança de 60%. Foram encontrados 36 itens frequentes distribuídos em um de tamanho zero L0[2], oito de tamanho um L1, dezenove de tamanho L2, seis de tamanho L3 e dois de tamanho L4. Além desses resultados, também foram encontradas 55 regras de associação. Na figura 4.7, apresentamos a parte do relatório de saída referente a descoberta dos itens frequentes. Cada número representa uma localidade mapeada. O padrão 13 por exemplo é um padrão de tamanho 2 e indica que em 45% das ocorrências, existe uma trajetória que passa pela localidade 0, no caso "*Airport*", e pela localidade 5, no caso "*University*". À medida que o tamanho dos itens frequentes vai aumentando, os padrões vão se tornando mais interessantes. No caso do padrão 36, temos que em 40% das ocorrências do banco, temos que uma trajetória passa pelo "*ShoppingCenter*", "*University*", "*Beach*"e "*Footbal Stadium*".

[2]Padrão do algoritmo para indicar o tamanho da base de dados. Semanticamente irrelevante.

```
------- FREQUENT ITEMSETS -------
L0
pattern 0:  support :  1 (500/500)
L1
pattern 1:  0 support :  0,62 (310/500)
pattern 2:  1 support :  0,61 (304/500)
pattern 3:  2 support :  0,68 (341/500)
(...)
L2
pattern 9:  0 1 support :  0,41 (203/500)
pattern 10:  0 2 support :  0,46 (229/500)
pattern 11:  0 3 support :  0,43 (214/500)
pattern 12:  0 4 support :  0,44 (222/500)
pattern 13:  0 5 support :  0,45 (224/500)
pattern 14:  0 6 support :  0,46 (230/500)
pattern 15:  0 7 support :  0,43 (214/500)
pattern 16:  1 2 support :  0,46 (232/500)
pattern 17:  1 3 support :  0,43 (215/500)
pattern 18:  1 4 support :  0,43 (213/500)
pattern 19:  1 5 support :  0,4 (202/500)
(...)
L3
pattern 29:  3 6 7 support :  0,48 (239/500)
pattern 30:  3 5 7 support :  0,46 (231/500)
pattern 31:  2 4 5 support :  0,49 (244/500)
(...)
L4
pattern 35:  2 3 5 7 support :  0,41 (206/500)
pattern 36:  4 5 6 7 support :  0,42 (212/500)
```

Figura 4.7: Itens fequentes gerados pelo algoritmo APRIORI aplicado às trajetórias de participantes. Fonte: Autoria própria

Na figura 4.8 podem ser observadas as regras de associação geradas. Essas regras se reportam a incidência de locais associados nas mesmas trajetórias. No caso dessas regras o algoritmo apresenta o suporte mínimo e a confidência que serve como mais um parâmetro de validação da regra.

```
------- All association rules -------
rule 0:  1  ==> 0 support :  0.406 (203/500) confidence :  0.6677631578947368
rule 1:  0  ==> 1 support :  0.406 (203/500) confidence :  0.6548387096774193
rule 2:  0  ==> 2 support :  0.458 (229/500) confidence :  0.7387096774193549
rule 3:  2  ==> 0 support :  0.458 (229/500) confidence :  0.6715542521994134
rule 4:  3  ==> 0 support :  0.428 (214/500) confidence :  0.6407185628742516
rule 5:  0  ==> 3 support :  0.428 (214/500) confidence :  0.6903225806451613
rule 6:  0  ==> 4 support :  0.444 (222/500) confidence :  0.7161290322580646
rule 7:  4  ==> 0 support :  0.444 (222/500) confidence :  0.6397694524495677
rule 8:  5  ==> 0 support :  0.448 (224/500) confidence :  0.6808510638297872
rule 9:  0  ==> 5 support :  0.448 (224/500) confidence :  0.7225806451612903
rule 10:  0  ==> 6 support :  0.46 (230/500) confidence :  0.7419354838709677
rule 11:  6  ==> 0 support :  0.46 (230/500) confidence :  0.6571428571428571
rule 12:  0  ==> 7 support :  0.428 (214/500) confidence :  0.6903225806451613
rule 13:  7  ==> 0 support :  0.428 (214/500) confidence :  0.654434250764526
rule 14:  1  ==> 2 support :  0.464 (232/500) confidence :  0.7631578947368421
rule 15:  2  ==> 1 support :  0.464 (232/500) confidence :  0.6803519061583577
(...)
```

Figura 4.8: Regras de associação geradas pelo algoritmo APRIORI aplicado às trajetórias de participantes. Fonte: Autoria própria

Na busca por padrões sequenciais, a segunda fase de teste foi realizada, agora aplicando o algoritmo GSP. Devido ao baixo número de sequências frequentes presentes na base e à dificuldade de análise de conhecimento nos dados, foi aplicado um suporte mínimo considerado baixo -10%- e um intervalo de tempo considerável, no caso 8 unidades de tempo, para que mais regras pudessem ser analisadas. Com a aplicação desses parâmetros de entrada ao algoritmo, 79 padrões sequenciais frequentes com restrição de tempo foram encontrados, divididos em 40 padrões de tamanho L1, 32 de tamanho L2 e 7 de tamanho L3.

Seguindo a mesma notação adotada no algoritmo APRIORI, o GSP utiliza o mapeamento dos locais em números e esses são minerados, melhorando assim o desempenho do algoritmo.

Como explicado no subtopico 3.2.3.6, o GSP busca padrões sequenciais frequentes, aplicando restrições de tempo. Podemos citar por exemplo o padrão 47 presente na figura 4.9. Esse padrão significa que em 14% das vezes, ou seja, em 70 ocorrências dos 500 registros da base de dados, no tempo zero ele passa pelo local 2 ou seja pelo "*Hotel*" e quatro unidades de tempo após, ele passa pelo local 4, ou seja, "*ShoppingCenter*". Um outro padrão interessante, também visualizado na figura 4.9 é o padrão 79. É um padrão L3, onde no instante zero a trajetória sai do "*TouristPlace*", no instante 3 passa pelo "*Hotel*" e no instante 4 passa pelo "*ConferenceCenter*". É importante salientar que no momento que as trajetórias foram transformadas em sequências temporais, as unidades de tempo foram transformadas em séries temporais, ou seja, cada deslocamento tem um peso associado ao seu tempo e esse peso foi convertido em unidade de tempo. Essa transformação fez-se necessária por questões de processamento do algoritmo e claridade na interpretação dos padrões minerados.

```
============  Algorithm - STATISTICS =============
 Total time ~ 484 ms
 Frequent sequences count : 79
 ----------FREQUENT SEQUENCES WITH TIME + CLUSTERING -------
  L0
  L1
  pattern 1:  {t=0, 3 1 }    support :  0,11 (53/500)
  pattern 2:  {t=0, 3 6 }    support :  0,11 (55/500)
  pattern 3:  {t=0, 3 4 }    support :  0,14 (71/500)
  pattern 4:  {t=0, 3 }    support :  0,67 (334/500)
  pattern 5:  {t=0, 2 1 }    support :  0,11 (56/500)
  pattern 6:  {t=0, 2 6 }    support :  0,1 (51/500)
  pattern 7:  {t=0, 2 3 }    support :  0,16 (79/500)
  pattern 8:  {t=0, 2 4 }    support :  0,11 (55/500)
  pattern 9:  {t=0, 2 }    support :  0,68 (341/500)
  pattern 10:  {t=0, 4 2 }    support :  0,15 (73/500)
  (...)
  L2
  pattern 41:  {t=0, 3 }{t=1, 4 }    support :  0,13 (65/500)
  pattern 42:  {t=0, 3 }{t=2, 5 }    support :  0,12 (59/500)
  pattern 43:  {t=0, 3 }{t=1, 6 }    support :  0,11 (57/500)
  pattern 44:  {t=0, 3 }{t=1, 7 }    support :  0,1 (51/500)
  pattern 45:  {t=0, 3 }{t=3, 0 }    support :  0,11 (56/500)
  pattern 46:  {t=0, 3 }{t=2, 2 }    support :  0,12 (58/500)
  pattern 47:  {t=0, 2 }{t=4, 4 }    support :  0,14 (70/500)
  pattern 48:  {t=0, 2 }{t=2, 5 }    support :  0,12 (60/500)
  pattern 49:  {t=0, 2 }{t=3, 0 }    support :  0,11 (54/500)
  pattern 50:  {t=0, 2 }{t=5, 1 }    support :  0,11 (56/500)
  (...)
  L3
  pattern 73:  {t=0, 5 }{t=2, 2 }{t=3, 0 }    support :  0,11 (54/500)
  pattern 74:  {t=0, 5 }{t=1, 3 }{t=2, 2 }    support :  0,1 (51/500)
  pattern 75:  {t=0, 4 }{t=4, 5 }{t=5, 1 }    support :  0,11 (53/500)
  pattern 76:  {t=0, 4 }{t=2, 6 }{t=4, 5 }    support :  0,11 (54/500)
  pattern 77:  {t=0, 4 }{t=2, 7 }{t=3, 2 }    support :  0,11 (55/500)
  pattern 78:  {t=0, 6 }{t=2, 2 }{t=4, 5 }    support :  0,11 (57/500)
  pattern 79:  {t=0, 7 }{t=3, 3 }{t=4, 4 }    support :  0,11 (54/500)
 ------------------------------- Patterns count : 79================================================
```

Figura 4.9: Padrões sequenciais frequentes com restrição de tempo. Fonte: Autoria própria

Devido ao padrão textual observado no relatório e a extensão do mesmo, o modelo proposto utiliza-se do poder de visualização dos grafos para facilitar o entendimento dos resultados dos padrões minerados. A figura 4.10. apresenta a visualização do padrão 10 de formato L1, presente na figura 4.9. Já a figura 4.11 apresenta o padrão 76 de formato L3, presente na figura 4.9.

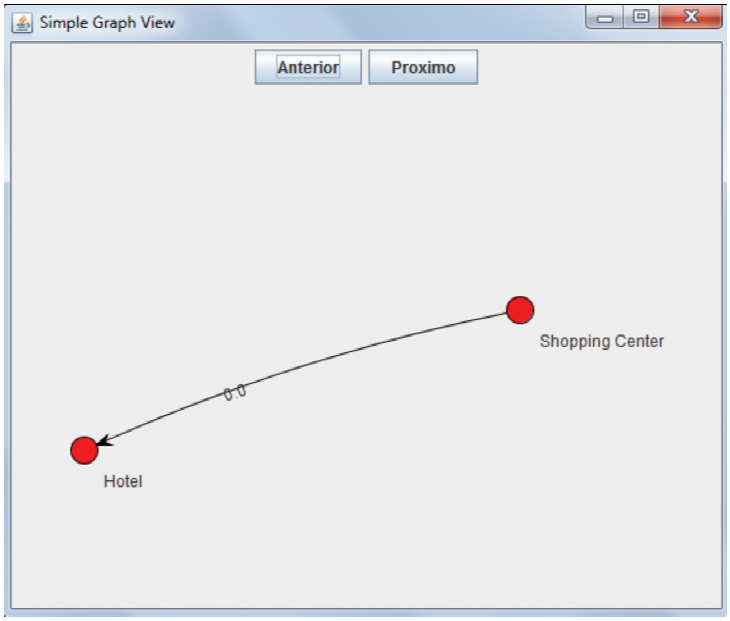


Figura 4.10: Padrões sequencial com restrição de tempo L2 minerado em formato de grafo. Fonte: Autoria própria

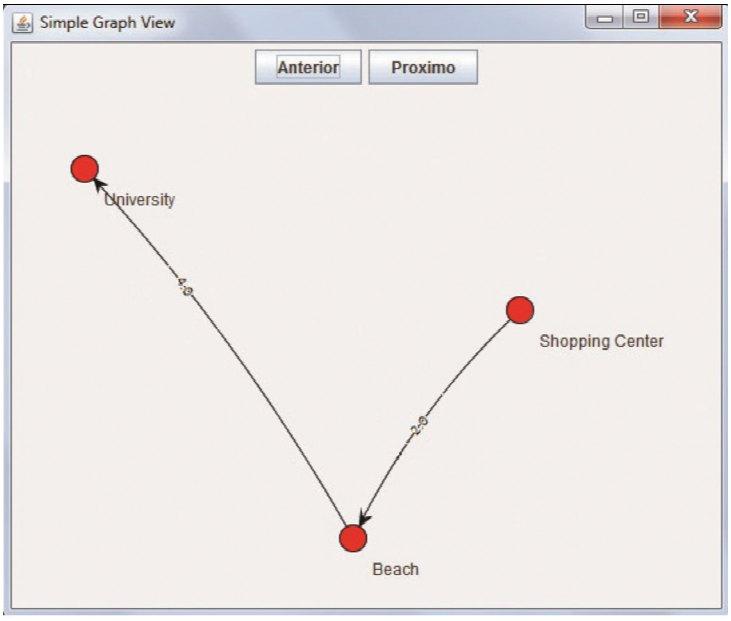


Figura 4.11: Padrões sequencial com restrição de tempo L3 minerado em formato de grafo. Fonte: Autoria própria

Dentre os vários padrões minerados, muitos não possuem um significado considerável ou relevância maior. Porém, dentre os padrões encontrados alguns apresentam uma semântica interessante e que pode ser aplicada em tomada de decisões gerenciais. Pro exemplo, as regras de associação possibilitam inferir

que os locais mais frequentados pela a amostra da nossa base são o "*Airport*", "*ConferenceCenter*" e o "*Hotel*". Essa informação pode ser útil para empresas de marketing, agencias de turismo ou mesmo órgãos governamentais interessados em organizar o trânsito da localidade.

Outro padrão que pode ser bem aproveitado são os padrões sequenciais com restrição de tempo. O padrão apresentado na figura 4.11 pode me levar a conclusão que pessoas se deslocam frequentemente, saindo do shopping center, passando pela praia e depois indo a universidade em um tempo especifico. Esse tipo de padrão pode ser útil para identificar o tipo de transporte que foi utilizado nos trajetos, levando em consideração o tempo de deslocamento, para organização do trânsito, melhoramento de estradas, instalação de posto de fiscalização policial ou mesmo centros de saúde como o SAMU - Serviço de Atendimento Móvel de Urgência.

Capítulo 5

Considerações finais e Trabalhos Futuros

Nos últimos anos, a mineração de dados tem crescido e apresentado ótimos resultados, embora ainda apresente grandes desafios. Uma das principais motivações desse trabalho foi propor uma abordagem para facilitar a extração de padrões e geração de conhecimento a partir da exploração automatizada de dados de objetos espaço-temporais. A fim de detectar padrões de movimentos e de se tentar prever o perfil dos deslocamentos, este trabalho direcionou-se para propor um modelo computacional de mineração das trajetórias que se beneficiasse do poder de simplificação e visualização da teoria dos grafos juntamente com as técnicas de mineração de dados de associação e padrões sequenciais frequentes.

O Modelo proposto neste trabalho mostrou-se uma solução eficiente quanto ao objetivo que lhe é destinado, isto é, fornecer um modelo de descoberta de conhecimento sob bases de dados de trajetórias que utilizasse técnicas de mineração de dados não convencionais para abordar todos os aspectos de movimento envolvidos nos mesmos. Portanto, verifica-se que os objetivos contemplados nesse estudo foram alcançados, podendo-se elencar, entre outros benefícios:

- A proposta do modelo de descoberta de conhecimento em trajetórias abordado no capitulo 3;
- A implementação do modelo apresentado através do desenvolvimento de algoritmos de manipulação, transformação e adaptação de dados, aliados as ferramentas de auxílio a mineração de dados e visualização dos mesmos em formato de grafos;
- Detecção e geração de padrões gerenciais úteis a gestores de controle urbano, turismo ou deslocamento logístico de pessoas em geral, através da descoberta de conhecimento. Esse objetivo foi alcançado ao longo do capitulo 4;
- Aplicação de um novo paradigma de mineração de objetos espaço-temporais, aproveitando-se das características preditivas dos algoritmos de aprendizagem de máquina não supervisionados, aliado ao poder de visualização e interpretação dos grafos;

Alguns desafios e dificuldades foram encontrados durante o desenvolvimento desse modelo de descoberta de conhecimento em bases de dados de trajetórias, entre eles:

- A dificuldade inicial neste trabalho foi a obtenção de bases de dados de trajetórias. Por tratar-se de um tipo de dado espacial relativamente novo, essas dados não estão facilmente disponíveis e muitas vezes, quando disponíveis não estão modelados no formalismo utilizados pelos mineradores;
- Por ser um trabalho que envolve outra área de conhecimento, no caso a física, se fez necessário um estudo específico de como se comporta um objeto espaço-temporal e suas peculiaridades;
- A principal dificuldade foi a obtenção de um minerador, que resolvesse o problema a que se propõe. Na maioria das vezes, as bibliotecas de mineradores não contemplavam todas as funcionalidades descritas na documentação ou não extraiam padrões satisfatórios, contemplando as duas dimensões propostas no modelo.

Após a conclusão desse estudo, recomenda-se algumas linhas de pesquisa a serem seguidas que podem nortear novos trabalhos. Alguns pontos merecem aprofundamento em pesquisas e trabalhos futuros. São eles:

- Utilização de outras técnicas de mineração de dados não contempladas nesse estudo, como por exemplo, redes neurais, algoritmos genéticos, dentre outras;
- O estudo de mecanismos inteligentes para detecção do tamanho da amostra de dados e dos parâmetros ideais a serem aplicados nos algoritmos de mineração de dados;
- Utilização de técnicas de geoprocessamento no sentido de visualização das trajetórias em sistemas de informação geográficas e apresentação de padrões minerados em mapas;
- Realização de pesquisas no intuito do desenvolvimento de um dispositivo semântico de auxílio à poda, interpretação e visualização de padrões minerados.

Finalmente, espera-se que esse trabalho contribua para o engrandecimento da área de mineração de dados, notadamente as dos dados espaciais e não convencionais, servindo de apoio para novos desafios acerca da problemática das entidades móveis e suas nuances.

Referências Bibliográficas

AGRAWAL, R.; IMIELINSKI, T.; SWAMI, A. Mining association rules between sets of items in large databases. *In Proc. of the ACM SIGMOD Conference on Management of Data*, D.C., v. 2, p. 207–216, May 1993.

AGRAWAL, R.; SRIKANT, R. Fast algorithms for mining association rules in large databases. In: BOCCA, J. B.; JARKE, M.; ZANIOLO, C. (Ed.). *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago de Chile, Chile*. [S.l.]: Morgan Kaufmann, 1994. p. 487–499. ISBN 1-55860-153-8.

AGRAWAL, R.; SRIKANT, R. Mining sequential patterns. In: *Proceedings of the Eleventh International Conference on Data Engineering*. Washington, DC, USA: IEEE Computer Society, 1995. (ICDE '95), p. 3–14. ISBN 0-8186-6910-1. Disponível em: <http://dl.acm.org/citation.cfm?id=645480.655281>.

ALVARES, L. O. et al. Dynamic modeling of trajectory patterns using data mining and reverse engineering. In: *In Twenty-Sixth International Conference on Conceptual Modeling - ER2007 - Tutorials, Posters, Panels and Industrial Contributions*. [S.l.: s.n.], 2007.

ALVAREZ, A. B.; LUQUE, B. *Rede Neural de Kohonen e Outras Técnicas para Treinamento Não-Supervisionado*. [S.l.], Abril 2003. 39 p.

AMO, S. Curso de data mining. *Programa de Mestrado em Ciência da Computação*, v. 1, 2003. Disponível em: <http://www.deamo.prof.ufu.br/CursoDM.html>.

AMO, S. Técnicas de mineração de dados. *Sociedade Brasileira de Computação, Universidade Federal da Bahia. (Org.). Jornadas de Atualização em Informática*, v. 2, p. 195–236, 2004.

ANDRIENKO, G. et al. Interactive visual clustering of large collections of trajectories. *IEEE Symposium on Visual Analytics Science and Technology (IEEE VAST)*, Atlantic city, New Jersey, USA, p. 3–10, Outubro 2009.

ARBEX, E. C.; SABOREDO, A. d. P.; MIRANDA, D. Implementação e estudo de caso do algoritmo apriori para mineração de dados. In: . Resende - Rio de Janeiro: Curso de Sistema de Informação, Associação Educacional Dom Bosco, Simpósio de Excelência em gestão de Tecnologia (SEGeT), 2004.

BANG-JENSEN, J.; GUTIN, G. *"Digraphs: Theory, Algorithms and Applications"*. [S.l.]: C Springer-Verlag 2nd Edition, 2009.

BARNETT, V.; LEWIS, T. *Outliers in Statistical Data.* 3rd Edition: Jhon Wiley and Sons, 1994.

BERNSTEIN, G. *Jung 2.0 tutorial.* 2009. [Acessado Junho-2011]. Disponível em: <http://www.grottonetworking.com/JUNG/JUNG2-Tutorial.pdf>.

BOAVENTURA, P. O.; JURKIEWICZ, S. *Grafos: Introdução e Prática.* [S.l.]: Editora Blucher, 2009.

BOGORNY, V. Tutorial on spatial and spatio temporal data mining. *ICDM 2010 The 10th IEEE International Conference on Data Mining*, 2010. Disponível em: <http://datamining.it.uts.edu.au/icdm10/index.php/tutorials>.

CARVALHO, L. A. V. *A mineração de dados no marketing, medicina, economia,engenharia e administração.* [S.l.]: Ciência Moderna, 2005.

CARVALHO, M. A. G. *Isomorfismo e Casamento de Grafos*. 2009. [Acessado Setembro-2011]. Disponível em: <http://www.ft.unicamp.br/ magic/ft024/isomorfismo.pdf>.

CHAKRABARTI, D. *Tools for Large Graph Mining*. Tese (Doutorado) — Carnegie Mellon University, 2005.

CâMARA, G. et al. *Banco de Dados Geográficos*. Editora MundoGEO, 2005. Acesso em 06 de Set. de 2011. Disponível em: <http://www.dpi.inpe.br/livros/bdados/capitulos.html>.

COSTA, C. N. et al. Descoberta de conhecimento em bases de dados. *Revista Eletronica*, v. 2, 2009.

CUBAS, U. B. *Grafos e Digrafos*. 2008. [Acessado Setembro-2011]. Disponível em: <http://www2.brazcubas.br/professores1/arquivos/9_migliano/Apostilas/grafos.pdf>.

DEVêZA, C. H. *Minerando Padrões Sequenciais para Bases de Dados de Lojas Virtuais.* 30 f. Monografia (Monografia) — Universidade Federal de Ouro Preto - UFOP, Ouro Preto, 2011.

EGENHOFER MAX J., F. R. D. Point-set topological spatial relations. *International Journal of Geographical Information Systems*, London, v. 5, p. 161–174, February 1991.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SHYTH, P. From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press*, p. 1–34, 1996.

FEOFILOFF, P.; KOHAYAKAWA, Y.; WAKABAYASHI, Y. *Uma Introdução Sucinta à Teoria dos Grafos.* [S.l.]: II BIENAL SBM, 2007.

FOURNIER-VIGER, P. *SPMF - Sequential Pattern Mining Framework*. 2008. [Acessado Agosto-2011]. Disponível em: <http://www.philippe-fournier-viger.com/spmf/>.

FRANçA, F. S. *Enriquecimento Semântico de Padrões minerados em Grafos Utilizando Ontologias.* Dissertação (Mestrado) — Universidade Estadual do Rio Grande do Norte, Universidade Federal Rural do Semiárido, Mossoró, RN, 2011.

GAZOLA, A.; FILHO, J. L. *Projeto Automatizado de Bancos de Dados Geográficos para Aplicações Small GIS.* [S.l.]: REIC. 2005 ,Ano V., 2005.

GIANNOTTI, F.; TRASARTI, R. *Mobility, Data Mining and Privacy: The GeoPKDD Paradigm.* 2000. PKDD 2000 Conference, Lyon, França. [Acessado Agosto-2011]. Disponível em: <http://eric.univ-lyon2.fr/ pkdd2000/>.

GüING, H.; ALMEIDA, T. de; DING, Z. Modeling and querying moving objects in networks. *The VLDB Journal*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, v. 15, p. 165–190, June 2006. ISSN 1066-8888. Disponível em: <http://dx.doi.org/10.1007/s00778-005-0152-x>.

GONçALVES, E. C. *Data Mining de Regras de Associação.* 2008. Software Livre -DevMedia. [Acessado Novembro-2011]. Disponível em: <http://www.devmedia.com.br/articles/viewcomp.asp?comp=7065>.

GONçALVES, E. C. Data mining com a ferramenta weka. *III FSLDC - Fórum de Software Livre de Duque de Caxias.*, v. 3, p. 60, 2011.

GüTING, R. H. An introduction to spatial database systems. *The International Journal on Very Large Data Bases*, v. 3, n. 4, p. 357–399, October 1994.

GUO, D. *VIS-STAMP - Visualization System for Space-Time and Multivariate Patterns.* 2009. [Acessado Março-2011]. Disponível em: <http://www.SpatialDataMining.org>.

GUO, D.; LIU, S.; JIN, H. A graph-based approach to vehicle trajectory analysis. *J. Locat. Based Serv.*, Taylor & Francis, Inc., Bristol, PA, USA, v. 4, p. 183–199, September 2010. ISSN 1748-9725. Disponível em: <http://dx.doi.org/10.1080/17489725.2010.537449>.

HAN, J.; KAMBER, M. *Data Mining:Concepts and Techniques.* [S.l.]: Morgan Kaufmann, 2001.

HäRRI, J.; BONNET, C.; FILALI, F. Kinetic graphs: a framework for capturing the dynamics of mobile structures in manet. In: *Proceedings of the 2nd ACM workshop on Performance monitoring and measurement of heterogeneous wireless and wired networks.* New York, NY, USA: ACM, 2007. (PM2HW2N '07), p. 88–91. ISBN 978-1-59593-805-3. Disponível em: <http://doi.acm.org/10.1145/1298275.1298294>.

JHON, G. H. Robust decision trees: Removing outliers from databases. In: *Proceedings of the 1st International Conference on Knowledge Discovery and Data Mining.* [S.l.]: AAAI Press, 1995. p. 174–179.

KANG, I.-S.; KIM, T.-w.; LI, K.-J. A spatial data mining method by delaunay triangulation. In: *Proceedings of the 5th ACM international workshop on Advances in geographic information systems.* New York, NY, USA: ACM, 1997. (GIS '97), p. 35–39. ISBN 1-58113-017-1. Disponível em: <http://doi.acm.org/10.1145/267825.267836>.

LEE, J.-G. et al. Traclass: trajectory classification using hierarchical region-based and trajectory-based clustering. *Proc. VLDB Endow.*, VLDB Endowment, v. 1, p. 1081–1094, August 2008. ISSN 2150-8097. Disponível em: <http://dx.doi.org/10.1145/1453856.1453972>.

MONREALE, A. et al. Wherenext: a location predictor on trajectory pattern mining. In: *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2009. (KDD '09), p. 637–646. ISBN 978-1-60558-495-9. Disponível em: <http://doi.acm.org/10.1145/1557019.1557091>.

NONG, Y. *The HandBook of Data Mining*. New Jersey, Arizona State University: Lawrence Erlbaum Associates, inc., 2003.

OLIVEIRA, A. L. T.; LIMA, D. R. Analise espacial do perfil dos alunos do ifpi campus picos. In: *Sistemas e Tecnologias de Informação*. Chaves,Portugal: [s.n.], 2011. v. 2, p. 368–373. ISBN 978-989-96247-5-7. Disponível em: <http://www.aisti.eu/cisti2011>.

OLIVEIRA, J.; DUTRA, L. V.; RENNó, C. Aplicação de métodos de extração e seleção de atributos para classificação de regiões. *Anais XII Simpósio Brasileiro de sensoriamento Remoto*, Goiânia - Brasil, v. 1, p. 16–21, Abril 2005. Disponível em: <http://marte.dpi.inpe.br/col/ltid.inpe.br/sbsr/2004/11.20.21.24/doc/4201.pdf>.

OLIVEIRA, P.; RODRIGUES, F.; HENRIQUES, P. R. Limpeza de dados: Uma visão geral. In: BELO, O.; LOURENçO, A.; ALVES, R. (Ed.). *Data Gadgets 2004 Int. Workshop - Bringing Up Emerging Solutions for Data Warehousing Systems*. Málaga, Espanha: [s.n.], 2004. p. 39–51.

ONG, R. et al. From pattern discovery to pattern interpretation in movement data. *2010 IEEE International Conference on Data Mining Workshops*, Sydney, Australia, p. 527–534, Dezembro 2010.

PANAGIOTAKIS, C.; PELEKIS, N.; KOPANAKIS, I. Trajectory voting and classification based on spatiotemporal similarity in moving object databases. In: *Proceedings of the 8th International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis VIII*. Berlin, Heidelberg: Springer-Verlag, 2009. (IDA '09), p. 131–142. ISBN 978-3-642-03914-0. Disponível em: <http://dx.doi.org/10.1007/978-3-642-03915-7_12>.

PEI, J. et al. Mining sequential patterns by pattern-growth: The prefixspan approach. *IEEE Trans. on Knowl. and Data Eng.*, IEEE Educational Activities Department, Piscataway, NJ, USA, v. 16, p. 1424–1440, November 2004. ISSN 1041-4347. Disponível em: <http://dx.doi.org/10.1109/TKDE.2004.77>.

PRASS, F. S. Kdd: Processo de descoberta de conhecimento em bancos de dados. *Grupo de Interesse Em Engenharia de Software, Florianópolis*, v. 1, p. 10–14, 2004.

RAHM, E.; DO, H. H. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, v. 23, p. 2000, 2000.

READ, R. C.; CORNEIL, D. G. The graph isomorphism disease. *Journal of Graph Theory*, Wiley Subscription Services, Inc., A Wiley Company, v. 1, n. 4, p. 339–363, 1977. ISSN 1097-0118. Disponível em: <http://dx.doi.org/10.1002/jgt.3190010410>.

SANTOS, I. J. P. dos; ALVARES, L. O. Tracts: Um método para a classificação de trajetórias de objetos móveis usando séries temporais. *CSBC - Congresso da Sociedade Brasileira de Computação*, Natal, RN, Brasil, p. 183–199, Julho 2011.

SANTOS, R. *Mineração e Visualização de Dados usando Java*. 2010. [Acessado Setembro-2011]. Disponível em: <http://www.lac.inpe.br/ rafael.santos/dmapresentacoes.jsp>.

SILVA, M. P. Mineração de dados - conceitos, aplicações e experimentos com weka. *Escola Regional de Informática RJ/ES*, Sociedade Brasileira de Computação, v. 1, p. 19–21, Novembro 2004.

SOUZA, C. R. B. de et al. Sometimes you need to see through walls: a field study of application programming interfaces. In: *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. New York, NY, USA: ACM, 2004. (CSCW '04), p. 63–71. ISBN 1-58113-810-5. Disponível em: <http://doi.acm.org/10.1145/1031607.1031620>.

SRIKANT, R.; AGRAWAL, R. Mining Sequential Patterns: Generalizations and Performance Improvements. In: *EDBT '96: Proceedings of the 5th International Conference on Extending Database Technology*. London, UK: Springer-Verlag, 1996. p. 3–17. ISBN 3-540-61057-X.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *"Introduction to Data Mining"*. [S.l.]: Boston: Addison-Wesley, 2006.

TEAM, J. D. *Jung - Java Universal Network/Graph Framework*. 2012. [Acessado Outubro-2011]. Disponível em: <http://jung.sourceforge.net/>.

TEAM, R. D. *RapidMiner*. 2008. [Acessado Janeiro-2011]. Disponível em: <http://www.rapidminer.com>.

WITTEN, I. H.; FRANK, E. *Weka - the waikato environment for knowledge analysis.* 2008. [Acessado Setembro-2010]. Disponível em: <http://www.cs.waikato.ac.nz/ml/weka/>.

WOLFSON, O. et al. Moving objects databases: issues and solutions. In: *Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on*. [S.l.: s.n.], 1998. p. 111 –122. ISSN 1099-3371.

YI, B. *Um modelo de dados para objetos móveis.* Dissertação (Mestrado) — Universidade Estadual de Campinas, Campinas, SP, 2004.

ZHAO, Q.; BHOWMICK, S. S. Sequential pattern mining: A survey. *ITechnical Report CAIS Nayang Technological University Singapore*, Citeseer, p. 1–26, 2003.

ZHENG, Y. et al. Learning transportation mode from raw gps data for geographic applications on the web. In: *Proceeding of the 17th international conference on World Wide Web*. New York, NY, USA: ACM, 2008. (WWW '08), p. 247–256. ISBN 978-1-60558-085-2. Disponível em: <http://doi.acm.org/10.1145/1367497.1367532>.