UNIVERSIDADE ESTADUAL PAULISTA

“Júlio de Mesquita Filho”

Pós-Graduação em Ciência da Computação

Daniel da Silva Gomes Lima

Extração de Conhecimento em Trajetórias Semânticas

UNESP

2017

Daniel da Silva Gomes Lima

Extração de Conhecimento em Trajetórias Semânticas

Orientador: Prof. Dr. Ivan Rizzo Guilherme

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Ciência da Computação - Área de Concentração em Matemática e Inteligência Computacional, como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

UNESP

2017

Lima, Daniel da Silva Gomes.
Extração de conhecimento em trajetórias semânticas / Daniel da Silva Gomes Lima. -- São José do Rio Preto, 2017
97 f. : il.

Orientador: Ivan Rizzo Guilherme
Dissertação (mestrado) – Universidade Estadual Paulista “Júlio de Mesquita Filho”, Instituto de Biociências, Letras e Ciências Exatas

1. Ciência da computação. 2. Aprendizado do computador. 3. Mineração de dados (computação) 4. Movimento - Estudo. 5. Trajetória semântica. I. Universidade Estadual Paulista "Júlio de Mesquita Filho". Instituto de Biociências, Letras e Ciências Exatas. II. Título.

CDU – 518.72

Daniel da Silva Gomes Lima

Extração de Conhecimento em Trajetórias Semânticas

Dissertação de Mestrado elaborada junto ao Programa de Pós-Graduação em Ciência da Computação - Área de Concentração em Matemática e Inteligência Computacional, como parte dos requisitos para obtenção do título de Mestre em Ciência da Computação.

Comissão Examinadora

Prof. Dr. Ivan Rizzo Guilherme
UNESP – Rio Claro
Orientador

Prof. Dr. Daniel Carlos Guimarães Pedronette
UNESP – Rio Claro

Prof. Dr. Evandro Eduardo Seron Ruiz
USP – Ribeirão Preto

UNESP

2017

Agradecimentos

Agradeço a minha esposa, Juliana Cristina Lopes Lima pela paciência, compreensão e incentivo irrestrito durante todo o período dedicado a esse trabalho. Sem você ao meu lado essa tarefa seria muito mais complicada.

Agradeço aos meus pais, Paulo Gomes Lima e Sônia Maria Souza da Silva Lima, e irmãos, Eduardo da Silva Gomes Lima e Paula Souza da Silva Gomes Lima, que sempre me apoiaram e torceram pelas minhas conquistas.

Agradeço ao meu orientador, Ivan Rizzo Guilherme, pela oportunidade que me foi dada, reconhecendo minhas limitações e me guiando com sabedoria pelos caminhos do mundo acadêmico.

Também gostaria de agradecer aos professores que tive contato durante todo o programa, seja através de disciplinas ou encontros ocasionais. Em especial para Verônica Oliveira de Carvalho e Renato Fileto que participaram da qualificação contribuindo para o amadurecimento e direcionamento do trabalho.

Por fim, gostaria de agradecer a todos os colegas que tive o prazer de conviver durante o programa. Em especial a Rodrigo Cesar Antonialli e Filipe Marcel Fernandes Gonçalves pelas boas discussões e sugestões de alternativas para solucionar obstáculos encontrados.

Resumo

O rápido avanço do uso de tecnologias que permitem a coleta dos dados de movimentação de indivíduos gerou como consequência um crescimento no volume de dados de trajetórias. Trabalhos que utilizam esses dados tem como principal objetivo a realização de análises para identificação de padrões que permitam explicar o comportamento do indivíduo durante o movimento. Para que a extração de conhecimento dos dados de trajetória possa ser gerada de uma forma que agregue conteúdo relevante, é necessário que exista um modelo de representação de trajetória que contemple as principais características do movimento no contexto de aplicação utilizado, além de um processo que transforme os dados brutos de trajetória na estrutura definida por esse modelo. Com isso é possível aplicar técnicas e algoritmos para exploração desses dados e geração de conhecimento. As técnicas de aprendizado de máquina em conjunto com a área de representação de conhecimento fornecem a base conceitual para que problemas desse tipo possam ser modelados e soluções possam ser desenvolvidas a fim de solucionar esses problemas. Este trabalho utiliza esses aspectos conceituais para apresentar uma proposta que permite a extração de conhecimento de trajetória. Esse conhecimento é obtido através da identificação dos locais de parada mais relevantes para um indivíduo e do movimento sequencial entres esses locais determinando o conjunto de comportamentos que representam o padrão de movimento do indivíduo em um período.

Palavras-chave: Trajetória Semântica, Aprendizado de Máquina, Mineração de Dados.

Abstract

The rapid advance of the use of technologies that allow the collection of the data of movement of individuals generate as a consequence an increase in the volume of trajectory data. Works that use this data have as main objective the accomplishment of analyzes for identification of patterns that allow to explain the behavior of the individual during the movement. In order to the knowledge extraction of the trajectory data to be generated in a way that aggregates relevant content, there must be a trajectory representation model that considers the main characteristics of the movement in the context of the application used, besides a process that transforms the raw trajectory data in the structure defined by this model. With this it is possible to apply techniques and algorithms for exploration of this data and generation of knowledge. Machine learning techniques in conjunction with the area of knowledge representation provide the conceptual basis for problems of this type to be modeled and solutions can be developed in order to solve these problems. This work uses these conceptual aspects to present a proposal that allows the extraction of knowledge of trajectory. This knowledge is obtained through the identification of the most relevant stop places for an individual and the sequential movement between these places determining the set of behaviors that represent the individual's movement pattern in a period.

Keywords: Semantic Trajectory, Machine Learning, Data Mining.

Lista de Figuras

Figura 1: Representação da Diferença entre Trajetória e Movimento.
Figura 2: Hierarquia de Algoritmos de Aprendizado de Máquina.
Figura 3: Processo de descoberta do conhecimento.
Figura 4: Fluxo de dados de trajetória aplicado na localização de veículos.
Figura 5: Estrutura modular de equipamentos para coleta de dados de trajetória.
Figura 6: Modelo de Paradas e Movimentos.
Figura 7: Modelo CONSTAnT de Trajetória Semântica.
Figura 8: Proposta de Ontologia para Trajetória Semânticas.
Figura 9: Ontologia Baquara.
Figura 10: Etapas enriquecimento semântico de trajetória.
Figura 11: Framework SeMiTri.
Figura 12: Framework Baquara.
Figura 13: Padrão Geométrico de Trajetória.
Figura 14: Padrão Semântico de Trajetória.
Figura 15: Visão geral do processo de descoberta de comportamentos.
Figura 16: Etapa de Agrupamento de Paradas.
Figura 17: Etapa de Sequenciamento de Grupos.
Figura 18: Exemplo de arquivos entrada e saída DBSCan.
Figura 19: Mapa de Paradas e centróide do Grupo.
Figura 20: Elementos retornados pesquisa Overpass.
Figura 21: Mapa elementos retornado Overpass e centróide do Grupo.
Figura 22: Mapa de raio de amplitude Grupo para identificação de Locais.
Figura 23: Exemplo de arquivos entrada e saída CM-SPADE.
Figura 24: Estudo de Caso - Fases.
Figura 25: Gráfico Quantidade de Comportamentos Extraídos.
Figura 26: Gráfico Consistência entre Fases.
Figura 27: Índice de Jaccard.
Figura 28: Gráfico Consistência dos Comportamentos.
Figura 29: Gráfico Consistência - Ranking 10 elementos.

Lista de Siglas

API - Application Programming Interface

BES - Begin End Stop

CEP - Código de Endereçamento Postal

CONSTAnT - CONceptual model of Semantic TrAjecTories

DBScan - Density Based Spatial Clustering of Application with Noise

GPRS - General Packet Radio Service

GPS - Global Positioning System

HTTP – Hypertext Transfer Protocol

IGN_OFF - Ignição Desligada

IGN_ON - Ignição Ligada

IP - Internet Protocol

JSON - JavaScript Object Notation

LBS - Location Based System

OSM - OpenStreetMap

POI - Point of Interest

SI - Sem Identificação

TCP - Transmission Control Protocol

VPN - Virtual Private Network

Sumário

Resumo

Abstract

1 Introdução 10

1.1 Contexto e Motivação 11

1.2 Objetivos 12

1.3 Organização do Trabalho 12

2 Aspectos Conceituais 14

2.1 Trajetória 14

2.2 Aprendizado de Máquina 16

2.3 Representação do Conhecimento 19

2.4 Contextualização do Problema 20

3 Trabalhos Relacionados 28

3.1 Modelos de Representação 28

3.2 Enriquecimento Semântico 34

3.3 Mineração de dados de trajetória 39

3.4 Conclusão do Capítulo 45

4 Extração de Conhecimento em Trajetórias 46

4.1 Visão Geral do Processo 46

4.2 Agrupamento de Paradas 47

4.3 Sequenciamento de Grupos 52

4.4 Conclusão do Capítulo 56

5 Implementação do Processo 57

5.1 Implementação 57

5.2 Conjunto de Dados 67

5.3 Metodologia 69

5.4 Estudo de Caso 72

5.5 Conclusão do Capítulo 82

6 Considerações Finais 84

Referências 87

Apêndice A 93

Apêndice B 95

Apêndice C 97

1 Introdução

Nos últimos anos houve um aumento na utilização de dispositivos que possuem tecnologias de posicionamento e comunicação para coleta e transmissão de informações sobre a localização de objetos em movimento. Essa expansão ocorreu em virtude dos seguintes aspectos: o barateamento dessas tecnologias; interesse de organizações na utilização desse conteúdo para fins informativo ou estratégico; e, popularização de ferramentas para visualização desse tipo de conteúdo no formato gráfico, sobretudo através de mapas, o que facilitou a compreensão sobre o comportamento desses objetos.

O conjunto de dados relativos ao deslocamento de um objeto em movimento descreve a trajetória desse objeto. Análises que utilizam esse tipo de conteúdo no formato bruto, geralmente composto por dados temporais e espaciais do deslocamento, geram resultados pobres no sentido semântico pois possuem poucas informações sobre o indivíduo, o meio e o contexto de aplicação.

Para realizar análises mais ricas no conteúdo de trajetória é necessário a adição de outros tipos de informações, específicas ao domínio de aplicação em um processo denominado de enriquecimento semântico. Trajetórias que são estruturadas em um modelo de representação que contemple esse conjunto adicional de conteúdo são denominadas de trajetórias semânticas.

Análises que utilizam o conteúdo de trajetórias semânticas facilitam o entendimento sobre o movimento do indivíduo mas esbarram em uma série de imprecisões que dificultam o enriquecimento semântico e geram incertezas nesse conteúdo. As principais dificuldades são relativas a aspectos técnicos do dispositivo utilizado e operacionais existentes durante a fase de coleta. Outras incertezas ocorrem em função do contexto da aplicação e das fontes utilizadas para o enriquecimento semântico.

O estudo de trajetórias semânticas engloba: (i) abordagens para representação desse tipo de conhecimento; (ii) mecanismos para realização do processo de enriquecimento semântico; (iii) e soluções para descoberta do conhecimento existente nesse conteúdo.

Trajetórias semânticas podem ser utilizadas em soluções de vários domínios como: análise do comportamento de pessoas; identificação de padrões de comportamento de animais ou grupos de animais em trajetos migratórios; soluções na área de logística e transportes; e, outras soluções que visem explorar o aspecto da localização desses indivíduos em movimento.

A extração de conhecimento sobre o movimento dos indivíduos é de grande interesse para as organizações. Esse conhecimento pode ser expresso na forma de padrões de comportamento que são gerados baseado nas características espaciais, temporais e semânticas dos locais no qual o movimento é realizado. A adição de conteúdos semânticos aos dados de trajetória permite que conhecimentos mais ricos sobre o comportamento possam ser obtidos.

Dessa forma este trabalho apresenta um processo que permite a extração de conhecimentos de trajetórias. Esse processo é composto por duas etapas que possibilita a identificação dos locais de parada mais relevantes para o indivíduo e dos principais padrões de movimento entre esses locais. Como resultado esse processo apresenta um conjunto com os principais comportamentos dos indivíduos que permite o entendimento sobre as principais características relativas ao movimento do indivíduo.

1.1 Contexto e Motivação

As técnicas convencionais que utilizam os dados de trajetória consideram apenas características temporais e espaciais do posicionamento do objeto em movimento. Análises mais complexas exigem um grande esforço computacional devido a natureza bruta desses dados. O grande volume de dados, gerado pela popularização dos equipamentos e aplicações que utilizam esses equipamentos, dificulta a exploração desse conteúdo e dessa forma as análises realizadas são apenas superficiais.

O conhecimento implícito nesse conteúdo acontece, sobretudo, na forma de padrões ocasionados por repetições de movimentos comuns aos hábitos dos indivíduos que podem ser uma pessoa, um animal, um veículo ou outro objeto. O indivíduo tem a tendência de repetir hábitos em uma escala de tempo diária, semanal, mensal, semestral e assim por diante. Esse hábito pode ser observado no deslocamento de uma pessoa no trajeto de casa para o trabalho, na migração de um pássaro de um continente para outro com o intuito de se reproduzir e até mesmo no roteiro de viagem realizado por um veículo na execução de um itinerário.

A identificação do conjunto de características que representam o padrão de comportamento do movimento do indivíduo é fundamental para que instituições, públicas ou privadas, possam desenvolver serviços inteligentes. Para descoberta desses comportamentos relativos ao movimento é necessário que as soluções lidem com uma série de imprecisões

comuns da operação e contexto utilizado e também de erros ocasionados da coleta desses dados.

Nesse sentido é necessário desenvolver uma solução que permita a extração de conhecimentos relativos aos comportamentos dos indivíduos utilizando um conjunto de trajetórias. Esse conhecimento, que representa as principais características do movimento realizado, deve ser apresentado em um formato significativo de conteúdo de forma a facilitar o entendimento sobre a movimentação.

1.2 Objetivos

O principal objetivo deste trabalho é apresentar um processo que permite a extração de conhecimentos de trajetórias possibilitando a identificação dos principais comportamentos que definem as características sobre o movimento de indivíduos. Também faz parte do objetivo a apresentação desse conhecimento em um formato mais significativo permitindo o entendimento sobre os aspectos do movimento realizado. Por fim, faz parte do objetivo a avaliação dos comportamentos identificados analisando se existe uma consistência na manutenção dos principais comportamentos que definem o perfil de um indivíduo ao longo do tempo.

1.3 Organização do Trabalho

Inicialmente no Capítulo 2 são abordados temas para fundamentação teórica requeridas para o desenvolvimento deste trabalho como trajetória, aprendizado de máquina e representação do conhecimento. Neste capítulo também é apresentado uma contextualização do problema evidenciando as principais características operacionais comuns em aplicações que utilizam dados de trajetórias. No Capítulo 3 é apresentado um levantamento de trabalhos do estado da arte relacionados a enriquecimento semântico e mineração de dados de trajetória. O Capítulo 4 apresenta o processo que permite a extração de comportamentos de trajetórias. No Capítulo 5 é apresentada a implementação do processo proposto com o objetivo de avaliar os principais aspectos conceituais aplicados no contexto de trajetórias de veículos. Neste capítulo

também são apresentados resultados de análises realizadas para avaliação da implementação e do processo proposto. Por fim, no Capítulo 6 são apresentadas as considerações finais deste trabalho.

2 Aspectos Conceituais

Neste capitulo são apresentados e discutidos os principais conceitos necessários para o entendimento da proposta para extração de conhecimento de trajetórias semânticas. Inicialmente serão apresentadas as definições de trajetória e trajetória semântica. Em seguida será apresentado uma fundamentação referente a descoberta de conhecimento e aprendizado de máquina. Na sequência será abordado fundamentos sobre representação de conhecimento. Por fim, será apresentada uma visão contextual de aplicações que utilizam dados de trajetória.

2.1 Trajetória

O trabalho de Spaccapietra et al. (2008) apresenta uma visão conceitual de trajetórias e define trajetória como um registro da evolução do posicionamento de um objeto que está se movendo no espaço durante um intervalo de tempo a fim de alcançar um determinado objetivo.

Os dados do posicionamento de um objeto possuem características espaciais e temporais. As características espaciais são representadas no formato de coordenada geográfica (x,y). A característica temporal é representada por um instante de tempo (t). Trajetória compostas exclusivamente por essas características espaciais e temporais são denominadas de trajetórias brutas.

O trabalho de Bogorny et al. (2013) faz a representação dessas características no formato de uma tupla (*x*,*y*,*t*), denominada de ponto, sendo *x* e *y* as coordenadas espaciais que representam o espaço e *t* o horário no qual o ponto foi coletado. Assim, uma trajetória é um conjunto de pontos (p_1, p_2, p_3….p_n) onde os elementos $p_i = (x_i, y_i, t_i)$ estão ordenados de forma temporal ascendente, ou seja, $t_1 < t_2 < t_3 < \ldots < t_n$.

Segundo Parent et al. (2013), o conceito de trajetória iniciou com a capacidade de coletar o movimento de um objeto que se desloca em um espaço geográfico durante um período de tempo definido. No entanto, somente pode ser considerado como trajetória a parte do movimento do objeto, delimitada por um ponto inicial e um ponto final.

Assim, uma trajetória é definida por dois pontos específicos no tempo e espaço de movimento do objeto. Esses pontos, denominados de Início e Fim da trajetória, representam respectivamente a primeira e a última posição do objeto nessa trajetória (SPACCAPIETRA et

al., 2008). Portanto trajetória é a parte do deslocamento de um indivíduo - pessoa, animal ou objeto - que é relevante ao contexto de aplicação ao qual esse indivíduo está contido.

A Figura 1 faz a representação de trajetória considerando somente a parte do movimento do indivíduo que é relevante.

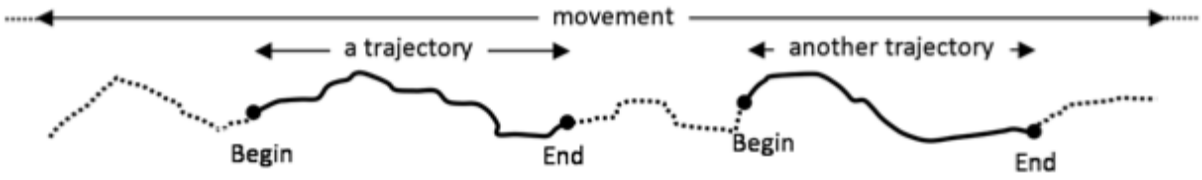

Figura 1: Representação da Diferença entre Trajetória e Movimento.
Fonte: Parent, Spaccapietra, Renso, Andrienko, Andrienko, Bogorny, Damiani, Gkoulalas-divanis, Macedo, Pelekis, Theodoridis e Yan (2013).

Na Figura 1 é possível observar que duas trajetórias são relevantes ao longo do movimento de um indivíduo. Cada trajetória tem um ponto de início (*Begin*) e um ponto de fim (*End*) identificados e que delimitam o espaço e o tempo de trajetória. Também é possível observar que existem pontos do movimento do indivíduo que não são relevantes e não fazem parte de nenhuma trajetória.

Alguns trabalhos como em Bogorny et al. (2013) representam trajetória como sendo um conjunto de sub trajetórias. Sendo a trajetória uma representação mais global do deslocamento do indivíduo ao longo de um período e uma sub trajetória um deslocamento específico realizado em parte do período. Uma sub trajetória possui as mesmas características de uma trajetória, sendo um segmento da trajetória que contém um conjunto de pontos ordenados temporalmente de forma ascendente. Dessa forma, conceitualmente pode-se dizer que uma sub trajetória também é uma trajetória.

A razão de subdividir a trajetória em segmentos é devido a características específicas do objeto, do meio utilizado para deslocamento e do ambiente ao qual esse se movimenta que possam ser utilizados como análise no contexto da aplicação.

Análises que utilizam trajetórias brutas, no geral, buscam encontrar padrões geométricos nesses dados como agrupamento de pontos ou sequência entre pontos. No entanto,

para que análises mais ricas, em conteúdo, sejam possíveis é necessário que algumas características adicionais sejam adicionadas a esses dados brutos.

O processo de adicionar conhecimento aos dados brutos de trajetória é denominado de enriquecimento semântico e será abordado no Capítulo 3. Trajetórias que possuem características adicionais sobre o indivíduo, meio ou outros dados relevantes para a aplicação, em conjunto com as características espaciais e temporais, obtidas do conteúdo bruto, são denominadas de trajetórias semânticas.

A finalidade de utilizar trajetórias semânticas é adicionar mais significado ao dado bruto em termos de conhecimentos de aplicação e contexto com o objetivo de obter resultados mais significativos nas análises realizadas sobre essas trajetórias (BOGORNY et al., 2013).

2.2 Aprendizado de Máquina

Atualmente o grande volume de informação gerada nas mais diversas áreas é impossível de ser processado por humano requerendo o uso de métodos computacionais dentre os quais o aprendizado de máquina. Aprendizado de máquina é um campo da inteligência artificial destinado ao desenvolvimento de algoritmos e técnicas que permitam ao computador adquirir conhecimento de forma automática sobre um determinado domínio.

Bishop (2007) define aprendizado de máquina como a área da Inteligência Artificial cujo objetivo é o desenvolvimento de técnicas computacionais inspiradas no processo de aprendizado humano.

Os métodos de aprendizado de máquina são baseados no uso das experiências acumuladas que podem ser utilizadas opcionalmente em conjunto com conhecimento de um especialista e permitem ao computador aperfeiçoar o entendimento e execução de determinada tarefa. Conforme exibido na Figura 2, as técnicas de aprendizado de máquina estão subdivididas em três classes de acordo com as experiências utilizadas:

- Aprendizado Supervisionado - Algoritmos desse tipo utilizam um conjunto de dados contendo as experiências com o valor da classe de saída rotulada. Uma parte do conjunto de dados é utilizado no treinamento para o ajuste e a calibração de atributos de funções. Uma outra parte são utilizadas

posteriormente para testes. Algoritmos desse tipo são utilizados para problemas de classificação e regressão de dados.

- Aprendizado Não Supervisionado - Algoritmos dessa categoria utilizam um conjunto de dados sem o valor da classe de saída rotulada. O conjunto dos dados é utilizado para determinar os padrões de dados e como os dados estão organizados (MITCHELL, 1997). Algoritmos desse tipo são utilizados para fazer agrupamento de dados (*clustering*), também sendo possível realizar associação e sumarização dos dados.
- Aprendizado Semi-Supervisionado - São formados por algoritmos que utilizam um conjunto de dados parcialmente rotulados durante o treinamento. Em virtude disso, essa abordagem é uma combinação entre as técnicas supervisionada e não supervisionada. São de grande utilidade em problemas na qual apenas parte do conjunto de treinamento está rotulado.

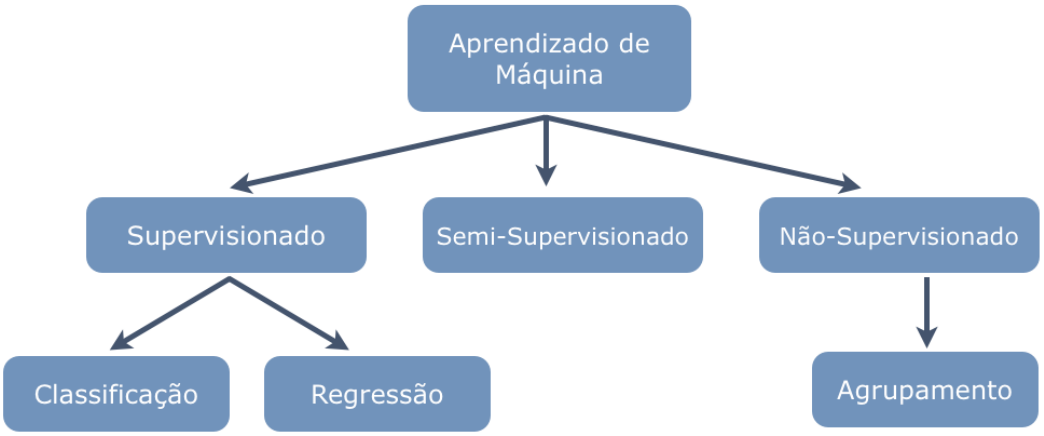

Figura 2: Hierarquia de Algoritmos de Aprendizado de Máquina.

O processo de descoberta de conhecimento é composto por etapas iterativas e sequenciais conforme exibido na Figura 3.

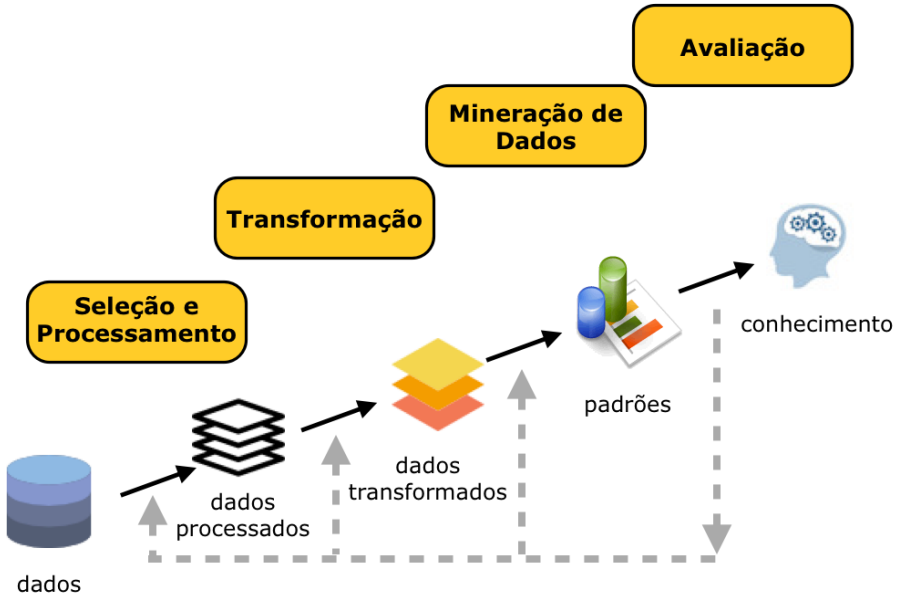

Figura 3: Processo de descoberta do conhecimento.

Fonte: Adaptado de Fayyad, Piatetsky-shapiro, Smyth e Uthurusamy (1996).

O processo de descoberta de conhecimento a partir de um conjunto de dados é composto por etapas iterativas e sequenciais conforme exibido na Figura 3. Em uma primeira etapa elementos do conjunto de dados são selecionados e processados. Em seguida o conjunto de dados processados na fase anterior são transformados em estruturas que facilitam a etapa de mineração.

A etapa de mineração de dados é formada por um conjunto de ferramentas e técnicas de aprendizado de máquina que permite a extração do conhecimento. Esse conhecimento, na forma de padrão, pode ser exibido no formato de agrupamentos, árvores de decisões e regras. Por fim, na última etapa é feita uma avaliação do conhecimento extraído. Essa análise pode ser utilizada para realimentar as etapas anteriores para que novas hipóteses possam ser testadas. O resultado esperado do processo de descoberta de conhecimento é o conhecimento estruturado e um conjunto de informações relevantes que serão utilizadas pelos tomadores de decisão.

O processo de extração de conhecimento tem sido aplicado a dados de trajetórias. Em razão do grande volume dados coletados e incerteza dos dados coletados têm requerido o desenvolvimento de abordagens para as fases de Seleção e Processamento e Transformação dos dados.

A aplicação de técnicas de aprendizado de máquina com dados relativos a trajetória é de grande interesse e utilidade para diversos tipos de aplicações. Algoritmos que fazem a mineração de dados de trajetória, permitindo a descoberta de padrões baseado não apenas na proximidade espacial entre os dados, mas também considerando a similaridade semântica entre os elementos, estão entre as áreas de maiores interesses nos estudos relacionados a trajetória semântica (BOGORNY et al., 2013).

2.3 Representação do Conhecimento

Na Inteligência Artificial a área denominada de representação do conhecimento, estabelece as formas de representar o conhecimento que possa ser entendido não apenas por humanos, mas também por máquinas. Assim, um formalismo para a representação do conhecimento é um conjunto de definições sintáticas e semânticas para permitir descrever os objetos e suas características em determinado contexto. Os principais formalismos de representação do conhecimento são: Lógica de Primeira Ordem, Regras, Quadros (Frames) e Redes Semânticas.

Uma metodologia para a representação do conhecimento deixa explícitos quais são os objetos, as características e relações que descrevem esses objetos e também as restrições específicas ao contexto. Para realizar a representação do conhecimento é necessário definir as características relevantes ao escopo ao qual o sistema engloba e também o nível de detalhamento, ou granularidade, de como esse conhecimento será representado. Essa representação deve ser de fácil entendimento, disponibilidade e utilização, permitindo ser utilizável por pessoas ou máquinas.

O fator determinante na representação do conhecimento é a forma como esse conhecimento será utilizado. A utilização abrange desde o formato como esse conhecimento será adquirido, estruturado e representado; passando pela etapa de consulta e recuperação desse conhecimento, até chegar na etapa de raciocínio que engloba técnicas de generalização, agrupamento, analogia entre outros.

Um importante aspecto relacionado a representação do conhecimento é o reuso e o compartilhamento do conhecimento representado. Neste contexto, as ontologias foram propostas. De acordo com Gruber (1993), uma ontologia é especificação formal e explícita de

uma conceitualização compartilhada. Nessa definição, uma "conceitualização" se refere a um modelo abstrato de um fenômeno e dos conceitos relevantes desse fenômeno, "especificação formal" se refere ao fato do conceito ser compreensível por todos, "explícita" pelo fato dos elementos e restrições estarem claramente definidos e, "compartilhada" pelo fato desse conhecimento ser aceito de forma consensual.

As ontologias têm sido utilizadas no desenvolvimento de trabalhos em várias linhas de pesquisa como: inteligência artificial, web semântica, engenharia de software, entre outras. Dessa forma, é natural que exista uma grande variação na definição desse conceito (BREITMAN, 2005).

Conforme Lacy (2005), uma ontologia permite transcrever um entendimento comum de um determinado domínio através da declaração da descrição semântica dos termos de uma forma explícita, e de forma expressiva, além de suportar o compartilhamento de informações. A declaração semântica dos termos de forma explícita, ao ser transposta para os conteúdos que estavam codificados em sistemas ou banco de dados, possibilita deixar visível e compreensível para outras partes interessadas, reduzindo a ambiguidade de conceitos.

2.4 Contextualização do Problema

A utilização de equipamentos que permitam a coleta de dados sobre a movimentação de pessoas, animais e objetos aumentou nos últimos anos com a popularização e barateamento dessas tecnologias. Análises sobre o significado desses movimentos são complexas devido: ao grande volume de dados; imprecisões ocasionadas por questões técnicas referentes ao tipo de equipamento utilizado; e, as especificidades do contexto de aplicação. Imprecisões técnicas dos equipamentos são causadas, sobretudo, devido a limitações energéticas e de comunicação dos dispositivos. Imprecisões de aplicação são oriundas do desconhecimento espacial por onde o indivíduo se movimenta, sobre a finalidade dos movimentos e também dos locais de paradas. Assim, os trabalhos que se propõem a analisar esse conteúdo precisam levar em consideração esse conjunto de imprecisões a fim de gerar resultados condizentes com a realidade do movimento.

A Figura 4 apresenta o exemplo de um cenário de análise de trajetória de veículos composto de duas fases: a coleta de dados e análise de dados. A fase de coleta de dados é feita

com a utilização de equipamentos para a coleta de dados sobre o movimento de veículos em um cenário comum de utilização para a identificação da localização e trajetória de veículos. Nesse tipo de contexto o equipamento é instalado no veículo conectado a bateria principal do mesmo e obtém as informações sobre o posicionamento utilizando um GPS (*Global Positioning System*) que através dos satélites em órbita permite a identificação da posição do veículo. Os dados são enviados via GPRS (*General Packet Radio Service*) utilizando uma rede de antenas de operadoras de celulares. Esses dados são transmitidos até um endereço IP definido no firmware interno dos equipamentos. O firmware é um programa de baixo nível embarcado no equipamento, responsável por toda a lógica interna do dispositivo. Seu princípio é similar a um sistema operacional e permite fazer o gerenciamento entre os diversos módulos que compreendem o hardware do equipamento. Os dados são enviados para um servidor que possui um programa denominado de *gateway* responsável pelo processamento e validação dessas mensagens. Após validação as mensagens são armazenadas em um repositório. A fase de análise consiste na definição de métodos para a consulta dos dados disponíveis no repositório que serão utilizados por aplicações que variam de acordo com o interesse do usuário nesse tipo de conteúdo possibilitando que o dado seja visualizado através de mapas, relatórios, gráficos e demais conteúdos que permitam análises e estatísticas.

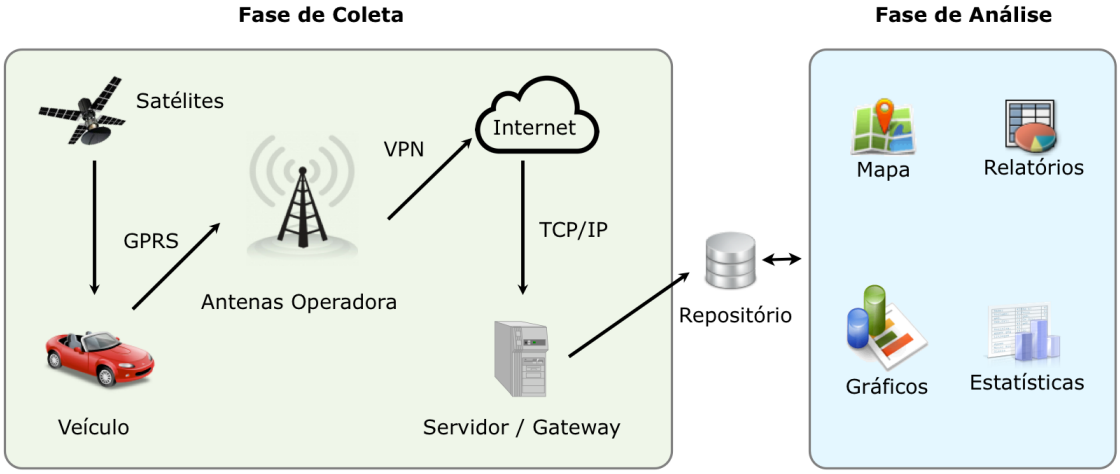

Figura 4: Fluxo de dados de trajetória aplicado na localização de veículos.

A Figura 4 retrata o fluxo de dados na forma convencional partindo do indivíduo analisado até o destino onde será consumido por diversas aplicações. Porém esse fluxo também pode funcionar de forma inversa. Nesse caso, um comando ou configuração podem ser enviados para o equipamento. A diferença nesse caso é que o *gateway* fica responsável por

formatar a mensagem de acordo com o protocolo aceito pelo equipamento e também por fazer o envio do conteúdo para o equipamento correto.

A Figura 5 exibe uma representação dos principais módulos que compõem um equipamento utilizado para coleta de dados do movimento de indivíduos. Esses equipamentos, de uma forma geral, são compostos por 6 módulos principais: posicionamento, comunicação, alimentação, armazenamento, processamento e o módulo de entrada e saída.

Figura 5: Estrutura modular de equipamentos para coleta de dados de trajetória.

O módulo de posicionamento contempla o mecanismo utilizado para identificação da posição do indivíduo no espaço. Dentre as soluções mais comuns está o sistema de posicionamento global (GPS) e o sistema baseado em localização, do inglês *Location Based System* (LBS). O posicionamento via GPS utiliza as informações enviadas por conjunto de satélites em órbita terrestre para obter a localização sendo necessárias no mínimo 3 satélites para que essa localização seja considerada válida.

O módulo de comunicação aborda a forma utilizada para transmissão dos dados coletados. Entre as alternativas mais comuns estão as tecnologias de radiofrequência e GPRS utilizados em grande parte por empresas de telefonia. Esse módulo permite a comunicação bidirecional dos dados entre o dispositivo e aplicações.

O módulo de alimentação corresponde aos componentes utilizados para manter o dispositivo ligado como baterias e outras fontes de energia ininterruptas. Equipamentos que não são conectados a fontes ininterruptas de energia possuem uma bateria interna que possibilita o funcionamento dos demais módulos. Em geral, quanto menor o equipamento menor será a autonomia energética em função do tamanho da bateria.

Os dispositivos possuem estruturas para armazenamento temporário ou definitivo dos dados como memórias e discos rígidos definidos no módulo de armazenamento. Esses módulos são importantes, sobretudo, para armazenar os dados coletados que excepcionalmente não

foram enviados devido alguma perda de comunicação ou caso o equipamento tenha sido concebido com esse propósito de armazenamento interno dos dados para posterior coleta.

Já o módulo de processamento é responsável por fazer a coleta dos dados entre os demais periféricos, realizar uma formatação desses dados e padronizar de acordo com as mensagens que serão enviadas seguindo um protocolo interno de comunicação utilizado pelo dispositivo.

Por fim o módulo de entrada e saída compreende um conjunto de sensores e atuadores que podem ser utilizados para coleta de informações específicas que deseja-se medir ou para execuções de ações que serão realizadas em determinadas circunstâncias.

A combinação das características desses módulos possibilita a elaboração de um equipamento que permite a coleta dos dados do movimento de indivíduos. A quantidade de dados coletados e a qualidade desses dados está diretamente relacionada ao escopo de utilização do equipamento. Em um cenário onde o equipamento não possui uma fonte ininterrupta de energia, ou seja, possua apenas uma bateria interna, o volume de mensagens enviadas provavelmente terá uma taxa de transmissão menor do que em situações onde o dispositivo permaneça ligado continuamente.

No exemplo de aplicação exibido na Figura 4 é utilizado um equipamento no indivíduo que será analisado - nesse caso um veículo - e dessa forma é possível obter os dados sobre o movimento e trajetórias realizados por esse veículo. Para isso é necessário que seja feita a instalação do equipamento de forma apropriada. Na instalação, o equipamento é conectado a bateria do veículo através de um componente eletromecânico que permite identificar quando o veículo teve a ignição ligada ou desligada. Nessas situações o equipamento é alimentado de forma correspondente ao estado de ignição do veículo evitando o consumo da bateria do veículo com o mesmo desligado. Esses momentos de transição da ignição geram eventos específicos nesses equipamentos sinalizando os estados de ignição ligada (IGN_ON) e ignição desligada (IGN_OFF).

Dessa forma esses eventos podem ser utilizados para representar respectivamente o início e fim dos movimentos e das paradas realizadas pelo indivíduo. Assim, uma parada é caracterizada pelo intervalo de tempo entre os eventos IGN_OFF e IGN_ON. Já um movimento é caracterizado pelo intervalo de tempo entre os eventos IGN_ON e IGN_OFF.

Além desses eventos que representam o estado do movimento do veículo existem um conjunto de outros eventos que podem ser enviados para sinalizar situações normais ou

anormais e que possam ser relevantes para um determinado contexto. Eventos anormais são gerados utilizando sensores que podem ser instalados de forma opcional para coletar alguma característica externa que deve ser medida. Exemplos de evento anormal são: um evento de bateria principal violada, evento de acionamento de alarme ou evento de abertura de porta. Dependendo do contexto de aplicação esses eventos podem ser utilizados para tomada de decisões.

Por outro lado, eventos normais são aqueles considerados comuns no ciclo de funcionamento do equipamento. Exemplos de eventos normais são: IGN_ON, IGN_OFF, Posicionamento, Sleep. O evento de posicionamento, também conhecido como tracking, sinaliza o local do veículo no momento da coleta dos dados. É através desses eventos que é possível identificar a localização do veículo em um determinado instante do presente ou passado. O evento de posicionamento é enviado de forma periódica durante tempos parametrizados no firmware do dispositivo. Aplicações que realizam o monitoramento de veículos de passeios utilizam o intervalo de tempo entre 1 a 3 minutos. Logo, a localização do veículo corresponde aos eventos de posicionamento enviados durante o movimento.

Se por um lado as aplicações que utilizam esse tipo de evento para obter o histórico do movimento perdem em precisão, devido a falta de dados entre dois eventos consecutivos, elas acabam ganhando em um volume menor de conteúdo para processamento e armazenamento pois uma parte desses dados são repetidos e com pequena variação espacial sobretudo em um ambiente urbano onde o veículo fica parado em semáforos ou congestionamentos. Para o contexto de aplicação de monitoramento e localização de veículos de passeio essa parametrização dos tempos dos eventos de posicionamento é suficiente.

O evento de Sleep é utilizado para registrar o momento exato que o dispositivo foi desligado. Isto ocorre como estratégia energética para evitar o consumo de bateria. Quando o equipamento entra nesse estado os módulos são desligados e somente serão ligados novamente quando houver um novo evento de ignição. É comum na lógica do firmware do equipamento que o evento de Sleep aconteça alguns minutos após o evento de IGN_OFF. Geralmente utiliza-se esse intervalo de tempo para realizar alguma configuração do equipamento ou para que o mesmo envie dados coletados e armazenados em memória que eventualmente ainda não foram enviados.

Embora essa estratégia de desligar o equipamento seja eficiente no aspecto energético isso acaba gerando um efeito colateral nos dados de localização que serão coletados e enviados.

Isso porque quando o equipamento é ligado seus módulos são reiniciados e o módulo de posicionamento requer um tempo de inicialização maior para conseguir obter as informações sobre os satélites em órbita a fim de gerar uma posição válida. Isso geralmente demora alguns minutos pois depende da localização do veículo e da posição dos satélites. Durante esse tempo de reinicialização os eventos de posicionamento enviados são imprecisos. Alguns equipamentos enviam uma posição padrão nula, outros enviam a última posição válida em memórias. Independente de qual estratégia utilizada é importante que as aplicações que utilizam esse tipo de conteúdo estejam preparadas para lidar com este tipo de imprecisão operacional pois a utilização desse conteúdo pode gerar resultados incorretos.

Além da imprecisão causada pela inicialização do módulo de posicionamento existem duas outras imprecisões técnicas muito comuns dos equipamentos. A primeira ocorre pela perda de comunicação. Isso acontece quando o veículo está em uma região onde não existe a rede de comunicação utilizada pelo equipamento para envio dos dados. Nesse cenário o equipamento armazena os dados em memória para envio posterior. Aplicações que necessitam desse conteúdo em tempo real são as mais impactadas por esse tipo de problema. Por outro lado, aplicações que analisam o histórico do movimento são menos impactadas desde que os dados sejam enviados em algum momento. Geralmente esse tipo de problema acontece em áreas não urbanas. A segunda causa de imprecisão é gerada pela imprecisão espacial do GPS. O GPS é um equipamento de princípio militar que é largamente utilizado na esfera civil. Para fins de segurança existe uma margem de tolerância da localização gerada. Dessa forma a posição real de um indivíduo possui uma margem de erro adicionada de forma proposital que geralmente é inferior a 5 metros. Os módulos de posicionamento dos equipamentos utilizados no monitoramento de veículos possuem uma margem de precisão de 3 metros sendo eficientes para maioria das necessidades.

Além das imprecisões técnicas levantadas existem algumas outras imprecisões contextuais que as aplicações precisam lidar ao utilizar os dados do movimento. Essas imprecisões são causadas, sobretudo, devido à complexidade urbana. O local onde um indivíduo estacionou o veículo pode estar alguns metros distante do real local de interesse do indivíduo. Dessa forma é complexo realizar essa identificação automática de qual local está associado a parada. Além disso na pesquisa do local associado a parada pode haver muitas opções disponíveis gerando uma ambiguidade de identificação. Também pode ocorrer de o local de interesse do indivíduo não constar no conjunto de locais pesquisados. Essas

imprecisões dificultam o processo de enriquecimento semântico das paradas e trajetórias e com isso, análises que utilizem esse tipo de característica podem apresentar resultados incorretos.

Existem alguns trabalhos que propõem alternativas para solucionar esse tipo de imprecisão utilizando informações de redes sociais, aprendizado de máquina e outras técnicas que utilizam fontes adicionais a fim de reduzir essa incerteza. O problema de tentar reduzir essa incerteza através de dados de redes sociais são basicamente dois: é preciso que o indivíduo tenha alguma rede social e também que ele execute alguma ação informando que esteve em determinado estabelecimento em um processo conhecido como check-in. Dessa forma é possível fazer uma correspondência que correlacione os dados identificando os check-ins realizados pelo usuário com os locais em horários próximos das paradas do veículo. Já a abordagem que utiliza aprendizado de máquina, no geral utiliza dados históricos de outros usuários e tenta identificar o local mais apropriado utilizando características comuns como horário, e outros atributos dos locais. A utilização de características de outros usuários pode reduzir a incerteza mas ainda poderá gerar um resultado incorreto.

Dessa forma acredita-se que a utilização dos dados históricos específicos do próprio indivíduo seja a melhor maneira de reduzir esse conjunto de imprecisões. Como o histórico das trajetórias de um indivíduo contempla o conjunto de paradas realizadas pelo mesmo em um período, o simples fato de reunir essas paradas em grupos, utilizando algum critério espacial, permite solucionar alguns dos casos de incertezas técnicas e semânticas. Esses grupos representarão as paradas mais frequentes do indivíduo no período. É natural que alguns grupos contemplem uma quantidade maior de elementos do que outros. Esses grupos com maior quantidade de paradas correspondem aos locais mais relevantes para o indivíduo. Além disso, quanto maior a quantidade de paradas do grupo maior a possibilidade do ponto médio do grupo representar o local exato de interesse do indivíduo. Com isso a utilização do ponto do médio do grupo para identificação do local reduz a incerteza semântica das paradas deste grupo.

Outro benefício da utilização dos grupos de paradas é a possibilidade de identificação de padrões sequenciais que representam os movimentos entre essas paradas. Isso somente é possível devido a essa abstração das paradas. Esses padrões acontecem na reincidência do movimento entre os locais mais relevantes durante um período de tempo.

Assim, a principal vantagem nesse tipo de abordagem é o fato de considerar somente as informações próprias do indivíduo para reduzir essas incertezas durante o processo de enriquecimento semântico. Isso porque grande parte dos movimentos dos indivíduos são

baseados em hábitos ou padrões que se repetem ao longo do tempo. O conjunto desses padrões define o perfil de comportamento do indivíduo. Embora a identificação desse perfil seja complexa, o conhecimento das principais características comportamentais sobre o movimento dos indivíduos é de grande interesse para o desenvolvimento de diferentes tipos de soluções.

3 Trabalhos Relacionados

Neste capítulo é apresentado a revisão dos trabalhos relacionados ao tema de trajetórias semântica. Nas seções seguintes serão apresentadas as principais abordagens levando em consideração os modelos de representação do conhecimento, o processo de enriquecimento semântico e as técnicas de mineração de dados de trajetória.

3.1 Modelos de Representação

O principal objetivo de um modelo de representação de trajetória é fornecer uma representação compreensível dos dados relativo ao movimento de um indivíduo contemplando informações adicionais que permitam que análises sobre esse conteúdo possam ser realizadas.

Os dados brutos de trajetória possuem um baixo nível de abstração, redundância de registros, inconsistências e pouca informação sobre o domínio de aplicação. Dessa forma, a concepção de um modelo de representação de trajetória, composto por entidades mais abstratas e que permita a inclusão de algum tipo de semântica ao movimento do indivíduo, é fundamental para realização de análises sobre esses indivíduos.

O primeiro modelo para representação de trajetórias que considera características semânticas foi proposto por Spaccapietra et al. (2008) é apresentado na Figura 6.

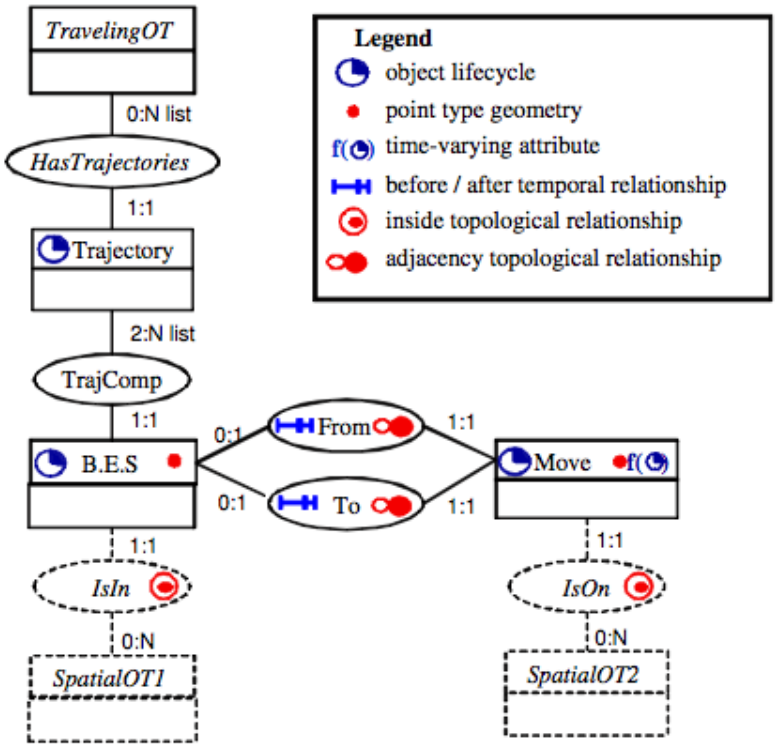

Figura 6: Modelo de Paradas e Movimentos.
Fonte: Spaccapietra, Parent, Damiani, Macedo, Porto e Vangenot (2008).

O modelo de paradas e movimentos apresentado na Figura 6 também é conhecido como modelo BES (*Begin End Stop*) devido aos tipos de paradas. Nesse modelo o indivíduo em movimento (TravelingOT) possui um conjunto de trajetórias (Trajectory) sendo esta composta por uma coleção de paradas (BES). As paradas estão associadas a um local onde o indivíduo permanece por um período de tempo. Um movimento (Move) é o deslocamento entre duas paradas. Dessa forma a trajetória pode ser representada como uma sequência de paradas e movimentos sendo que a principal característica semântica apresentada por esse modelo é referente ao conjunto de locais associados as paradas e aos movimentos.

Esse modelo pode ser estendido e características adicionais da trajetória podem ser incluídas diretamente a uma parada, a um movimento, a trajetória ou ao indivíduo. Esse modelo possui como limitação o fato de não permitir a representação de características que ocorrem em uma determinada posição ao longo do movimento.

Para solucionar esse tipo de restrição foi proposto o modelo CONSTAnT (*CONceptual model of Semantic TrAjecTories*), apresentado por Bogorny et al. (2013). Esse modelo é mais flexível e permite ao usuário adicionar qualquer tipo de informação semântica a diferentes

partes da trajetória. Esse modelo, apresentado na Figura 7, introduz um novo conceito no qual uma trajetória semântica é composta por diferentes subtrajetórias semânticas.

As subtrajetórias semânticas podem ser geradas baseadas nos objetivos, meios de transporte ou comportamento do objeto em movimento (BOGORNY et al., 2013).

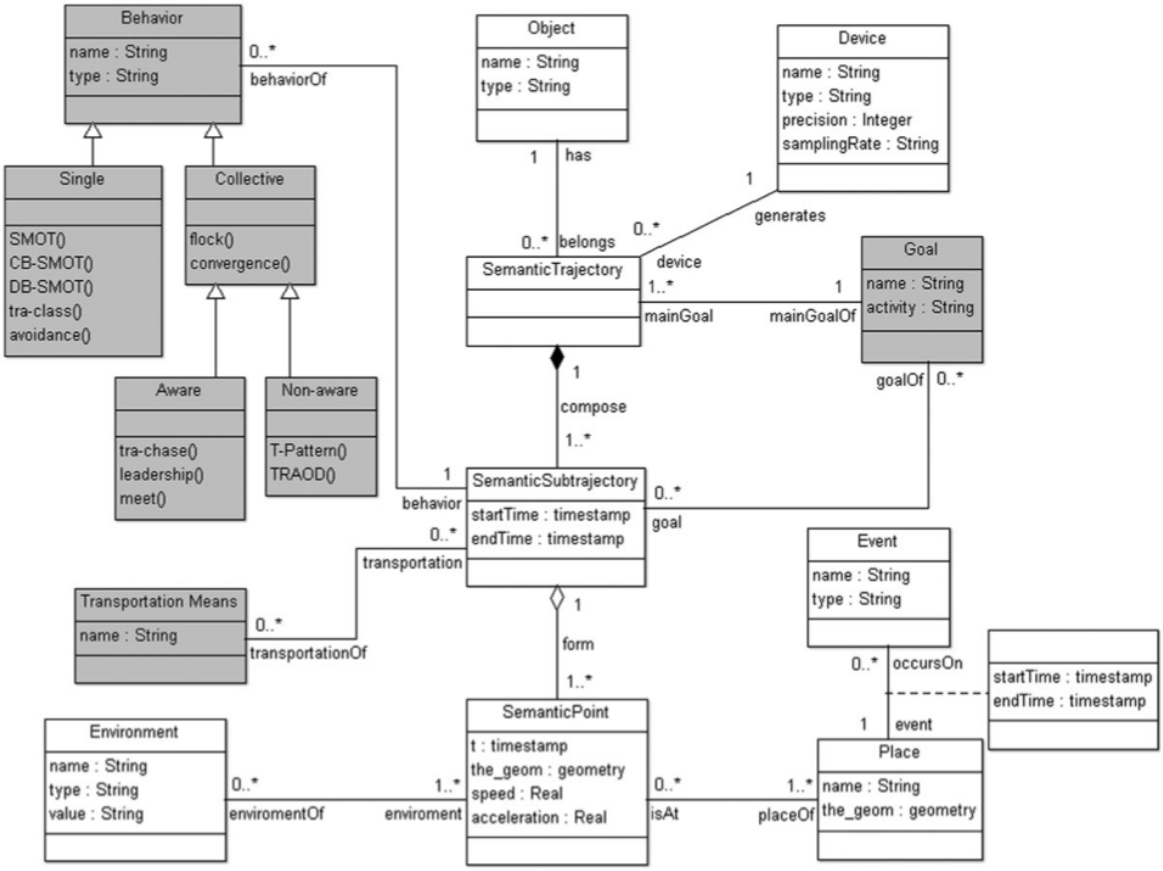

Figura 7: Modelo CONSTAnT de Trajetória Semântica.
Fonte: Bogorny, Renso, Aquino, Siqueira e Alvares (2013).

O modelo CONSTAnT, apresentado na Figura 7, pode ser subdivido em 2 partes. A primeira, com as entidades em branco, representa os conceitos relacionados ao indivíduo (Object) que está em movimento, ao dispositivo (Device) que gera a trajetória, a trajetória semântica (SemanticTrajectory), subtrajetórias (SemanticSubtrajectory) e outras entidades.

A segunda parte, com as entidades em escuro, representa conceitos adicionados ao modelo para serem utilizados em métodos avançados de exploração dos dados, como técnicas de mineração de dados. Essas entidades representam o objetivo (Goal) da trajetória, meios de transporte (Transportation Means) e comportamento (Behavior) das trajetórias semânticas.

A possibilidade de representação do comportamento é uma característica importante em uma trajetória semântica, pois um comportamento apresenta um conjunto de características relativas ao movimento que identificam um modo peculiar de um indivíduo ou de vários indivíduos. Comportamento é algo que geralmente não está explícito nos dados e dessa forma essas informações são geradas por algoritmos e técnicas de mineração de dados.

Alguns trabalhos utilizam ontologia para representação do conhecimento de trajetórias semânticas. Como a anotação semântica e o enriquecimento de trajetórias de objetos em movimento é realizado em um nível semântico, é fundamental a utilização de ontologias (DENTLER et al., 2011).

No trabalho de Yan et al. (2008) é apresentada a proposta de uma ontologia (Figura 8) para modelagem de trajetórias semânticas composta por três módulos: módulo geométrico (*Geometric Trajectory Ontology*), módulo geográfico (*Geography Trajectory Ontology*), módulo de aplicação (*Application Trajectory Ontology*). A maior vantagem de usar uma estrutura modular é devido a facilidade de manutenção.

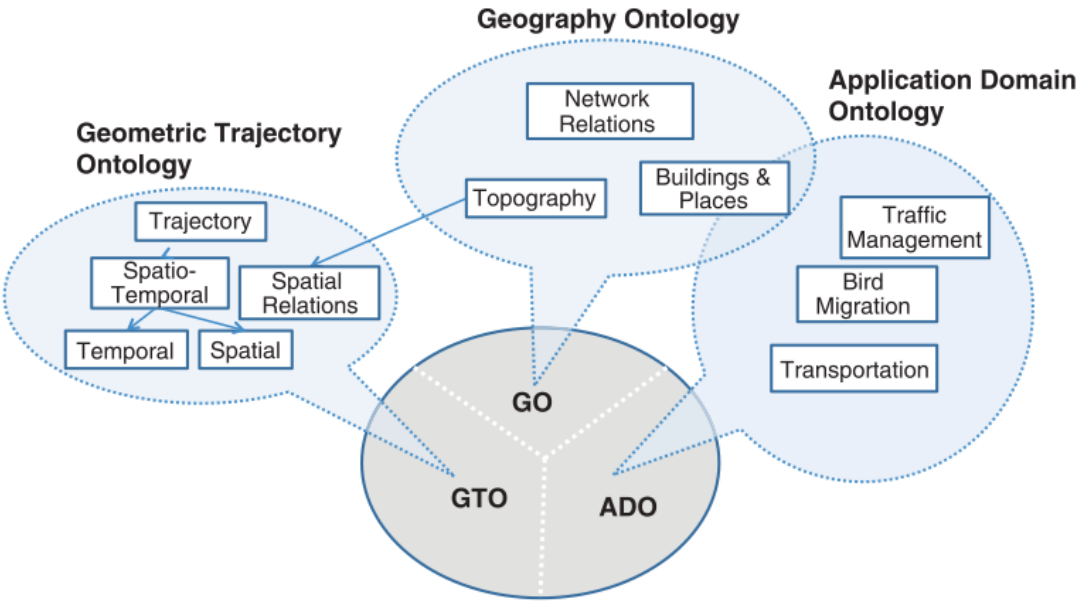

Figura 8: Proposta de Ontologia para Trajetória Semânticas.
Fonte: Yan, Parent, Macedo e Spaccapietra (2008).

O módulo geométrico (Figura 8) abrange um conjunto de aspectos genéricos para descrição de componentes geométricos da trajetória. Neste módulo estão definidos os aspectos espaciais e temporais necessários para uma descrição completa dos dados da aplicação. Por exemplo, uma cidade pode ser representada no formato de um ponto ou de uma região. Os

conceitos típicos para representação de trajetória, como as entidades que definem paradas e movimentos, estão inclusas nesse módulo.

O módulo geográfico abrange conceitos relacionados a topografia natural ou artificial, paisagens, construções, vegetação, entre outros. Um importante aspecto a ser destacado é que alguns elementos dessa ontologia estão fortemente relacionados com a parte geométrica, pois os elementos devem ser representados nesse formato. Por exemplo, uma escola adicionada em conjunto com outros Pontos de Interesse, ou simplesmente POI (*Point of Interest*), é representada no formato geométrico de ponto.

Por último, o módulo de aplicação representa todos os conceitos específicos da aplicação em análise. Dependendo de qual contexto de aplicação a ontologia for utilizada, podendo incluir aspectos relacionados ao gerenciamento de tráfego, informações de migrações de pássaros, entre outras. A combinação dos três módulos ontológicos fornece uma completa descrição semântica para aplicações que necessitam utilizar trajetórias em diversos contextos.

A combinação desses 3 módulos, geométrico, geográfico e de aplicação geram a ontologia semântica de trajetória (*Semantic Trajectory Ontology*). Essa ontologia final fornece uma descrição semântica completa das trajetórias, com aplicação relevante e significado semântico em um domínio específico (YAN et al., 2008).

O trabalho de Baglioni et al. (2009) apresenta uma proposta de ontologia modelando os aspectos de trajetória e os padrões de trajetória gerados. O trabalho de Fileto et al. (2013) apresenta a ontologia Baquara que fornece um modelo conceitual para representar trajetórias, episódios, padrões e outros aspectos referente ao movimento. Esses conceitos e propriedades permitem a descrição do movimento usando *Linked Data* (FILETO et al., 2013). A Figura 9 apresenta os principais conceitos e relações dessa ontologia.

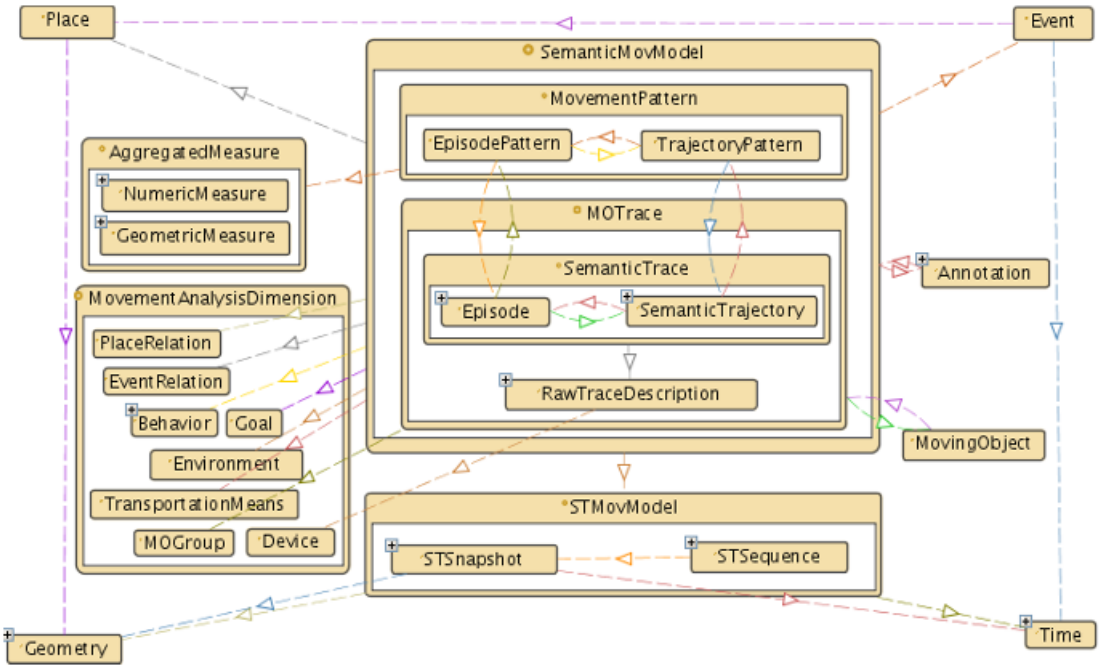

Figura 9: Ontologia Baquara.
Fonte: Fileto, Krüger, Pelekis, Theodoridis, Renso (2013).

A ontologia Baquara foi desenvolvida para servir como um modelo conceitual para descrever trajetórias semânticas em vários domínios. Ela possui as principais relações e atributos necessários para descrever e analisar trajetória e pode ser adaptada para algum domínio de aplicação específico.

Os trabalhos relacionados a representação de conhecimento de trajetórias semânticas, seja através da utilização de ontologias ou outras técnicas, apresentam uma dificuldade na representação das características que definem o comportamento dos objetos em movimento.

Devido à falta de padronização e do consenso na definição dos conceitos taxonômicos de comportamento, é essencial uma definição precisa sobre esses atributos para análises desses comportamentos relevantes na aplicação de mobilidade (PARENT et al., 2013).

O processo de modelagem, no formato de ontologia, dos aspectos do comportamento de elementos em trajetória é complexo em virtude da falta de padronização da nomenclatura para descrever esses conceitos. (LAUBE, 2009).

O trabalho de Wood e Galton (2009a; 2009b) propõe uma descrição para caracterização de diferentes comportamentos e das relações entre os indivíduos na análise de um comportamento em grupo.

Outros trabalhos relacionados com a definição do comportamento são de Laube e Imfeld (2002), Dodge (et al., 2008), Thériault (et al., 1999), que utilizam informações espaciais e temporais na descrição desses conceitos. A proposta apresentada por Andrienko e Andrienko (2007) leva em consideração características do terreno, como obstáculos, e outras anotações semânticas, como meio de transporte, na definição da semântica do comportamento.

3.2 Enriquecimento Semântico

Na seção anterior foram apresentados alguns trabalhos que permitem representar o conceito de trajetórias semânticas, independente do contexto de aplicação, de acordo com alguns modelos de conhecimento utilizados para o desenvolvimento de análises sobre as trajetórias. Para transformar o dado bruto de trajetória, em seu formato original, em um modelo de dados enriquecido é necessário a realização de um processo, denominado de enriquecimento semântico, composto por um conjunto de etapas com o objetivo de limpar, estruturar e adicionar significado aos dados.

A primeira etapa no processo de enriquecimento semântico é necessária para que seja feita uma limpeza e processamento inicial dos dados brutos. Isso acontece porque conforme apresentado por Yan (2009), os dados brutos do movimento não são totalmente confiáveis, possuindo uma série de registros imprecisos e com ruídos. Dessa forma, é necessário realizar um prévio processamento desses dados a fim de detectar, corrigir e remover esses registros para que os dados de movimento possam ser corretamente analisados.

Esse processamento prévio utiliza algoritmos que realizam a remoção dos registros corrompidos, correção de erros de GPS e interpolação de pontos. Também é comum, dependendo do contexto da aplicação, que seja realizado um processo de compressão e redução dos dados. Isso acontece em cenários onde o volume gerado por cada dispositivo é muito grande e parte desse volume acaba sendo desnecessário para determinadas análises. O conjunto de ações definidas neste processamento preliminar é denominado de limpeza de dados (*data cleaning*).

Em seguida, no processo de enriquecimento semântico, os dados são organizados conforme os principais conceitos do modelo de representação de trajetória utilizado. O objetivo dessa etapa é organizar o conjunto de dados em trajetórias e demais estruturas que a compõe

como paradas e movimentos. Nessa etapa de estruturação de trajetórias são utilizados técnicas de segmentação - para reduzir a complexidade do problema dividindo o conjunto de dados em subconjuntos - e algoritmos de identificação de paradas.

Por fim, a última etapa do processo de transformação de trajetória bruta do movimento em trajetória semântica consiste na associação da trajetória estruturada com informações relevantes ao contexto de aplicação. Nessa fase os dados podem ser correlacionados com informações geográficas - como redes de rodovias, pontos de interesse e regiões de interesse - e também podem ser relacionados com dados de redes sociais ou outros conjuntos de dados relevantes ao domínio de aplicação desenvolvido.

O trabalho de Yan et al. (2010) apresenta um processo de transformação dos dados brutos de movimento em trajetórias semânticas. Esse fluxo de transformação pode ser observado na Figura 10.

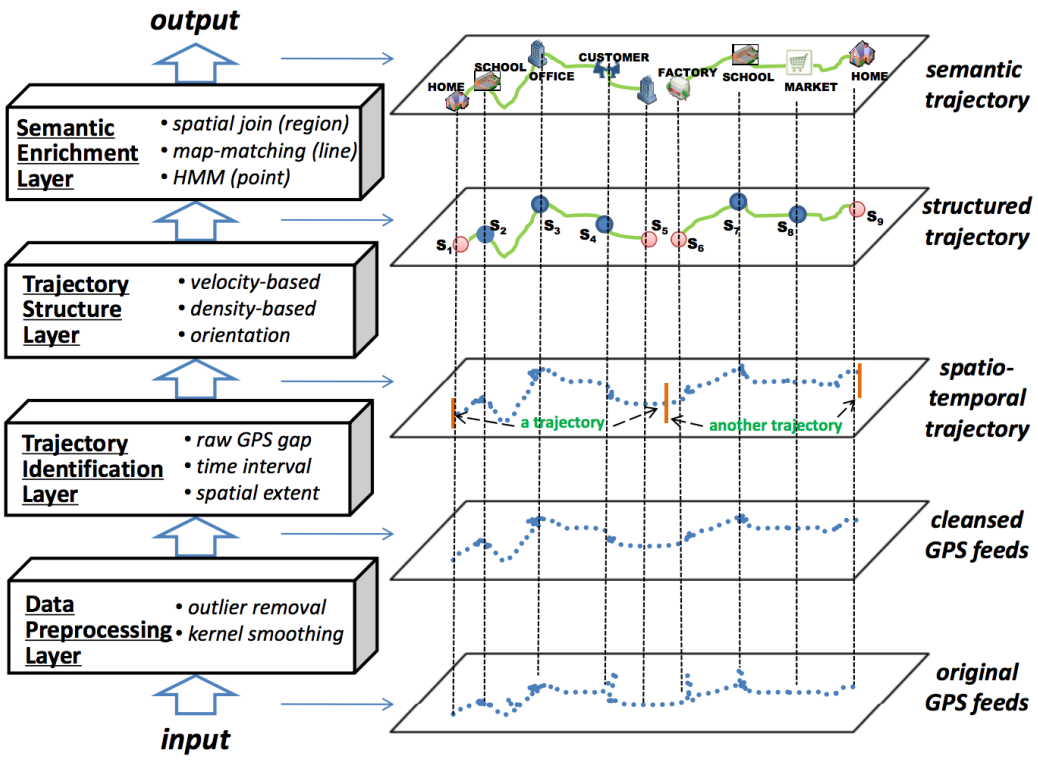

Figura 10: Etapas enriquecimento semântico de trajetória.
Fonte: Yan, Parent, Spaccapietra e Chakraborty (2010).

A Figura 10 apresenta o fluxo para enriquecimento semântico em um processo realizado em 4 etapas disponíveis na forma de camadas: Camada de Processamento dos Dados

(Data Preprocessing Layer), Camada de Identificação de Trajetórias (Trajectory Identification Layer), Camada de Estruturação de Trajetória (Trajectory Structure Layer) e Camada de Enriquecimento Semântico (Semantic Enrichment Layer). Na primeira etapa, Processamento dos Dados, os dados brutos recebidos são processados e são aplicados algoritmos para remoção de incertezas e impurezas tornando esse conteúdo limpo e apto a ser utilizado na sequência. Em seguida, na etapa de Identificação de Trajetórias, é feita a segmentação do conjunto de dados utilizando algum critério espacial ou temporal resultando em um conjunto de trajetórias segmentadas. Na sequência, na etapa de Estruturação de Trajetória, são aplicados algoritmos para identificação das paradas e movimentos para cada um dos segmentos gerados na etapa anterior. Diversas abordagens podem ser utilizadas para identificar as paradas entre elas técnicas baseadas na velocidade, densidade ou orientação. Ao final dessa etapa as trajetórias estão estruturadas. Por fim, na etapa de Enriquecimento Semântico, são utilizados conjunto de dados externos combinado com técnicas computacionais que permitem a identificação dos locais associados as paradas, regiões ou vias pelo qual o indivíduo percorreu. Ao final dessa etapa, e do fluxo como um todo, as trajetórias são representadas de forma semântica.

Como o enriquecimento semântico é complexo e composto por um conjunto de etapas onde cada uma delas abrange algoritmos e técnicas diferentes para manipulação e transformação dos dados, foram desenvolvidos alguns frameworks que realizam esse trabalho de forma integrada e contínua.

A Figura 11 apresenta o framework SeMiTri (YAN et al., 2011), que foi desenvolvido com o propósito de ser uma solução completa e integrada para permitir o enriquecimento semântico dos dados de trajetórias em diferentes formatos.

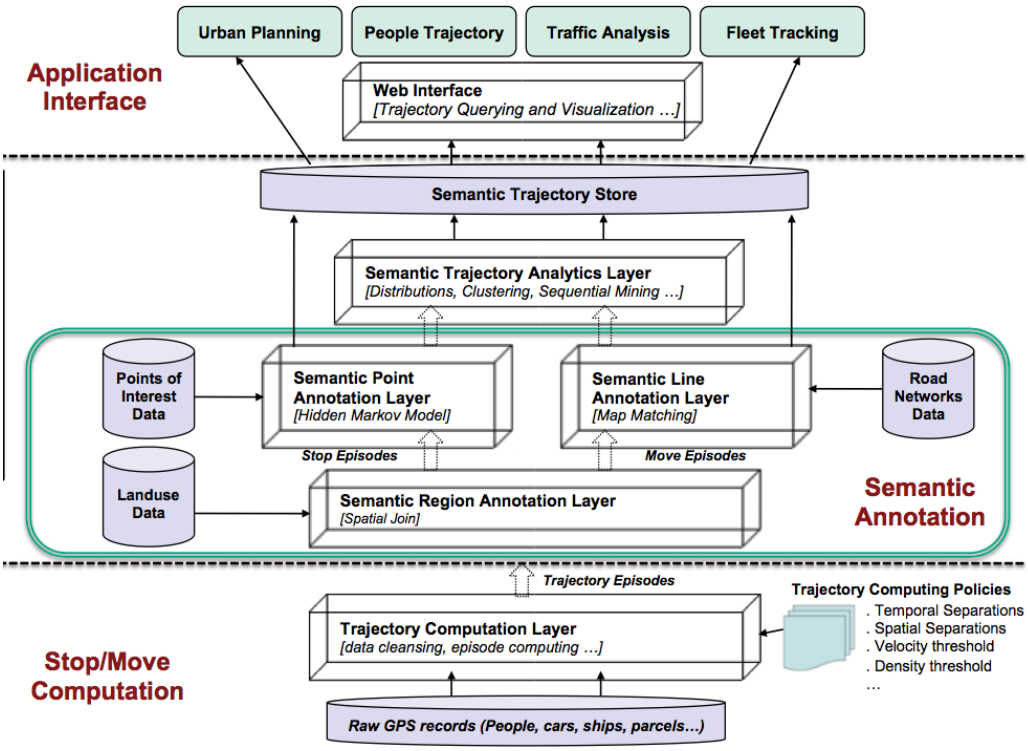

Figura 11: Framework SeMiTri.
Fonte: Yan, Chakraborty, Parent, Spaccapietra e Aberer (2011).

SeMitri é uma solução unificada para anotar trajetória com informações semânticas que prontamente podem ser exploradas por aplicações que necessitam dos dados de movimento (YAN et al, 2011). Esse framework é composto por 3 camadas: Stop/Move Computation que realiza o pré-processamento, limpeza, segmentação e estruturação das trajetórias; Semantic Annotation que possibilita o enriquecimento semântico utilizando bases de dados externas como rede de rodovias, regiões e pontos de interesse; e, Application Interface com um conjunto de interfaces para consulta e acesso aos repositório de trajetórias semânticas que podem ser utilizadas por diversos tipos de aplicação dependendo do contexto.

O framework Baquara, apresentado na Figura 12, é outra solução que foi desenvolvida e que permite realizar o enriquecimento semântico através da utilização de dados ligados.

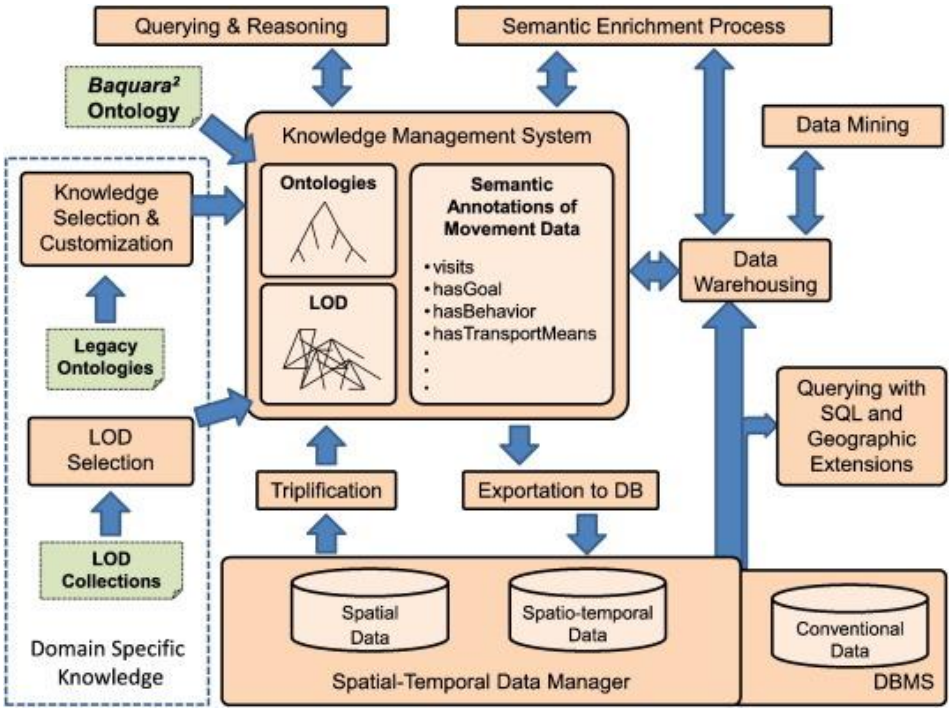

Figura 12: Framework Baquara.
Fonte: Fileto, May, Renso, Pelekis, Klein, Theodoridis (2015).

O framework Baquara apresentado na Figura 12 utiliza o modelo de representação de trajetórias de mesmo nome apresentado na Figura 9 da Seção 3.1 (Modelos de Representação). Nessa proposta o enriquecimento semântico é realizado utilizando anotações textuais realizadas em cima de dados de movimentos. Essas anotações textuais podem ser obtidas através de aplicações que utilizam redes sociais.

O trabalho de Ilarri et al. (2015) apresenta o framework SemanticMove, organizado de forma distribuída, que o diferencia dos demais trabalhos. Nessa proposta, cada objeto em movimento possui uma infraestrutura local que permite coletar, organizar e analisar os dados. Os dados são compartilhados entre os objetos e são sincronizados para um repositório central.

Esses frameworks fornecem um conjunto de ferramentas e algoritmos que permitem a transformação de forma progressiva dos dados brutos de trajetórias em dados enriquecidos. O trabalho de Parent et al. (2013) apresenta um conjunto de algoritmos desenvolvidos para realizar o processo de enriquecimento semântico. São apresentadas soluções para limpeza dos dados, correção espacial dos dados de trajetória, compressão de dados, identificação de paradas e segmentação de trajetórias. Grande parte dessas técnicas são utilizadas nesses frameworks apresentados.

3.3 Mineração de dados de trajetória

Os dados de trajetórias formatados, conforme os modelos de representação de conhecimento, e semanticamente enriquecidos, possuem um conjunto de características representativas sobre o movimento realizado pelo indivíduo. Essas características podem ser exploradas por aplicações com o objetivo de identificar padrões no movimento realizado por um indivíduo em específico ou em um grupo de indivíduos.

Análises que utilizam dados de trajetória enriquecidos semanticamente permitem a descoberta de padrões mais úteis para o contexto de aplicação como a causa do trânsito em determinada região ou o comportamento relativo a um grupo de objetos em movimento baseado não somente na proximidade espacial mas também em similaridade semântica (BOGORNY et al., 2013).

A descoberta de conhecimento em dados de movimentos tem sido uma das comunidades de pesquisa mais produtivas, gerando resultados científicos substanciais como pode ser observado pela grande quantidade de algoritmos e métodos desenvolvidos na última década (GIANNOTTI et al., 2007).

A descoberta, a representação e a análise de padrões de trajetória desafiam a comunidade científica com relação a métodos para agregar, generalizar e explicar padrões. Nesse cenário, técnicas de mineração de dados exercem um papel fundamental pelo fato de mineração de dados ser um mecanismo para extração de padrões de dados através da aplicação de algoritmos específicos (BAGLIONI et al., 2009).

A maioria dos métodos que extrai padrões de trajetória são focados em características geométricas da trajetória, o que gera resultados pouco representativos para o usuário final, além da possibilidade de gerar análises e conclusões incorretas. Esse fato pode ser demonstrado analisando a Figura 13.

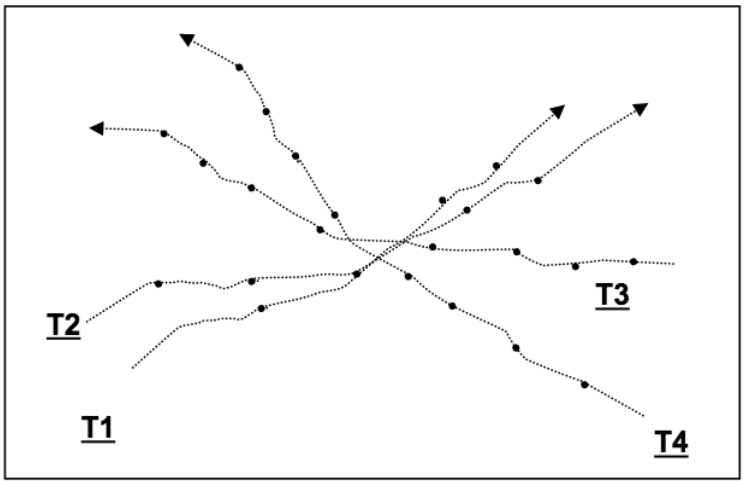

Figura 13: Padrão Geométrico de Trajetória.
Fonte: Bogorny (2016).

Analisando a Figura 13 é possível observar que 4 trajetórias são representadas e identificadas por T_1, T_2, T_3 e T_4. Analisando apenas o aspecto geométrico e espacial dessas trajetórias, algoritmos de agrupamento podem gerar como conclusão que as trajetórias T_1 e T_2 possuem características similares. O mesmo pode ser concluído sobre as trajetórias T_3 e T_4. Essa similaridade acontece baseado na característica espacial e de direção das trajetórias.

No entanto, ao adicionar conteúdo semântico nos dados de trajetória, como pode ser observado na Figura 14, no qual foram adicionadas informações sobre o local de saída e destino das trajetórias, além de um local intermediário onde as trajetórias atravessaram, é possível observar que a similaridade, no sentido semântico, acontece entre as trajetórias T_1 e T_3. Nesses casos, a trajetória iniciou em um local identificado como H (Hotel) com destino a um local identificado com R (Restaurante). Uma outra similaridade acontece com as trajetórias T_2 e T_4. Para esses casos o padrão semântico identificado é de trajetórias com destino a um local identificado como C (Cinema) tendo passado por uma região identificada como SC (Shopping Center).

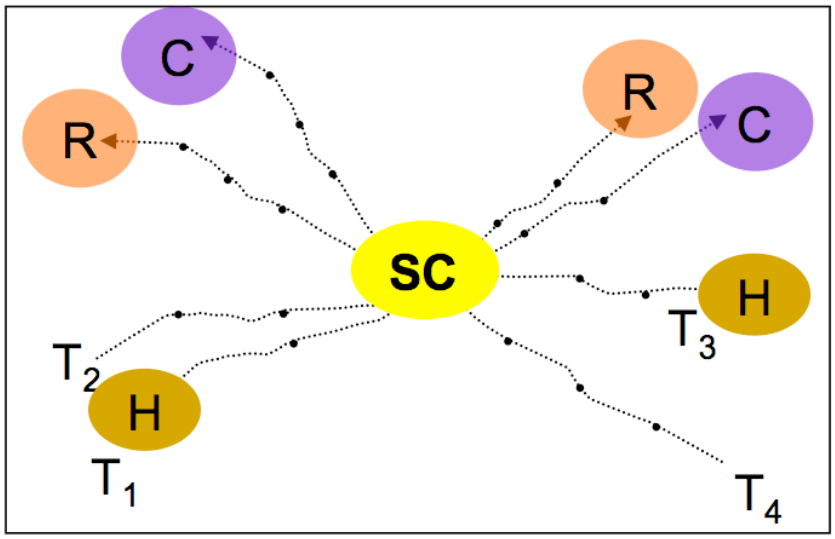

Figura 14: Padrão Semântico de Trajetória.
Fonte: Bogorny (2016).

O trabalho de Zheng (2015) apresenta uma visão geral das principais pesquisas relacionadas a mineração de dados de trajetória como padrões sequenciais de trajetória, classificação de trajetórias, identificação de trajetórias fora do padrão (outliers) e agrupamento de trajetórias similares que compartilham características em comum.

O trabalho de Giannotti et al. (2007) propõe um algoritmo, denominado de T-Pattern, que identifica padrões de trajetórias tendo como base uma sequência de regiões visitadas por um objeto e pelo tempo de permanência nessas regiões.

Muitos dos padrões de trajetória gerados através dessas técnicas e algoritmos de mineração de dados acontecem na forma de comportamento. O comportamento na trajetória é um conjunto de características que indicam o modo peculiar de como um objeto ou conjunto desses se comportam durante um movimento. Esse comportamento é definido através de um atributo que indica se uma trajetória observada possui ou não o comportamento analisado (PARENT et al., 2013).

Os padrões de comportamento de trajetória estão relacionados a aspectos espaciais, temporais e semânticos que o objeto possui ao longo do percurso. Aspectos espaciais são relativos aos locais frequentes de parada, início e fim de trajetória. Por outro lado, aspectos temporais são relativos a sequência realizada durante a trajetória identificando se existem uma determinada ordem e horário. Por último, os aspectos semânticos englobam as características

e os padrões de trajetória de um indivíduo em específico ou de um grupo considerando a semântica do conteúdo da trajetória.

Esses comportamentos podem ser de dois tipos: individuais ou coletivos. Um comportamento individual é uma característica presente em uma ou mais trajetórias de um indivíduo em específico. Da mesma forma, um comportamento coletivo é uma característica presente em um conjunto, não vazio, de trajetórias de múltiplos indivíduos. Os trabalhos realizados por (ALVARES et al., 2011; DODGE et al., 2008; LAUBE et al., 2005; NANNI et al., 2008; WOOD e GALTON, 2009a) apresentam os principais comportamentos individuais ou coletivos de trajetórias dentre os quais vale citar: encontro, bando, liderança, recorrência.

Os resultados dessas análises de comportamento oferecem grandes indicadores sobre as características do indivíduo, onde eles vivem, para onde eles vão, os locais que eles visitam e com quais indivíduos se relacionam. Algumas dessas características são sensíveis no aspecto da privacidade do indivíduo. Enriquecimento semântico de trajetórias potencializa o risco de violação de privacidade porque informações sobre o comportamento do indivíduo são extraídas e representadas em formato explícito e compreensível.

Alguns países criaram leis e normas para controlar o processo de coleta e compartilhamento de informação pessoal. Porém, o fato de existir uma política que regulamenta a privacidade para esse tipo de conteúdo não impede que atividades maliciosas utilizem esses dados. Dessa forma, algumas pesquisas recentes abordam técnicas para proteger e ofuscar características que identifiquem tanto o indivíduo quanto alguns locais de parada que requerem um maior grau de privacidade como hospitais, locais de cunho religioso, político ou que caracterizam algum comportamento sexual.

Dentre as abordagens mais comuns para fazer essa camuflagem existe a técnica que permite ofuscar uma posição exibindo ao invés da localização exata do indivíduo um conjunto de posições próximas. Como contrapartida essa técnica traz a desvantagem de perder a precisão geográfica além de não garantir que o local de parada seja identificado. Para solucionar esse problema de identificação do local de parada, a estratégia mais comum é ao invés de associar a parada a um local em específico relacionar a um conjunto de locais possíveis dentro de um raio de tolerância configurado nos parâmetros de privacidade que a solução utilizará.

Um estudo sobre o aspecto de segurança e privacidade é proposto por Monreale et al. (2011). Nesse estudo é apresentada uma estratégia que utiliza informações semânticas para generalizar alguns dados de localização de forma a deixar esse conteúdo anônimo. O trabalho

apresentado por Parent et al. (2013) utiliza uma alternativa de generalização que leva em conta aspectos semânticos, definidos na taxonomia da ontologia dos locais de parada, também denominados de pontos de interesse, ao invés de utilizar exclusivamente informações do nível espacial. Um exemplo aplicado dessa estratégia pode ser observado na seguinte situação. Ao invés de associar uma parada a um local definido como "Catedral Metropolitana de São Paulo", também conhecida como Catedral da Sé, essa parada poderia ser associada a um local definido como "Local Turístico". Assim, essa estratégia de generalização consegue abstrair algumas características específicas do lugar, que viola a privacidade do indivíduo, ao mesmo tempo que mantém a semântica do local, fundamental para a extração de conhecimentos e comportamentos mais ricos.

Existem diversas abordagens de mineração de dados de trajetórias para identificação de padrões sendo as mais comuns as técnicas para agrupamento e identificação de padrões sequenciais.

3.3.1 Algoritmos de Agrupamento

Os algoritmos de agrupamento são métodos de aprendizado não supervisionado utilizados com o objetivo de agrupar um conjunto de elementos e encontrar alguma semelhança entre esses elementos. São utilizados para identificação de conjunto de trajetórias baseado na similaridade da trajetória como um todo ou em partes específicas da trajetória (HADAJ et al., 2015).

Os algoritmos de agrupamento são organizados de acordo com a estratégia de formação dos grupos e podem ser classificados em: hierárquicos, particionamento, baseado em modelos, baseado em grades e baseado em densidade. A escolha do algoritmo e da estratégia de formação do grupo depende da natureza dos dados e do conhecimento sobre a distribuição desses dados no espaço podendo impactar na quantidade de grupos identificados e de pontos que não pertencem a nenhum grupo.

No contexto de trajetórias os algoritmos de agrupamento são utilizados com o propósito de identificar locais de paradas, encontrar regiões com grande incidência de algum evento que pode ter acontecido durante o movimento ou agrupar trajetórias utilizando algum critério de

similaridade. Nessas abordagens os algoritmos mais utilizados são o K-means, DBScan e algumas outras variações do DBSCan.

DBScan é a abreviação do termo *Density Based Spatial Clustering of Application with Noise* é um algoritmo baseado em densidade (ESTER et al., 1996). Este algoritmo foi utilizado na implementação do processo de agrupamento definido neste trabalho. A principal razão para escolha deste algoritmo para realização do agrupamento de paradas deve-se ao fato de não haver a necessidade de informar a quantidade de *clusters* previamente. Desta forma, o algoritmo identifica e separa os dados em grupos conforme dois parâmetros principais: quantidade mínima de pontos; e, a distância máxima de vizinhança (ε).

3.3.2 Algoritmos de Padrões Sequenciais

Mineração de padrões sequenciais é a tarefa de encontrar todas as subsequências frequentes em uma base dados sequenciais. Uma subsequência é um padrão sequencial se e somente se a quantidade de incidências for superior a um valor mínimo de suporte (*minsup*) definido pelo usuário (SRIKANT e AGRAWAL, 1996).

Os algoritmos para a identificação de padrões de sequência são organizados de acordo com o tipo de estratégia, dentre as quais temos: baseados no Apriori (AprioriAll, GSP); algoritmos baseados em estratégia vertical (Spade, Spam); e, algoritmos recursivos (FreeSpan, PrefixSpan) (FOURNIER-VIGER et al., 2014).

De uma forma geral, esses algoritmos geram o mesmo resultado final caso venha ser utilizado o mesmo conjunto de dados e valor mínimo de suporte. Assim, o que diferencia um algoritmo de um outro não é o resultado, mas a estratégia utilizada para a descoberta desses padrões. Dependendo da estratégia utilizada e do tamanho do conjunto de dados, alguns algoritmos são mais eficientes que outros (FOURNIER-VIGER et al., 2017).

As técnicas que permitem a identificação de padrões sequenciais são aplicadas em diversos trabalhos que utilizam dados de trajetórias, bem como: em soluções para a recomendação de viagens (ZHENG e XIE, 2011); na predição da próxima localização (MONREALE et al., 2009); na comparação de trajetórias utilizando técnicas de similaridades (LI et al., 2008); e, na compressão de dados de trajetória (SONG et al., 2014). O trabalho de

Zhang et al. (2014) apresenta uma proposta, denominada de Splitter, que possibilita a identificação de padrões sequenciais utilizando trajetórias semânticas.

Algoritmos de mineração de padrões sequenciais são utilizados para descoberta de sequências que representam um comportamento regular realizado por um indivíduo ou por um conjunto de indivíduos.

3.4 Conclusão do Capítulo

Neste capítulo foi apresentado o estado da arte dos trabalhos que envolvem trajetórias semânticas. Os estudos dessa área estão organizados em trabalhos que abordam o aspecto de modelagem e representação do conhecimento, soluções que permitem o enriquecimento semântico dos dados brutos de trajetória, e trabalhos que envolvem o processo de descoberta do conhecimento dessas trajetórias.

Para cada um dos aspectos levantados foram apresentadas as principais abordagens e características que distinguem esses trabalhos. Os conceitos apresentados neste capítulo são fundamentais para o entendimento desse trabalho e serviram de base para a identificação de aspectos chaves essenciais para o desenvolvimento de um processo que permite a descoberta de conhecimento utilizando dados de trajetória. Esse conhecimento, obtido através do agrupamento de paradas e do movimento sequencial entre esses grupos, representa algumas características comuns sobre o comportamento regular realizado por um indivíduo ao longo do tempo podendo indicar um perfil ou hábitos de movimentos do indivíduo. Detalhes sobre o processo desenvolvido será apresentado no próximo capítulo.

4 Extração de Conhecimento em Trajetórias

Neste capítulo é apresentada a abordagem desenvolvida composta por um processo que permite a extração de comportamentos de trajetórias. Na próxima seção é apresentada uma visão geral do processo e nas seções seguintes detalhes de cada uma das etapas são descritas abordando as principais características de sua concepção.

4.1 Visão Geral do Processo

A visão geral do processo é apresentada na Figura 15. O processo é composto por duas etapas: agrupamento de paradas e sequenciamento de grupos. Como entrada do processo é fornecido um conjunto de trajetórias estruturadas ou semânticas. O requisito principal é que esse conjunto de trajetória esteja estruturado em um modelo baseado em paradas e movimentos. A primeira etapa, denominada agrupamento de paradas, tem como objetivo identificar os grupos que representam os locais na trajetória onde o indivíduo esteve parado com maior frequência no período correspondente as trajetórias coletadas. Na segunda etapa são geradas as sequências utilizando os grupos descobertos na etapa anterior, para identificar os possíveis padrões de movimento entre esses locais de alta relevância para o indivíduo. Como saída do processo é gerado um conjunto de comportamentos sequenciais baseado na frequência e movimento entre as paradas contidas nas trajetórias. Detalhes sobre o funcionamento interno de cada uma das etapas serão descritas nas seções seguintes.

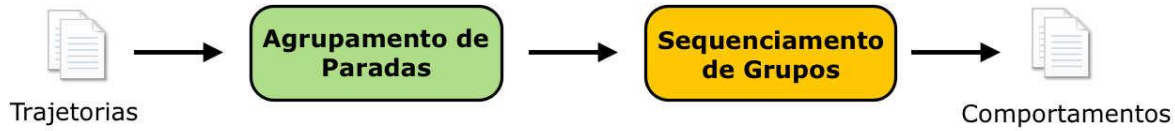

Figura 15: Visão geral do processo de descoberta de comportamentos.

4.2 Agrupamento de Paradas

Na etapa de agrupamento de paradas é identificado os principais locais de parada do indivíduo realizando um agrupamento através de critérios definidos na implementação podendo ser espacial, temporal ou semântico. Os dados de entrada desta etapa representam um conjunto de trajetórias. Como saída é gerada uma lista contendo um conjunto de trajetórias utilizando as sequências diárias do movimento entre os grupos identificados.

Considerando que cada trajetória do conjunto de entrada compreende uma coleção sequencial de paradas, o agrupamento realizado nessa etapa tem como finalidade a identificação de grupos que reúnem paradas com similaridades de acordo com os critérios definidos na implementação. Geralmente, utiliza-se a similaridade espacial e dessa forma reúne-se no mesmo grupo elementos que estejam espacialmente próximos e dentro de um limite de distância pré-estabelecido. Após a identificação dos grupos as trajetórias de entrada são transformadas substituindo as paradas originais por esses grupos descobertos.

De acordo com os critérios adotados para agrupamento das paradas e do conjunto de entrada utilizados, os grupos descobertos podem conter uma quantidade diferente de elementos. Quanto maior a quantidade de elementos de um grupo maior sua densidade e maior sua relevância para o indivíduo no contexto e período observado. Assim, os grupos com uma grande quantidade de paradas representam os locais de maior relevância para o indivíduo no período, compreendendo os locais onde o indivíduo esteve com maior frequência.

Como resultado do agrupamento além do conjunto de grupos descobertos pode existir também um conjunto de elementos que não pertencem a nenhum grupo. Os elementos que não pertencem a nenhum grupo representam valores pouco comuns no conjunto de dados da amostra utilizada. Na prática isso significa que uma parada que não pertence a nenhum grupo é pouco relevante para o indivíduo no período observado ou apenas que não existiu uma quantidade suficiente de incidências que permitisse agrupá-la de acordo com os critérios considerados. Dessa forma, os critérios utilizados para definição dos grupos são fundamentais para que um bom resultado seja gerado nesta etapa, impactando diretamente na quantidade de grupos identificados e dos elementos que não pertencem a nenhum grupo.

A etapa de agrupamento de paradas é composta por alguns componentes com propósitos específicos e que são exibidos na Figura 16. Nas próximas seções serão apresentadas

a concepção e as principais tarefas realizadas pelos componentes internos da etapa de agrupamento de paradas.

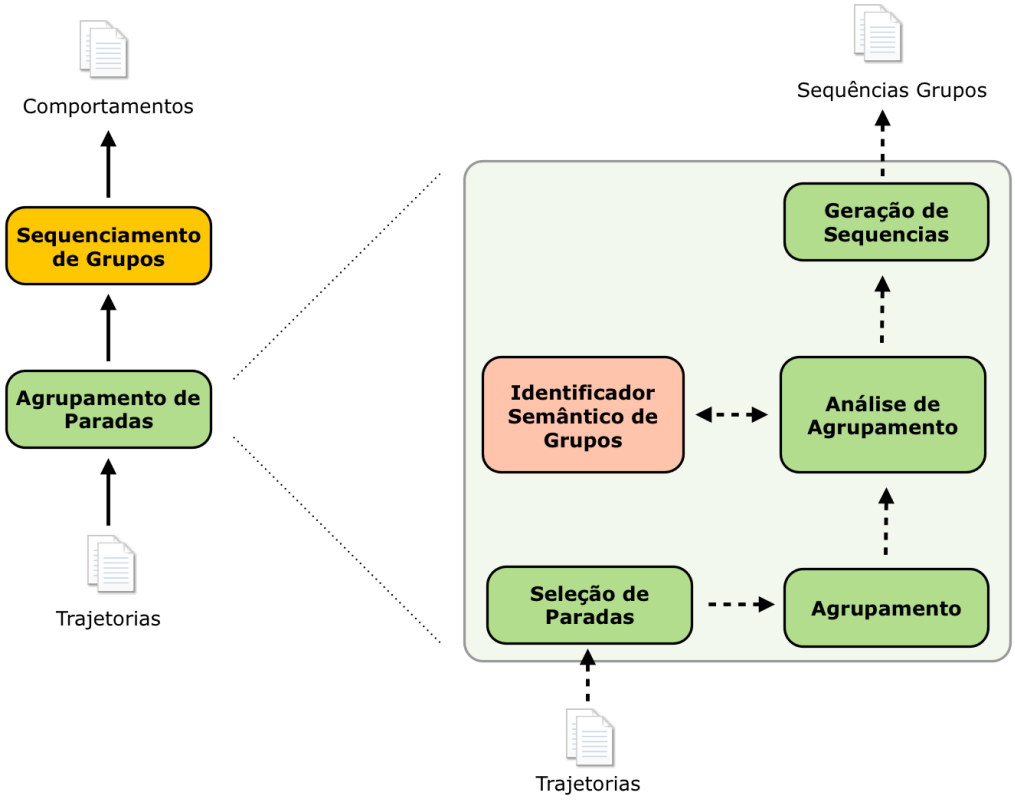

Figura 16: Etapa de Agrupamento de Paradas.

4.2.1 Seleção de Paradas

O objetivo principal do componente de Seleção de Paradas é selecionar as paradas que serão utilizadas no processo de identificação dos grupos. Esse é o primeiro componente da etapa de agrupamento de paradas e recebe como entrada um conjunto de trajetórias. Dessa forma, a principal atribuição desse componente é percorrer a lista de trajetórias e selecionar as paradas que serão utilizadas no agrupamento. Neste componente também é realizada a padronização e a estruturação desse conteúdo de acordo com o formato que será utilizado no componente seguinte denominado Agrupamento.

O motivo de realizar a seleção de paradas, ao invés de utilizar todo o conjunto de paradas contido nas trajetórias, é eliminar eventuais inconsistências técnicas e operacionais que algumas paradas podem conter e que será melhor detalhado no capítulo 5.

4.2.2 Agrupamento

Este componente é responsável por fazer o agrupamento das paradas identificando aquelas que estão espacialmente próximas e reunindo-as em um mesmo grupo. Para geração de um grupo de paradas é utilizado um algoritmo de agrupamento e são utilizados critérios específicos definidos conforme necessidade de implementação. Os principais critérios utilizados para identificação de um grupo estão: a quantidade mínima de paradas, ou pontos, que definem um grupo, e a distância máxima entre essas paradas que delimitam a amplitude do grupo. Assim, esse componente é responsável pela identificação dos grupos que representam os locais onde o indivíduo esteve com maior regularidade e frequência no período contemplado pelas trajetórias. Os parâmetros utilizados nos critérios de agrupamento impactam diretamente na quantidade de grupos identificados. Caso seja utilizado um valor alto como critério de quantidade de elementos para identificação do grupo poderá resultar em uma pequena quantidade de grupos identificados. Da mesma forma, a utilização de um valor alto de distância entre os elementos do grupo poderá reunir dentro de um mesmo grupo elementos que pertençam a grupos distintos.

4.2.3 Análise de Agrupamento

O objetivo desse componente é fazer o gerenciamento e o controle dos grupos identificados no componente anterior e avaliar quais paradas selecionadas no componente de Seleção de Paradas (seção 4.2.1) pertencem a esses grupos. Neste componente também é analisado a densidade do grupo baseado na quantidade de elementos do mesmo. Para cada grupo identificado é gerado um ponto, denominado de centróide, que representa o ponto médio de todas as paradas contidas neste grupo. Através do centróide é possível analisar a amplitude do grupo cujo valor é calculado utilizando a distância entre o centróide e o elemento mais distante. A amplitude permite avaliar se os elementos dentro do grupo estão compactos ou

esparsos. As paradas que não pertencem a nenhum grupo são classificadas como exceções e representam locais de menor relevância para o indivíduo no período correspondente as trajetórias utilizadas. Uma parada deste tipo pode ter sido gerada por uma parada ocasional realizada pelo indivíduo em um local esporádico ou um local cuja regularidade e frequência não permitiu a identificação de um grupo utilizando os elementos do conjunto de dados analisado.

4.2.4 Geração de Sequências

A principal tarefa deste componente é padronizar os dados das trajetórias em um formato que represente a sequência de paradas realizadas pelo indivíduo. Porém, ao invés de utilizar a parada propriamente dita será utilizado o grupo ao qual a parada pertence conforme definido no componente anterior. As sequências geradas representam o movimento realizado pelo indivíduo entre os grupos e paradas de cada trajetória. Ao invés de utilizar o identificador individual das paradas utiliza-se, para as paradas que pertencem a um determinado grupo, o identificador do grupo. Com isso a etapa de Sequenciamento de Grupos (seção 4.3) utilizará essa informação para identificação de possíveis padrões sequenciais. Na Tabela 1 é apresentado um exemplo de como é feito essa formatação de saída. Dada uma trajetória em um determinado dia com 6 paradas identificadas como P_1 a P_6, e os grupos como G_1 e G_2 identificados no componente de Agrupamento, sendo que o grupo G_1 contempla as paradas P_1, P_4 e P_6 e o grupo G_2 contempla as paradas P_3 e P_5. A sequência de paradas do dia ao invés de ser descrita como $P_1 \rightarrow P_2 \rightarrow P_3 \rightarrow P_4 \rightarrow P_5 \rightarrow P_6$ será representada como $G_1 \rightarrow P_2 \rightarrow G_2 \rightarrow G_1 \rightarrow G_2 \rightarrow G_1$. Nesse exemplo a parada P_2 não pertence a nenhum grupo e no momento de formatação deverá ser utilizado o próprio identificador da parada.

Tabela 1: Paradas e Grupos.

Parada	Grupo
P_1	G_1
P_2	-
P_3	G_2
P_4	G_1
P_5	G_2
P_6	G_1

Como esse é o último componente da etapa de Agrupamento de Paradas o conteúdo de saída deste componente será utilizado como entrada da etapa seguinte de Sequenciamento de Grupos.

4.2.5 Identificador Semântico de Grupos

Nesse componente é realizado o enriquecimento semântico dos grupos identificados na seção 4.2.2. Ao utilizar os grupos descobertos anteriormente como critério de identificação do local de parada, as incertezas geográfica e semântica da parada do indivíduo são reduzidas utilizando informações históricas exclusivas do próprio indivíduo. Isso porque grupos com maior densidade de paradas representam os locais de maior frequência do indivíduo e assim as imprecisões geográficas causadas pelo equipamento ou por alguma outra condição espacial são reduzidas ao longo do tempo. Consequentemente o centróide do grupo, identificado no componente Análise de Agrupamento (seção 4.2.3), tende a representar com maior precisão o local correto da parada. Ao realizar o enriquecimento semântico dos grupos, e consequentemente das paradas associadas, o resultado final do processo de extração de comportamentos apresenta um resultado mais significativo para os usuários que farão a análise e interpretação dos comportamentos gerados. Este componente também pode ser utilizado de forma opcional para o enriquecimento semântico das paradas e trajetórias. Nessa situação

deverá existir algum procedimento - adicional ao escopo deste trabalho - para persistir para todas as paradas que pertencem ao grupo o local identificado por esse componente.

4.3 Sequenciamento de Grupos

A etapa de sequenciamento de grupos tem como principal objetivo identificar os comportamentos mais comuns do indivíduo avaliando a ordem, a frequência e também a sequência das paradas presentes nas trajetórias. Para isso, utiliza-se como entrada um conjunto de elementos que correspondem a sequência diária de paradas de um determinado indivíduo sendo que essas paradas estão abstraídas na forma de grupos identificados na etapa anterior. Dessa forma aumenta a possibilidade de identificação de comportamentos recorrentes e consecutivos entre paradas. Isso porque acredita-se que os indivíduos ao longo do tempo realizam movimentos repetitivos referentes a seus hábitos mais comuns. Esses hábitos representam padrões que são descobertos aplicando-se alguma técnica de mineração de padrões sequenciais. Como saída, esse processo apresenta os principais padrões sequenciais descobertos e presentes no conjunto de entrada. Este conjunto de padrões sequenciais identificados determinam o comportamento do indivíduo. A identificação desses comportamentos depende dos parâmetros adotados na implementação dessa etapa e da estratégia de execução utilizada. Além disso, como esses comportamentos são baseados na frequência, também é necessário um conhecimento do conjunto de dados utilizado no momento da definição dessas questões.

A etapa de Sequenciamento de Grupos é composta por alguns componentes conforme Figura 17. Nas próximas seções serão apresentadas a concepção e as principais tarefas realizadas por esses componentes.

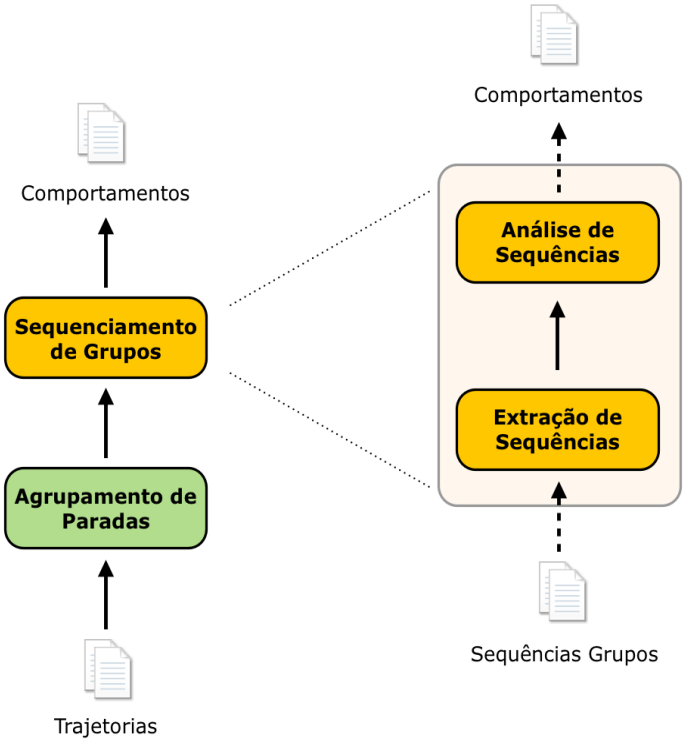

Figura 17: Etapa de Sequenciamento de Grupos.

4.3.1 Extração de Sequências

Esse componente tem como principal função extrair os comportamentos sequenciais do conjunto de entrada recebido. Nesse componente são definidos o algoritmo de mineração de padrões sequenciais e a estratégia utilizada para execução do mesmo. Dependendo da necessidade da aplicação o algoritmo poderá ser executado múltiplas vezes com subconjuntos do conjunto de entrada possibilitando a descoberta de padrões específicos para esses subconjuntos. Sendo assim, para cada execução serão selecionadas as sequências que serão utilizadas pelo algoritmo na pesquisa dos comportamentos comuns ao conjunto de entrada adotado.

Como padrões de comportamento estão associados a hábitos regulares do indivíduo, a execução utilizando características peculiares ao dia da semana poderá gerar resultados interessantes. Um exemplo seria separar os elementos em 2 subconjuntos distintos sendo um deles correspondente a trajetórias realizadas em dias úteis da semana e o outro correspondente a trajetórias realizadas aos finais de semana. Da mesma forma poderiam ser gerados subconjuntos para analisar possíveis padrões regulares existentes em dias específicos da

semana, como as quartas-feiras. Adicionalmente podem ser gerados conjuntos para identificar comportamentos presentes em dias intercalados que representam atividades realizadas 2 ou 3 vezes por semana e assim por diante.

Ao final de cada execução os comportamentos sequenciais descobertos são acumulados em uma lista que armazena o comportamento identificado e a quantidade de incidências. Essa lista é processada pelo componente seguinte.

A Tabela 2 apresenta um exemplo de conjunto que corresponde a sequência de paradas realizadas por um indivíduo entre os dias 24/01/2017 a 29/01/2017. Neste período o indivíduo realizou trajetórias representada pelo movimento entre as paradas abstraídas na forma dos grupos. Os grupos são representados pela notação G_x e as paradas que não pertencem a nenhum grupo são representadas pela notação P_y sendo x e y identificadores individuais dos grupos e paradas. A seta (→) representa o movimento entre duas paradas consecutivas.

Tabela 2: Sequências de paradas abstraídas como grupo

Dia	Sequência de Paradas por Grupo
24/01/2017	$G_1 \rightarrow P_2 \rightarrow G_2 \rightarrow G_1 \rightarrow G_2 \rightarrow G_1$
25/01/2017	$G_1 \rightarrow G_3 \rightarrow G_1$
26/01/2017	$G_1 \rightarrow P_4 \rightarrow G_1$
27/01/2017	$G_1 \rightarrow G_3 \rightarrow P_7 \rightarrow G_1$
28/01/2017	$G_1 \rightarrow G_4$
29/01/2017	$G_4 \rightarrow G_2 \rightarrow G_1$

Para o conjunto de entrada da Tabela 2 alguns padrões sequenciais podem ser observados. O mais comum é que existem 5 incidências do padrão $G_1 \rightarrow G_1$ em 4 dias distintos, sendo que no dia 24/01/2017 esse padrão acontece 2 vezes. Também é possível observar que no dia 25/01/2017 e no dia 27/01/2017 existe um comportamento comum $G_1 \rightarrow G_3 \rightarrow G_1$. Outro comportamento observável é que para todo movimento de G_2 leva a G_1. A complexidade na identificação desses comportamentos está diretamente relacionada ao tamanho do conjunto de

entrada e também a quantidade de paradas de cada elemento do conjunto. A Tabela 3 apresenta o resultado gerado por esse componente.

Tabela 3: Incidências de Comportamentos

Comportamentos	Incidência
$G_1 \rightarrow G_1$	5
$G_1 \rightarrow G_3 \rightarrow G_1$	2
$G_2 \rightarrow G_1$	3

4.3.2 Análise de Sequência

O principal objetivo desse componente é formatar o conjunto de comportamentos descoberto anteriormente para permitir análise dos resultados de acordo com o objetivo da aplicação.

Inicialmente esse componente ordena a lista de comportamentos utilizando como critério o conjunto de incidências gerando uma espécie de ranking. Dessa forma, no topo dessa lista vão estar os comportamentos mais frequentes e no final os menos frequentes. Em seguida esses comportamentos são apresentados no formato sintático ou semântico. A apresentação sintática descreve a sequência dos agrupamentos conforme Tabela 2. Na apresentação semântica as características semânticas dos grupos são utilizadas na formatação do resultado.

Também é papel desse componente relacionar o comportamento descoberto ao conjunto de trajetórias de entrada do processo. Dessa maneira será possível identificar em quais trajetórias existe um comportamento pesquisado. Também poderá ser adicionado a esse componente funcionalidades que permitam a exploração de características adicionais relativas aos comportamentos descobertos. Entre as características adicionais pode-se citar condições temporais relativas ao tempo de permanência nas paradas e o horário mais provável de identificação do padrão encontrado. Essas características adicionais, além de enriquecer o comportamento encontrado, permite que pesquisas utilizem esses critérios no momento da realização das consultas.

Este componente é o último da etapa de Sequenciamento de Grupos e também do processo como um todo. Assim, como saída deste componente será apresentado um conjunto de comportamentos descobertos.

4.4 Conclusão do Capítulo

Neste capítulo foram apresentados as principais etapas e componentes que compreendem o processo de extração de comportamentos sequenciais de trajetórias. A estratégia de reunir as paradas em grupos permite a redução das imprecisões espaciais e operacionais comuns desse tipo de dado. Além disso, a proposta de utilização desses grupos na etapa de mineração padrões sequenciais aumenta a possibilidade de identificação de padrões gerados na forma de comportamentos comuns aos hábitos de movimento do indivíduo. No próximo capítulo será apresentada uma proposta de implementação desse processo além da aplicação dessa implementação em um estudo de caso.

5 Implementação do Processo

Neste capítulo é apresentado detalhes da implementação do processo descrito no capítulo anterior. A implementação foi desenvolvida com o objetivo de identificar os comportamentos que representam o padrão de movimentos de veículos em um contexto de seguro de veículo. Um conjunto de dados de trajetórias foi utilizado para avaliar o sistema proposto. Além do detalhamento da implementação e da metodologia utilizada para aplicação desta implementação no contexto de veículos com seguro serão apresentados as características do motivo da escolha deste conjunto de dados e os resultados encontrados após realização de um estudo de caso.

5.1 Implementação

Como forma de validar o processo apresentado no capítulo anterior foi desenvolvida uma implementação para avaliar as principais funcionalidades concebidas pelas etapas e componentes definidos. A implementação foi aplicada em um contexto de seguro de veículos permitindo a avaliação dos resultados gerados. Para o desenvolvimento desta implementação foi utilizada a linguagem de programação Java em conjunto com algumas APIs (*Application Programming Interface*) que executam funções requeridas na proposta. Os detalhes da implementação serão apresentados em cada uma das etapas do processo definido.

Um conjunto de dados de trajetórias de veículos foi utilizado para avaliar o sistema proposto com o objetivo de identificar os principais comportamentos que definem o perfil de movimento do indivíduo e utilização do veículo. O perfil de utilização do veículo é de grande utilidade para que empresas avaliem o grau de risco e exposição do bem assegurado. Esse perfil de utilização é composto por um conjunto de características que definem os hábitos do indivíduo observados na forma de comportamentos recorrentes e que podem ser expressos pelo conjunto de paradas, movimentos, e também a sequência e ordem das paradas realizadas. Dessa forma os resultados gerados pelo processo de descoberta de comportamentos de trajetória permitem a verificação do perfil de utilização do veículo confirmando ou desmentindo as informações disponibilizadas no momento da aquisição do seguro.

5.1.1 Agrupamento de Paradas

Na implementação do componente de Seleção de Paradas da etapa Agrupamento de Paradas foi utilizado um critério para filtragem e eliminação de paradas subsequentes a movimentos de tempo inferior a 3 minutos. Para isso, durante a seleção das paradas obtidas das trajetórias do conjunto de entrada foram avaliados os tempos relativo aos movimentos entre essas paradas.

No componente de Agrupamento de Paradas foi utilizado o algoritmo DBScan implementado na biblioteca SPMF[1]. SPMF é uma biblioteca de código aberto que contempla uma série de algoritmos para mineração de dados. Os algoritmos disponíveis nesta biblioteca podem ser utilizados em problemas que envolvem descoberta de padrões sequenciais, regras de associação, agrupamento entre outros.

O algoritmo DBScan permite a identificação de grupos através de dois parâmetros principais: quantidade mínima de pontos e distância máxima de vizinhança (ε). O parâmetro quantidade mínima de pontos delimita o número mínimo de elementos necessários para formação de um grupo e por consequência, os grupos identificados necessariamente vão possuir essa quantidade mínima de elementos. Por meio do parâmetro distância máxima de vizinhança é possível avaliar os pontos vizinhos a um elemento pertencente a um grupo. Assim para qualquer elemento pertencente a um grupo existirá obrigatoriamente no mínimo um outro elemento a uma distância inferior ao delimitado pela distância máxima de vizinhança.

Inicialmente, em decorrência do problema e dos dados utilizado foi realizado um estudo para a definição do valor desses parâmetros relativos ao número de paradas nos grupos e a distância entre os pontos de paradas. O conjunto de dados, que abrange o total de 6 meses de coleta, foi dividido em 2 subconjuntos cada qual contemplando 3 meses de dados de forma complementar como será melhor explicado na seção 5.3 Metodologia. Cada um desses subconjuntos foi executado no processo e com isso utilizou-se esse período de 3 meses como referência para o estabelecimento do valor do parâmetro quantidade mínima de pontos. Foi considerado que um local seria relevante para um indivíduo se ao menos uma vez por mês existisse uma trajetória contemplando uma parada no local e dessa forma foi estabelecido o

[1] http://www.philippe-fournier-viger.com/spmf/index.php

valor de 3 como quantidade mínima de pontos para identificação de um grupo. Para o parâmetro distância máxima de vizinhança foi estabelecido o valor de 5 metros. O valor desse parâmetro impacta na quantidade de grupos identificados. Quanto maior o valor de ε maior é a amplitude do grupo agregando elementos esparsos que distorciam a posição central do grupo. Um exemplo pode ser observado na Tabela 4 que demonstra o aumento da amplitude conforme variação da distância máxima de vizinhança para um caso analisado.

Tabela 4: Comparativo entre a distância máxima de vizinhança e amplitude

ε (Metros)	Amplitude (metros)
5	96
10	142
20	343
30	600

A Tabela 4 apresenta o resultado de um experimento no qual foi selecionado um grupo em específico e verificado a variação da amplitude conforme o aumento do valor de ε. Como é possível observar nessa tabela, ao utilizarmos o valor de ε de 5 metros o valor da amplitude fica em torno de 96 metros. Aumentando o valor de ε para 30 metros o valor da amplitude salta para 600 metros demonstrando que no grupo foram selecionados elementos a 600 metros do centróide do grupo.

Como o objetivo da utilização de um algoritmo de agrupamento é a identificação de locais de paradas relevantes para o indivíduo que contemplem uma quantidade mínima de incidências, a identificação desses grupos reduz as incertezas relativas aos locais de paradas. Quanto mais compacto o grupo, ou seja, quanto menor a amplitude do grupo maior é a probabilidade do centróide identificado representar o local correto de parada. Dessa forma foi utilizado um valor moderado de ε para não causar grandes variações na amplitude dos grupos ocasionando a inclusão de paradas distantes não pertencentes ao grupo e que pudessem distorcer a identificação do local.

A utilização desse algoritmo na biblioteca SPMF requer uma formatação específica dos dados de entrada e de saída e para isso também foi implementado no componente de

Agrupamento o processo de formatação dos dados de entrada e saída utilizado no algoritmo, conforme apresentado na Figura 18. O arquivo de entrada utiliza um identificador para cada elemento utilizado, que no caso é o identificador da parada, em conjunto com os respectivos valores da latitude e longitude. O arquivo de saída reúne em cada linha os elementos pertencentes ao mesmo grupo.

Figura 18: Exemplo de arquivos entrada e saída DBSCan.

O componente Análise de Agrupamento utiliza o resultado do processamento do arquivo de saída, obtido do componente anterior, para avaliar a densidade dos grupos identificados, calculando a quantidade de elementos existentes em cada um. Para cada grupo obtido é gerado um número de identificação. Também é calculado o centróide de cada grupo através do ponto médio utilizando a coordenada (latitude e longitude) de todos os elementos pertencentes ao grupo. Através do centróide identificado é calculado a distância máxima entre esse ponto e os demais elementos do grupo, definindo dessa forma a amplitude máxima do grupo. Por fim, é verificado quais elementos do conjunto total de entrada que não pertencem a nenhum grupo representando os casos definidos como *outlier*. A Figura 19 exemplifica o resultado da geração de um grupo e identificação do centróide. Os elementos em azul representam as paradas que pertencem ao grupo e o elemento em vermelho representa o centróide calculado como ponto médio das latitudes e longitudes das paradas.

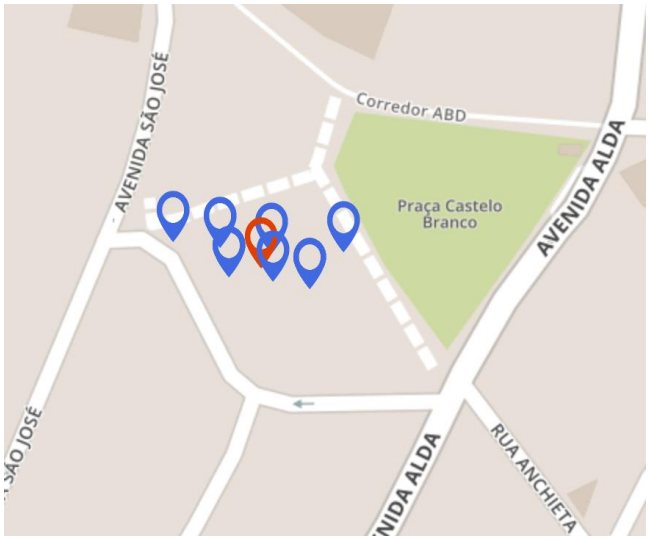

Figura 19: Mapa de Paradas e centróide do Grupo.

Na implementação do componente Identificador Semântico de Grupos foi utilizado a API Overpass[2] que é uma interface que permite a consulta de elementos cadastrados no OpenStreetMap[3] (OSM). OpenStreetMap é uma iniciativa de código aberto e licença livre que fornece um mapa de compartilhamento global que permite a colaboração de usuários para adição e gerenciamento de elementos. Esta iniciativa possui uma base com mais de 3,9 bilhões de elementos cadastrados. Cada elemento possui uma coordenada geográfica, um identificador único e um conjunto de atributos que definem as características desse elemento. Através do Overpass é possível realizar consultas para pesquisa de elementos da base do OSM através de filtros específicos. Dessa forma, foi utilizado essa interface para realizar o enriquecimento semântico das paradas, associando a cada parada o identificador do estabelecimento mais próximo.

Na consulta ao OSM foi utilizado como filtro de pesquisa uma distância máxima de 1.000 metros de raio, tendo o ponto do centróide como centro dessa circunferência e o conjunto com os tipos que classificam esses elementos. Para esta implementação foram considerados os seguintes tipos: estabelecimento (*amenity*), construções (*building*), lazer (*leisure*), escritório (*office*), comércio (*shop*), esportes (*sport*) e turismo (*tourism*). Para cada grupo foi realizado uma consulta como uma requisição HTTP conforme o exemplo abaixo:

[2] http://overpass-api.de/
[3] https://www.openstreetmap.org/

http://www.overpass-api.de/api/interpreter?data=[out:json];node[~%22amenity|building|leisure|office|shop|sport|tourism%22~%22.%22](around:1000,-23.670638050359706,-46.74689758992806);out;

O resultado da consulta ao OSM pode ser obtido em diversos formatos e deve ser informado no momento da consulta. Na implementação foi utilizada o formato JSON (*JavaScript Object Notation*) por se tratar de uma formatação simples e de fácil interpretação. A Figura 20 apresenta um exemplo de conjunto retornado contendo 3 elementos distintos do tipo estabelecimento definido pelo atributo *tags*.


```
{
  "type": "node",
  "id": 3247693030,
  "lat": -23.6719447,
  "lon": -46.7412828,
  "tags": {
    "amenity": "bank",
    "atm": "yes",
    "name": "Caixa EconÃ´mica Federal"
  }
},
{
  "type": "node",
  "id": 3247693038,
  "lat": -23.6659580,
  "lon": -46.7437387,
  "tags": {
    "amenity": "police",
    "name": "Base ComunitÃ¡ria GCM",
    "operator": "Guarda Civil Metropolitana do MunicÃpio de SÃ£o Paulo"
  }
},
{
  "type": "node",
  "id": 3259907562,
  "lat": -23.6651185,
  "lon": -46.7468881,
  "tags": {
    "amenity": "school",
    "name": "C.E.I. Manga Rosa"
  }
},
```


Figura 20: Elementos retornados pesquisa Overpass.

Diversas estratégias podem ser utilizadas para definir qual elemento do conjunto retornado será considerado no enriquecimento semântico. A estratégia adotada considerou o elemento mais próximo ao centróide do grupo. Na Figura 21 é apresentado um exemplo dos elementos retornados e a distância em relação ao centróide do grupo pesquisado. Nesta figura, o elemento em vermelho representa o centróide do grupo e os elementos em laranja representam os locais retornados conforme filtros utilizados no Overpass.

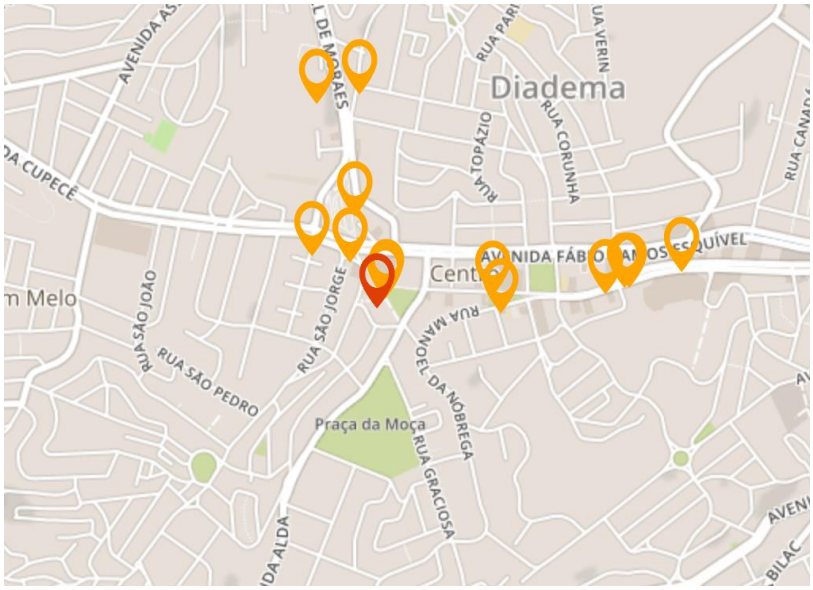

Figura 21: Mapa elementos retornado Overpass e centróide do Grupo.

No estudo prévio da definição da estratégia ocorreram casos onde o elemento mais próximo retornado estava relativamente distante das paradas formadoras do grupo. Para resolver esse tipo de situação foi aplicado um critério adicional, permitindo o vínculo do elemento mais próximo ao grupo desde que este elemento estivesse dentro de um raio máximo definido pela amplitude do grupo. A Figura 22 exemplifica esse cenário.

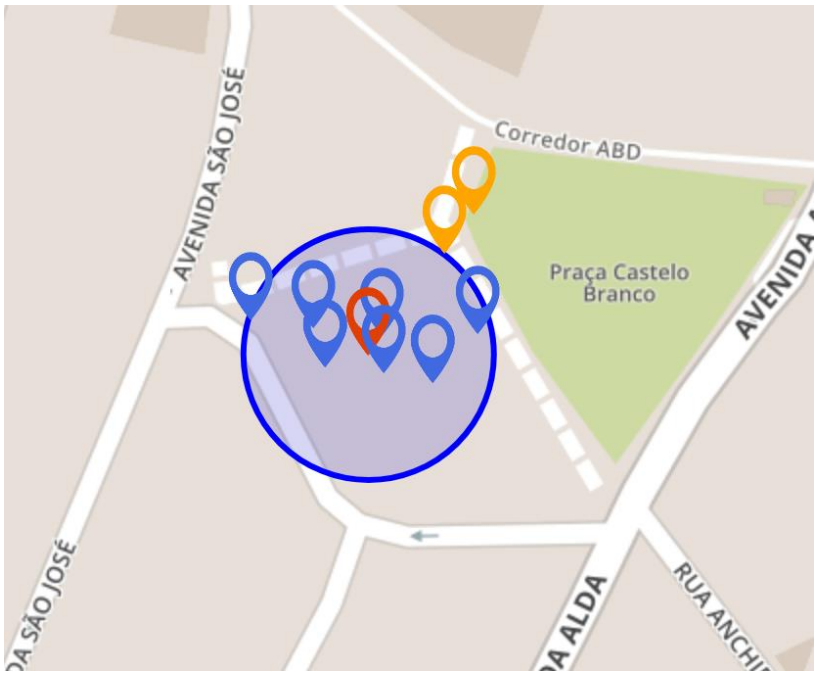

Figura 22: Mapa de raio de amplitude Grupo para identificação de Locais.

O objetivo desse componente Identificador Semântico de Grupos é permitir que os resultados gerados possam ser expressos em um formato mais significativo e representativo para os usuários que utilizam esse conteúdo. Com isso, através desse componente é possível apresentar o comportamento de um indivíduo expresso de forma semântica ao invés de sintática. Para exemplificar, ao invés de demonstrar um comportamento comum de um indivíduo as quartas-feiras como $G_1 \rightarrow G_2 \rightarrow G_3$ poderá ser apresentado como CASA→ BANCO→ ESCOLA.

Mesmo com os benefícios obtidos pela utilização dessa funcionalidade de enriquecimento semântico dos grupos e paradas, existem situações nas quais não foi possível resolver a questão da imprecisão semântica. Foram encontrados casos onde não houve retorno de elementos na chamada a biblioteca Overpass, situações onde foram retornados elementos, porém distantes da amplitude máxima do grupo e casos onde foram encontrados elementos dentro da amplitude máxima do grupo, porém com associação incorreta. Esses casos ocorrem em virtude de imprecisões comuns da base de dados. Muitos estabelecimentos que existem no mundo real não estão cadastrados nas bases e também é comum ocorrerem situações de dados desatualizados. Também pode acontecer de em uma região densa de locais existir um tipo de imprecisão comum da vida real. Em uma rua de comércio popular que possui uma grande quantidade de locais mapeados pode acontecer a situação do indivíduo parar o veículo em um local distante do estabelecimento real ao qual ele se dirigiu. Nesses casos é comum acontecer a associação semântica, porém de forma incorreta. Em geral, nos experimentos realizados foi observado um resultado satisfatório quando o local associado a parada do veículo é referente a um tipo de estabelecimento com estacionamento largo como em casos de shopping, supermercados, clubes, hospitais, escolas e faculdades. Também é observado um bom resultado em estabelecimentos que possuem um estacionamento próprio como postos de gasolina, farmácia, academias e estacionamentos privados.

Assim a definição do conjunto de dados que será utilizado para realizar esse enriquecimento semântico é fundamental para que os resultados sejam gerados de forma confiável. Além disso esse conjunto de dados é específico do contexto da aplicação podendo variar a qualidade dos resultados de solução para solução. Esse talvez seja o grande desafio na utilização desse tipo de abordagem. Neste trabalho mesmo considerando o OSM que é o maior conjunto de dados disponível e que engloba uma grande quantidade de elementos cadastrados

observou várias situações incomuns nos resultados das pesquisas. No desenvolvimento do estudo de caso foi desenvolvido uma abordagem que resolve essa situação para uma parada previamente conhecida do perfil do indivíduo.

O componente de Geração de Sequências transforma as trajetórias em sequências, substituindo as paradas originais pelos grupos identificados no componente de Agrupamento de Paradas e gerando um conjunto de sequências que será utilizado na próxima etapa.

5.1.2 Sequenciamento de Grupos

O componente Extração de Sequências da etapa de Sequenciamento de Grupos utiliza o algoritmo CM-SPADE implementado na biblioteca SPMF. Esse algoritmo utiliza como entrada um conjunto de dados e um valor mínimo de suporte (*minsup*) que representa a frequência mínima para identificação de um padrão sequencial.

A razão da escolha do algoritmo CM-SPADE foi em virtude desse algoritmo utilizar uma estrutura auxiliar que armazena as co-ocorrências encontradas tornando eficiente o processo de poda da busca em profundidade utilizada pelo algoritmo. O algoritmo teve bom desempenho quando comparado com outras alternativas, sendo também vantajoso quando utilizado com conjuntos de entradas pequenos. No contexto desta implementação, em que foi utilizado um conjunto de dados de um período de 3 meses, serão gerados dados de até no máximo 92 dias por indivíduo - considerando que no período possa existir dados referente a dois meses com 31 dias e um mês com 30 - e como cada dia representa uma sequência do conjunto de entrada que será utilizado no algoritmo podemos considerar que o conjunto não é grande.

O valor utilizado no parâmetro de suporte mínimo (*minsup*) que determina a quantidade mínima de incidências de uma sequência para identificação de um padrão foi de 30%. Nos testes realizados para avaliar o algoritmo esse valor apresentou bons resultados para geração de comportamentos.

Na implementação deste componente foi utilizado como estratégia a execução por dia da semana sendo feita a separação do conjunto de dados em sete subconjuntos relativos a cada dia da semana, começando com Domingo e terminando no Sábado. O objetivo dessa estratégia é identificar padrões de movimentos sequências comuns nesses respectivos dias. A expectativa

da utilização desta estratégia era da identificação de comportamentos recorrentes realizados pelo indivíduo em cada um desses dias. Dessa forma, o algoritmo de geração de padrões sequenciais foi executado sete vezes. Para cada execução foi realizada uma seleção das sequências específicas ao dia definido na estrutura de repetição.

Para a execução do algoritmo disponível na biblioteca SPMF é necessária uma formatação específica do conjunto de entrada. Também é preciso processar o arquivo de saída para obtenção dos valores obtidos. A Figura 23 demonstra o formato de entrada e saída utilizado no algoritmo. O arquivo de entrada, que representa o conjunto de sequências do dia da semana, utiliza como delimitador de elemento o valor "-1" e como delimitador final de sequência o valor = "-2". O arquivo de saída que reúne o conjunto de padrões identificados apresenta o padrão e a quantidade de incidências identificadas após delimitador "#SUP".

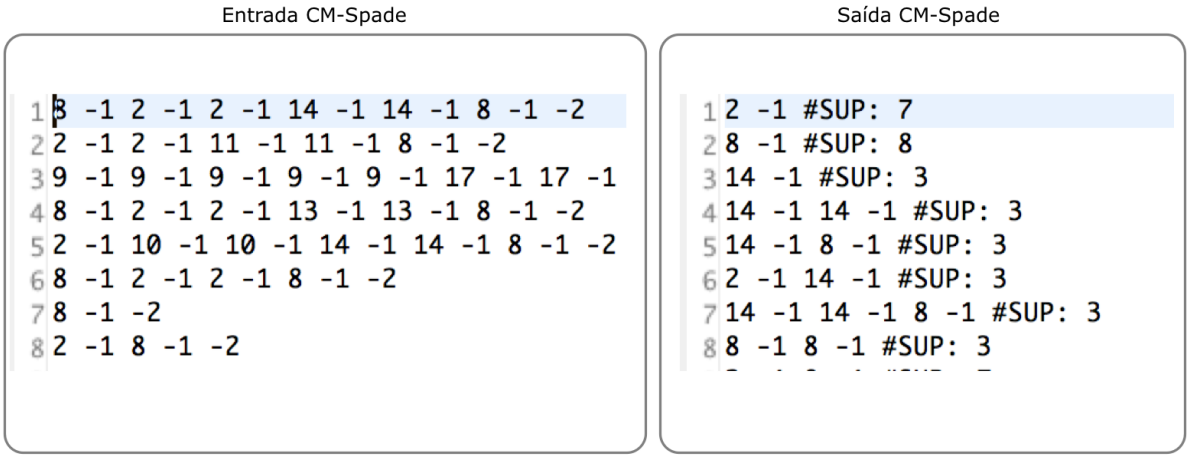

Figura 23: Exemplo de arquivos entrada e saída CM-SPADE.

Ao final da execução do componente Extração de Sequências, as sequências obtidas são armazenadas em uma lista que contém os padrões descobertos e o total de incidências. Esta lista é utilizada no componente seguinte Análise de Sequência.

O componente Análise de Sequência foi implementado para apresentar os resultados tanto no formato sintático como no formato semântico. Este componente primeiramente ordena a lista de padrões na forma de ranking, tendo nos primeiros elementos os padrões mais comuns. Este componente também vincula cada padrão identificado nos respectivos elementos do conjunto de entrada. O objetivo disso é realizar uma associação permitindo identificar a sequência de quais dias são encontradas em um padrão específico. Por fim, o componente

utiliza os resultados obtidos do processo de enriquecimento semântico dos grupos para apresentar um resultado mais significativo para o usuário. Nesta implementação o resultado é apresentado individualmente por dia da semana para facilitar a identificação dos padrões comuns de um dia em específico.

5.2 Conjunto de Dados

A escolha de um conjunto de dados para testar a abordagem proposta envolveu 3 condições. A primeira delas, e talvez mais importante, é referente a natureza dos dados. O conjunto de dados deveria contemplar trajetórias de indivíduos cujos movimentos fossem regulares durante o período, e ao mesmo tempo não fossem totalmente regrados. Em outras palavras o conjunto de dados não poderia ser de indivíduos com movimentos aleatórios como no caso de veículos de táxis pois nesse cenário seria difícil a identificação dos padrões de movimento no período. Ao mesmo tempo esses movimentos não poderiam ser estritamente regulares como é comum em transporte coletivo onde o veículo possui um itinerário definido. Nessa situação o comportamento é constante conforme roteiro pré-estabelecido e consequentemente não haveria uma descoberta significativa dos padrões sequenciais realizados. Também seria necessário que o conjunto de dados abrangesse várias trajetórias do mesmo indivíduo ao longo de um período e que esse conjunto fosse grande o suficiente para que hábitos rotineiros fossem identificados. Como essa abordagem utiliza informações históricas específicas de cada indivíduo para descoberta dos padrões sequenciais de movimento no caso de um conjunto pequeno de trajetórias, como o período de uma semana ou 15 dias, impossibilita a descoberta desses padrões.

A segunda condição relevante para escolha do conjunto de dados envolvia as características técnicas do equipamento utilizado para coleta e geração dessas trajetórias. Isso porque as particularidades técnicas e operacionais na utilização dos equipamentos geram imprecisões nos dados que poderiam resultar em interpretações equivocadas. Como abordado na seção 2.4 de contextualização do problema existem algumas condições onde o módulo de posicionamento gera localização incorreta. Utilizando esse tipo de dados possibilita que comportamentos incorretos sejam extraídos ou até mesmo que não sejam descobertos. Com isso a utilização de um conjunto de dados cujo desconhecimento das características técnicas do

equipamento e de possíveis imprecisões da etapa de coleta dificultaria a extração de comportamentos sequenciais confiáveis.

A terceira condição engloba o regionalismo do conjunto de dados. Essa condição é relevante sobretudo devido a etapa de enriquecimento semântico envolver características espaciais dos locais próximos às paradas. Nessa etapa são utilizadas fontes externas para identificação dos locais próximos às paradas e consequentemente associação entre elas. Utilizando dados de uma região desconhecida possibilitaria a geração de resultados que não poderiam ser confirmados ou questionados.

Os conjuntos de dados utilizados nos trabalhos levantados no estado da arte em sua maioria não contemplavam essas 3 condições de forma conjunta. Essa foi uma das grandes dificuldades encontradas nesse trabalho. Dessa forma optou-se por utilizar um conjunto de dados privados obtidos de uma empresa de rastreamento de veículo. Esses dados compreendem o conjunto de trajetórias de 50 indivíduos obtidos pelo período de 6 meses. Durante esse período foram gerados um total de 7.208 trajetória e 34.445 paradas. Os indivíduos selecionados para o estudo de casos compreendem veículos de passeios localizados no Brasil que utilizam esses equipamentos como medida de segurança para situações de roubo ou furto. Além disso esses veículos possuem seguro e dessa forma foi utilizada a informação do endereço, provavelmente utilizado na geração da apólice do seguro, para verificação dos grupos e comportamentos descobertos no estudo de caso e posterior avaliação de perfil.

Esse conjunto de dados atendeu as 3 condições estabelecidas. O fato dos indivíduos compreenderem veículos de passeio aliado ao período de coleta compreender 6 meses de conteúdo atende a primeira condição. O entendimento técnico-operacional do equipamento utilizado na coleta cumpre a segunda condição. O regionalismo pelo fato dos dados serem obtidos de trajetórias realizadas no Brasil satisfaz a terceira condição. Além disso, o fato do veículo possuir seguro permite que o CEP (Código de Endereçamento Postal) e endereço associado ao cadastro do veículo possa ser utilizado para confirmação dos grupos gerados nas paradas e comportamentos obtidos conforme será detalhado no estudo de caso.

5.3 Metodologia

O conjunto de dados utilizado neste trabalho contempla informações referente a 50 veículos selecionados de uma empresa que atua na área de segurança veicular. Esses veículos estão acoplados de um equipamento que permite sua localização em caso de roubo ou furto. A escolha dos indivíduos também utilizou como condição o requisito dos veículos terem seguro permitindo que a informação do CEP e endereço informado na geração da apólice, pudesse ser utilizado em uma das análises do estudo de caso para confirmação e correspondência aos grupos identificados na etapa de agrupamento de paradas.

Para cada um dos indivíduos foram pesquisados o total de eventos em um período de 6 meses. Após coleta dos dados foi aplicado uma etapa para filtragem de eventos, eliminação de ruídos e geração de trajetórias diárias estruturadas que foram utilizadas no processo. Tomou-se o cuidado de utilizar o mesmo período para coleta dos dados dos indivíduos garantindo que alguma condição temporal, como feriado prolongado, impactasse todos os indivíduos da mesma forma. Também houve a precaução de garantir que todos os veículos dos indivíduos analisados possuíssem a mesma tecnologia utilizada para coleta dos dados aumentando o controle sobre as imprecisões geradas.

A forma utilizada para avaliação dos resultados foi através da comparação dos resultados obtidos da divisão do conjunto de dados em 2 subconjuntos correspondente às fases 1 e 2 conforme Figura 24. Cada uma das fases compreende o período de 3 meses de dados de cada um dos 50 indivíduos. Os dados das fases 1 e 2 são complementares e totalizam o conjunto extraído. A separação do conjunto em duas fases permitiu que os comportamentos gerados ao final de cada fase pudessem ser comparados considerando sempre os resultados específicos de cada indivíduo. Dessa forma não foi realizado nenhum tipo de comparação entre os comportamentos entre indivíduos distintos.

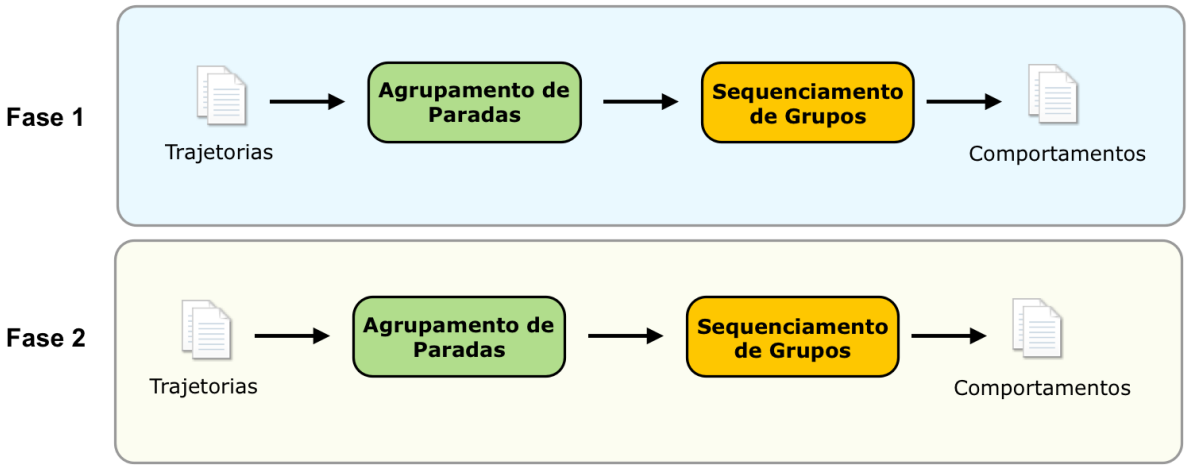

Figura 24: Estudo de Caso - Fases.

O objetivo principal do estudo de caso, detalhado na próxima seção, é avaliar se a solução desenvolvida satisfaz a proposta de identificação de comportamentos que representem um conjunto de movimentos sequenciais que caracterizem o perfil de movimentação de um indivíduo. Nesse caso o perfil abrange um conjunto de comportamentos sequenciais identificados pelas trajetórias do indivíduo no período. Com a divisão do conjunto de dados em 2 subconjuntos foi possível avaliar se o conjunto de comportamentos identificados na fase 1 também estava presente na fase 2. Além disso, o estudo de caso permitiu avaliar se os comportamentos mais relevantes, ou seja, aqueles com maior incidência em cada uma das fases eram comuns representando dessa forma os comportamentos mais relevantes no período como um todo e confirmando a regularidade e manutenção dos padrões de movimento.

A utilização de dados de comportamento para avaliação de perfil de indivíduos é algo de grande interesse para empresas que gerenciam riscos. Para esse tipo de abordagem, como no ramo de seguros, confirmar se a utilização de um bem está de acordo com o informado no momento da aquisição minimiza a possibilidade de fraudes. No segmento de seguro de veículos, durante o processo de aquisição, são solicitadas algumas informações sobre o indivíduo que está contratando o serviço e sobre o bem assegurado. As informações fornecidas são utilizadas para algumas verificações internas comuns desse segmento, para precificação do seguro e também para posterior confronto em caso de suspeita de fraude. Na aquisição do seguro são solicitadas informações sobre o tipo de utilização do veículo para averiguar se o proprietário utiliza para uso doméstico, comercial ou outra finalidade. Também é solicitado o CEP de pernoite do veículo e perguntado se neste local existe garagem. Essas informações são

utilizadas para averiguação do grau de exposição de risco do veículo. Em uma eventual situação de sinistro de roubo ou furto as empresas desse segmento avaliam se o motivo do sinistro foi por alguma inconsistência do perfil de utilização informado causada por negligência, omissão, ou falsidade de informações. Caso confirmado alguma hipótese que caracterize fraude o valor do prêmio pode inclusive não ser pago.

Como não existe outra alternativa para averiguação do perfil de utilização do veículo as empresas desse segmento utilizam essas informações para cálculo do preço do seguro. Em geral consideram índices históricos de sinistralidade da marca e modelo do veículo e também da região de pernoite do veículo como agravantes que encarecem o preço do serviço. Também são utilizados alguns outros critérios relativos ao proprietário do bem, como idade, estado civil e sexo para definir os riscos de exposição do bem.

Esse tipo de abordagem acaba gerando alguns tipos de incoerências permitindo que uma pessoa que utilize pouco o veículo seja prejudicada e pague mais caro pelo seguro simplesmente pelo fato de possuir um modelo de veículo com alto índice de sinistralidade ou por viver em uma região com alta incidência de roubo e furto de veículos. Com isso, a possibilidade de utilizar os comportamentos extraídos das trajetórias desses veículos permite que parte dos riscos de utilização possam ser avaliados e o perfil informado na aquisição do seguro possa ser confirmado ou desmentido resolvendo os casos de inconsistências levantados. Porém, essa possibilidade de verificação e confronto entre o perfil descoberto com o perfil informado não será possível no momento da aquisição do seguro pois o perfil descoberto é gerado em cima dos comportamentos extraídos do histórico de utilização, e isso somente será possível após algum tempo de coleta de dados.

Assim, as informações geradas pelo processo de descoberta de comportamentos de trajetórias somente poderão ser utilizadas no momento da renovação do seguro. Inclusive esse conteúdo poderá ser utilizado como parâmetro de ajuste do valor do seguro e também para uma gestão eficiente da carteira de clientes. Além disso, o entendimento do perfil de utilização permite o desenvolvimento e o oferecimento de serviços específicos e condizentes com as características de movimento do indivíduo.

Na próxima seção será detalhado o estudo de caso realizado e apresentado os principais resultados encontrados.

5.4 Estudo de Caso

Nesta seção são apresentados os resultados de análises realizadas utilizando a implementação do processo para identificação de comportamentos que representem o padrão do movimento de indivíduos aplicados no conjunto de dados selecionado utilizando a metodologia e contexto definidos anteriormente. As análises foram realizadas para avaliação do processo partindo de sua concepção geral até chegar em resultados específicos dos comportamentos identificados.

5.4.1 Aspecto Funcional

A primeira fase do estudo de caso tem como objetivo analisar o aspecto funcional do processo implementado avaliando se essa abordagem pode ser utilizada para obtenção de comportamentos de trajetórias. Essa análise possibilita avaliar a concepção do processo como um todo e verificar se a parametrização dos algoritmos e demais critérios utilizados na implementação permite a geração de resultados. Os resultados obtidos de cada indivíduo nas duas fases distintas estão disponíveis nos Apêndices A e B. Em cada um desses apêndices é apresentado o total de trajetórias realizadas, total de paradas identificadas, total de grupos gerados pela etapa de agrupamento de paradas, total de paradas que não pertencem a nenhum grupo e total de sequências extraídas na etapa de sequenciamento de grupos.

Embora tenham sido utilizados os mesmos períodos de pesquisa em cada uma das fases para coleta dos dados dos indivíduos é comum observar uma variação na quantidade de dados entre os indivíduos. Dessa forma, é possível que para um indivíduo tenha sido coletado dados referente a 182 dias e para outro apenas 35 dias. Também pode ocorrer a variação de dados entre as fases para um mesmo indivíduo. Nessa situação um indivíduo pode ter gerado dados em 87 trajetórias - equivalente a 87 dias - durante a fase 1 e apenas 72 trajetórias - correspondente a 72 dias - durante a fase 2. O motivo dessa variação é complexo e não foi abordado neste trabalho, mas abrange desde questões técnicas do aparelho que impactam na coleta dos dados, como operacional ocasionado em função de condições específicas de utilização do veículo nos períodos observados, como um período de férias, permanência em oficina mecânica, ou outro.

Para avaliação deste estudo de caso foi utilizado a quantidade de comportamentos extraídos de cada indivíduo. Conforme pode ser observado no Figura 25, todos os indivíduos tiveram comportamentos extraídos em ambas as fases. A quantidade média de comportamentos por indivíduo foi 18 no período como um todo sendo que para fase 1 a média foi de 24 comportamentos e para fase 2 a média foi de 13 comportamentos. O motivo principal desta diferença entre a quantidade de comportamentos descobertos entre as fases se deve em função da quantidade de trajetórias. Como o comportamento é baseado em frequência histórica de movimentos que se repetem ao longo do período, quanto maior a quantidade de trajetórias maior a probabilidade de incidências. Por essa razão, os comportamentos identificados na fase 1, representados pelas barras em azul, possuem valores superiores aos identificados na fase 2, representados pelas barras em verde, para maioria dos indivíduos.

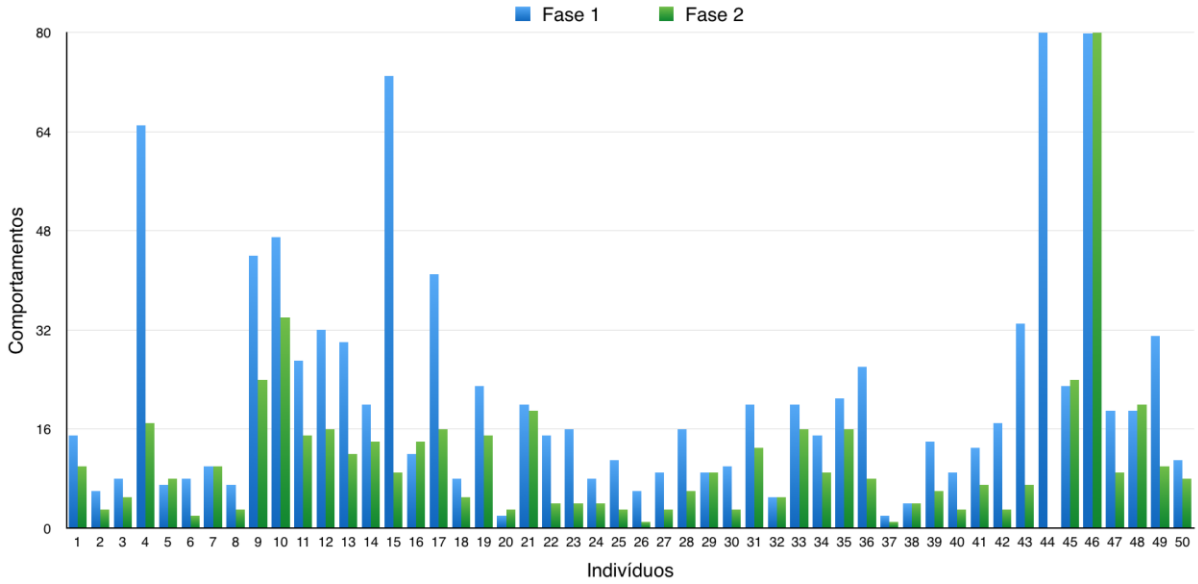

Figura 25: Gráfico Quantidade de Comportamentos Extraídos.

Análises comparativas entre os resultados de indivíduos distintos são complexas devido as condições específicas do movimento de cada um. Em um contexto como esse espera-se que os indivíduos tenham comportamentos distintos que representem seus hábitos. Por isso optou-se pela realização de análises comparativas entre os dados de um mesmo indivíduo em 2 fases distintas. Avaliando por essa forma é possível observar que os indivíduos possuem resultados próximos nas duas fases. Alguns casos específicos como os indivíduos 4, 15, 17, 44 e 46 apresentam valores desproporcionais aos demais. Existem três causas principais que justificam a variação dos resultados observados nesses indivíduos. A primeira delas é decorrente da

grande quantidade de paradas. Indivíduos com muitas paradas aumentam a possibilidade de identificação de um ou mais padrões no movimento entre elas. A segunda causa é referente a baixa quantidade de trajetórias. Caso um indivíduo tenha uma pequena quantidade de trajetórias permite que os padrões sequenciais sejam descobertos com poucas incidências de registros. A terceira condição é referente a própria variação dos resultados entre as fases. Uma variação muito grande resulta em distorções dos valores. Assim, esta fase do estudo de caso permitiu avaliar o aspecto funcional e conceitual do processo desenvolvido garantindo que é possível a utilização dessa abordagem para identificação de comportamentos sequenciais de trajetórias de indivíduos considerando as particularidades específicas do movimento de cada um.

5.4.2 Consistência entre Fases

A segunda etapa do estudo de caso tem como proposta analisar os resultados dos indivíduos avaliando se existe consistência entre os resultados obtidos nas duas fases distintas. O objetivo dessa análise é garantir que o processo como um todo é regular considerando a variação dos resultados como consequência exclusiva do movimento do indivíduo ou de alguma condição de coleta dos dados. Para avaliação deste estudo de caso foi utilizado como critério o módulo da diferença da quantidade de grupos e sequências extraídas nas duas fases para cada indivíduo. Na Figura 26 é possível observar que para maioria dos indivíduos o valor desse módulo da diferença fica próximo do valor de zero o que significa que houve uma pequena variação entre os grupos e sequências identificados. Essa pequena variação representa que os resultados são constantes e que possivelmente que os padrões gerados são comuns e representam os hábitos regulares de movimento dos indivíduos. Alguns indivíduos identificados pelos números 4, 15, 44 e 46 tiveram uma grande variação em relação a quantidade de sequências entre as duas fases e por isso se destacam no gráfico. A provável causa dessa variação é decorrente de uma mudança de comportamento pois os demais indicadores coletados como quantidade de trajetórias, paradas e grupos identificados no período não tiveram grande variação nas fases.

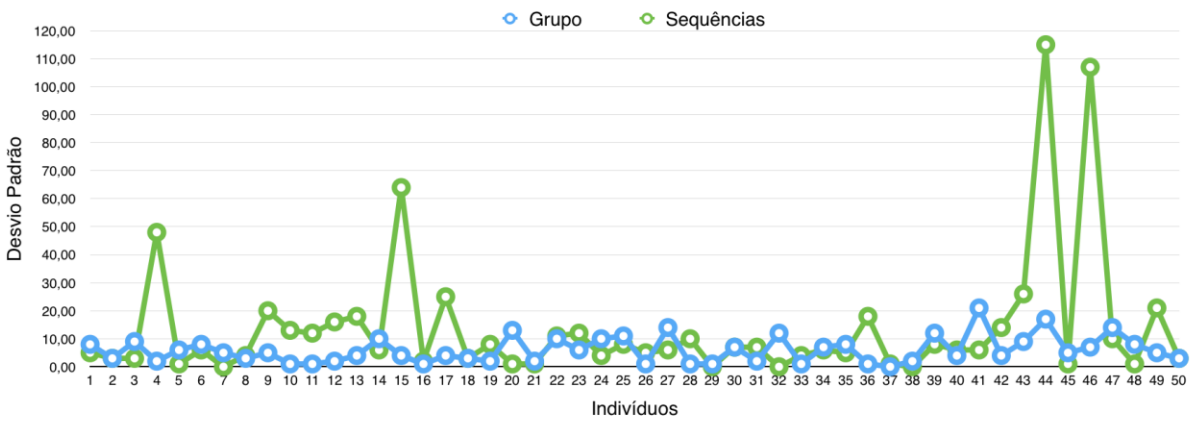

Figura 26: Gráfico Consistência entre Fases.

Embora os resultados dessa análise possam ser impactados pelas particularidades dos movimentos de cada indivíduo, utilizar a variação dos valores entre fases é um bom meio de avaliar a consistência do processo desenvolvido. Esta etapa do estudo de caso permitiu avaliar a consistência do processo possibilitando a identificação de resultados similares em fases distintas. No entanto, essa consistência do processo não garante que o indivíduo manteve exatamente os mesmos comportamentos, ou seja, essa análise apenas garante que nas fases analisadas o indivíduo manteve algumas características comuns. Na próxima fase do estudo de caso será avaliado a consistência desses comportamentos identificados, verificando se eles realmente são comuns em ambas as fases de execução e confirmando ou desmentindo a existência de um padrão do movimento ao longo do tempo.

5.4.3 Consistência dos Comportamentos

Esta fase do estudo de caso tem como objetivo analisar a consistência dos comportamentos descobertos e avaliar se o indivíduo possui comportamentos recorrentes e constantes nas duas fases. Através dessa análise é possível confirmar se o indivíduo ao longo do tempo tem a tendência de manter os mesmos comportamentos e demais características do movimento, como grupos e sequências identificadas. A manutenção de um comportamento ao longo do tempo possibilita análises preditivas sobre o movimento realizado em um determinado dia e por essa razão é grande o interesse por esse tipo conhecimento.

Para avaliação deste estudo de caso foi utilizado um indicador que mede a quantidade de grupos e sequências semelhantes identificados em ambas as fases. Foi utilizado como métrica o índice de Jaccard que é uma medida que indica a proporção de elementos comuns de duas amostras. Essa métrica é calculada dividindo-se o total de elementos comuns, relativos a intersecção entre dois conjuntos, pela soma total dos elementos, relativos a união de dois conjuntos conforme a Figura 27.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Figura 27: Índice de Jaccard.

Na Figura 28 é possível observar o resultado dessa avaliação. Quanto mais próximo de 1 maior a consistência sinalizando que o indivíduo teve um comportamento constante entre as fases. Por outro lado, quanto mais próximo de 0, mais variável é o comportamento do indivíduo. O motivo dessa variação pode ser devido a questões de coleta, questões temporais ocasionados por uma mudança de hábito ou simplesmente pelo fato do indivíduo não possuir movimentos regulares.

Para os indivíduos selecionados não houve nenhum caso de regularidade total considerando os grupos e sequências identificados. Os indivíduos 7, 32 e 38 tiveram os mesmos grupos em ambas as fases sendo os mais constantes nesse aspecto. O indivíduo 44 não foi constante tanto para grupo quanto para sequência. A razão do resultado para esse indivíduo deve-se ao fato de não existir resultados durante a fase 2. Uma outra característica fácil de observar através do gráfico é que para maioria dos indivíduos a consistência das sequências, que representam os movimentos, é maior do que a dos grupos.

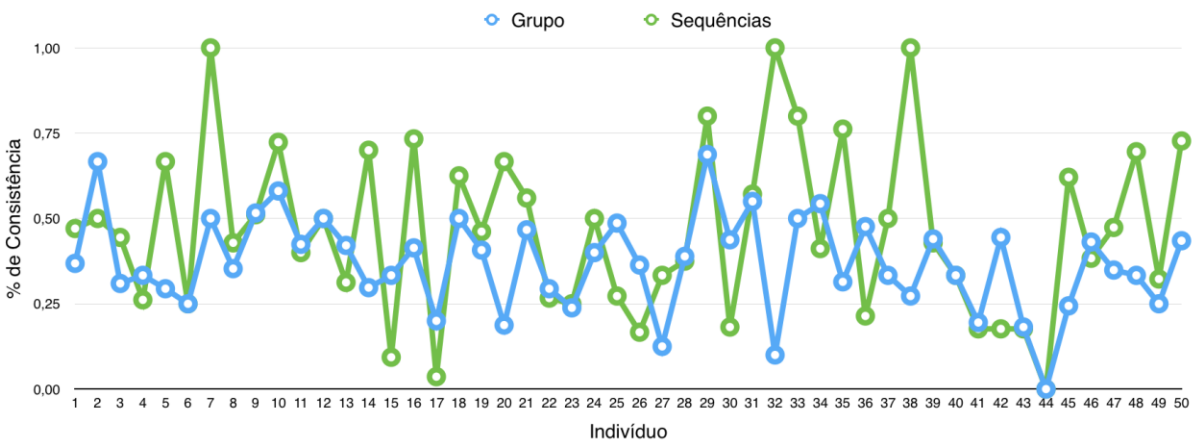

Figura 28: Gráfico Consistência dos Comportamentos.

Para os indivíduos analisados a consistência média dos grupos foi de 37% e das sequências 47%. Dessa forma, através desse estudo de caso foi possível observar que existe uma regularidade na manutenção dos comportamentos. Um pouco mais de um terço dos locais de grande concentração de paradas foram mantidos ao longo das fases, que engloba seis meses de coleta de dados sobre o movimento do indivíduo. A manutenção desses locais refletiu na manutenção das sequências que representam os movimentos entres esses grupos. Para o contexto ao qual essa implementação foi desenvolvida isso significa que esse processo pode ser utilizado para análises sobre o perfil dos indivíduos em decorrência dessa manutenção dessas características do movimento.

5.4.4 Similaridade dos Comportamentos

O estudo de caso anterior analisou o grau de consistência dos indivíduos avaliando os comportamentos recorrentes em ambas as fases, porém a análise realizada não levou em conta a quantidade de incidência desses comportamentos, ou seja, se em uma fase houve 90 incidências de um comportamento e na outra apenas 2, a análise anterior considerava que esse comportamento era consistente apenas em função da recorrência. Esta fase do estudo de caso tem o objetivo de avaliar se os principais comportamentos identificados em ambas as fases são os mais comuns levando em consideração a quantidade de incidência desses comportamentos. Para isso foram ordenados os resultados com os comportamentos encontrados em ambas as fases de forma que os comportamentos com maior incidência estivessem no começo dessa lista.

Dessa forma foi possível avaliar o ranking dos principais comportamentos encontrados nas duas fases.

A Figura 29 apresenta a consistência nas 10 primeiras posições dos rankings e dessa forma avalia os 10 comportamentos com maior incidência em ambas as fases. Quanto mais próximo de 1 mais comum são os comportamentos. O valor médio de consistência foi de 83%, ou seja, dos 10 comportamentos com maior frequência identificados nas fases 8 eram comuns a ambas.

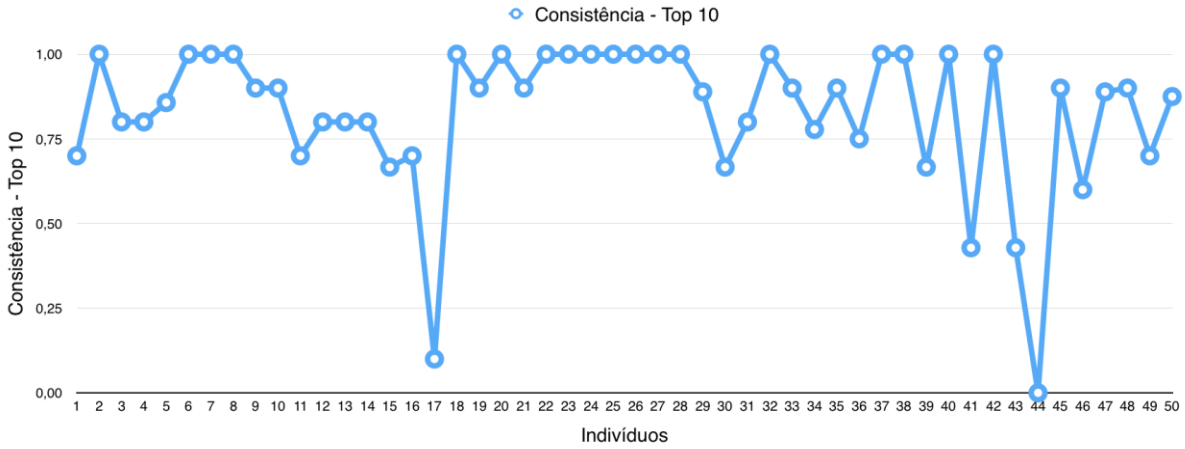

Figura 29: Gráfico Consistência - Ranking 10 elementos.

Esse tipo de análise é útil para confirmação da similaridade dos comportamentos encontrados possibilitando a utilização de técnicas que permite a intersecção em diversos níveis do ranking podendo ser consideradas não apenas os 10 comportamentos mais comuns e com isso podem ser analisados o ranking com os 20, 30, 50 comportamentos mais frequentes dos indivíduos. Utilizar uma medida de similaridade entre rankings também permite a comparação e a classificação entre indivíduos distintos. Quanto maior o índice de similaridade mais constante é o comportamento do indivíduo ao longo do tempo. Esse tipo de estratégia é útil e pode ser utilizado para avaliar a variação dos comportamentos ao longo do tempo pois é natural que novos grupos e sequências sejam criados em função de novos hábitos ou mudanças de movimentos. Esta etapa do estudo de caso permitiu avaliar a similaridade entres os comportamentos mais frequentes identificados em duas fases diferentes de coleta para cada indivíduo. A identificação de comportamentos comuns e com alta frequência de incidência possibilita a seleção de um conjunto que representa os movimentos mais prováveis realizados

por um indivíduo. Esse conjunto com os comportamentos mais prováveis podem ser utilizados para análises preditivas, comparativas ou combinada com outras abordagens para seleção e classificação dos indivíduos.

5.4.5 Aspectos Semânticos

Uma característica importante de ser analisada é referente aos resultados com conteúdo semântico. Dessa forma, esta etapa do estudo de caso tem o objetivo de apresentar alguns resultados com essa formatação semântica utilizando as características dos locais identificados durante a etapa de Agrupamento de Paradas. A apresentação do resultado nesse formato facilita o entendimento sobre os comportamentos identificados permitindo que análises mais estruturadas possam ser realizadas.

A Tabela 5 exibe alguns resultados obtidos ao utilizar as características semânticas para apresentação dos comportamentos descobertos. Para esse estudo de caso foram utilizadas as características semânticas relativas ao tipo do local de parada associado ao grupo e também o dia da semana de ocorrência do comportamento. A Tabela 5 também apresenta a frequência de incidência do comportamento descoberto nos respectivos dias da semana para cada indivíduo. Grupos que não tiveram um local associado estão denotados como "SI" representando os locais como "Sem Identificação". A apresentação dos resultados neste formato facilita o entendimento sobre o comportamento realizado.

Tabela 5: Resultado Apresentado de Forma Semântica.

Indivíduo	Comportamento	Frequência	Dia da Semana
8	Estacionamento → Casa	40%	Dom
8	Parada de Ônibus → Casa	30%	Dom
32	Casa → Casa → Casa	41%	Seg
46	Estacionamento → Parque→ SI → Casa	30%	Dom
46	Estacionamento → Parque → SI	30%	Dom
46	Estacionamento → Parque → Casa	30%	Dom
46	Parque → SI → Casa	30%	Dom
2	Supermercado → Casa	33%	Qua
12	Clínica → Casa	55%	Seg / Ter / Qua / Qui / Sex
48	Centro de Oração	32%	Seg / Qui
1	Posto de Combustível → Casa	63%	Dom
1	Casa →Universidade → Casa	36%	Ter
1	Casa → Casa → Casa	46%	Sab
3	Restaurante → Casa	30%	Seg
3	Banco → Casa	30%	Sex
9	Casa → Fast Food	30%	Qui
9	Fast Food → Casa	30%	Qui
9	Casa → Posto de Combustível	30%	Ter

Um dos grandes problemas identificados neste trabalho envolveu a incerteza semântica dos locais associados às paradas descrito na seção 4.2.5 (Identificador Semântico de Grupos) relativo a utilização de fontes externas para identificação dos locais de paradas. Embora tenha sido identificado um local associado a grande maioria dos grupos descobertos, foi observado uma incerteza sobre a conformidade e exatidão desse local. Um local identificado como uma clínica poderia ser na realidade um posto de gasolina, uma escola ou algum outro local. Esse tipo de incerteza, denominado de incerteza semântica, prejudica a apresentação e interpretação dos resultados. Assim foi realizado uma avaliação dos grupos descobertos com o objetivo de

verificar a possibilidade de existência dentre os grupos identificados algum equivalente ao endereço utilizado pelo indivíduo no momento da aquisição do seguro e com isso realizar um ajuste ou correção eliminando parte das incertezas semânticas.

No contexto de seguro de veículos o endereço informado representa o local de pernoite do veículo e assim é esperado que ao longo do tempo que a quantidade de paradas realizadas nesse local seja suficiente para identificação de um grupo por se tratar de um local de grande relevância para o indivíduo. Também é esperado que esse grupo identificado seja um dos mais densos, ou seja, com maior quantidade de incidências de paradas em virtude da frequência de movimentos realizados tendo como origem ou destino esse local. Assim, nesta análise foi verificado se dentre o conjunto de grupos descobertos existe algum correspondente ao endereço do indivíduo. Para isso foi utilizada a coordenada do endereço do indivíduo e calculada a distância entre esta coordenada e o centróide dos grupos. Caso essa distância fosse inferior a amplitude máxima do grupo então o grupo era considerado o endereço do indivíduo e marcado como residência (Casa).

A coordenada do endereço foi obtida através do processo de geocodificação utilizando a API do Google Maps[4]. O processo de geocodificação permite a obtenção de uma coordenada geográfica equivalente a um endereço pesquisado. Para isso foi feito um processo para envio da requisição e processamento do resultado conforme especificação dessa API. Um exemplo da requisição e resultado obtido pode ser visualizados no Apêndice C.

Do total de 50 indivíduos analisados 16 não tiveram o endereço identificado entre os grupos específicos de cada indivíduo descobertos nas duas fases. Os demais 34 indivíduos foram identificados em um dos grupos gerados na etapa de agrupamento de paradas do processo em ambas as fases analisadas. O motivo da não identificação acontece em função de duas causas principais. A primeira é decorrente de alguma imprecisão ocasionada pela rotina de geocodificação reversa na qual o endereço pode ter sido convertido para uma latitude ou longitude incorreta. A segunda causa ocorre em virtude de alguma inconsistência cadastral utilizada na aquisição do seguro. Não foram realizadas análises individuais para entendimento do motivo da não identificação do grupo equivalente ao endereço desses 16 casos.

Analisando o resultado dos 34 indivíduos que tiveram o endereço identificado apenas o indivíduo 9 teve o endereço associado a um grupo gerado somente na fase 2. Todos os demais

[4] http://maps.google.com/maps

tiveram o endereço identificado durante a fase 1. Além disso, 27 dentre os 34 indivíduos identificados, o grupo identificado como local de endereço ou pernoite era o mais denso representando o local de maior de incidência de parada e relevância para esses indivíduos no período. Consequentemente grande parte dos comportamentos sequenciais extraídos contemplaram esse grupo que representa o local de provável residência do indivíduo. Dessa forma a abordagem proposta pode ser utilizada para casos onde seja necessário a confirmação de perfil utilizando o endereço do indivíduo como forma de consistência. Este estudo de caso permitiu avaliar o processo no aspecto semântico analisando o tipo de resultado que pode ser obtido ao expressar os comportamentos em um formato mais representativo e também análises que podem ser realizadas utilizando esse conteúdo como para confirmação de perfil.

5.5 Conclusão do Capítulo

Neste capítulo foi apresentado a implementação do processo para extração de comportamentos de trajetórias. Através desta implementação foi possível avaliar algumas particularidades comuns do contexto para o qual essa proposta foi desenvolvida e verificar as dificuldades na utilização desse processo.

Em paralelo com a implementação foram executados testes para avaliar a flexibilidade do processo de modo a permitir a utilização de diferentes estratégias para manipulação de conjunto de dados e geração de comportamentos. Essa flexibilidade também ocorre na seleção dos algoritmos de agrupamento e sequenciamento. A utilização do algoritmo de agrupamento DBSCan e do algoritmo de sequenciamento CM-SPADE, ocorreu em virtude da estratégia utilizada para identificação dos comportamentos, da metodologia adotada e do conjunto de dados selecionado. Essa implementação também permitiu observar algumas dificuldades comuns que fazem parte do contexto de utilização dessa proposta referente ao enriquecimento semântico utilizado nos grupos, avaliando o impacto que a base de dados utilizada com esse propósito reflete nos resultados.

Um conjunto de dados foi selecionado de acordo com alguns critérios definidos permitindo a geração de comportamentos que foram avaliados em diferentes análises do estudo de caso. Os comportamentos identificados por essa abordagem são baseados na frequência e

recorrência de movimentos e por essa razão é fundamental que o conjunto de dados contemple as condições dos requisitos levantados.

Para realização do estudo de caso foi utilizada uma metodologia que permitiu a separação do conjunto de dados em 2 fases possibilitando a comparação dos resultados extraídos após processamento realizado em duas fases. Esta metodologia permite acompanhar a evolução dos comportamentos identificados em fases distintas possibilitando avaliar se o indivíduo possui comportamentos que permanecem constante ao longo do tempo. Embora esta metodologia tenha considerado duas fases o mesmo conceito pode ser aplicado por um período maior e dessa forma a avaliação sobre a consistência dos resultados pode ser estendida.

Um estudo de caso foi desenvolvido com o objetivo de avaliar o aspecto funcional da proposta permitindo demonstrar a viabilidade da utilização deste processo para identificação de comportamentos sequenciais utilizando trajetórias. Também foram realizadas análises para garantir que o processo é constante e que dessa forma o fator determinante para variação dos resultados é exclusivo ao conjunto de dados utilizados. Foram realizadas análises para avaliar a consistência dos comportamentos identificados, para avaliar se os indivíduos possuem comportamentos constantes e a relevância desses comportamentos. Algumas análises foram feitas de forma específica para observar os motivos que levaram a algumas variações observadas nos resultados gerados.

Por fim, foi realizada uma análise com o objetivo de demonstrar a possibilidade da utilização de características semânticas para apresentação de resultados que facilitam o entendimento sobre o movimento dos indivíduos. Embora nesta etapa do estudo de caso tenha sido utilizada as características semânticas relativas ao local de parada e ao dia da semana outros conteúdos podem ser utilizados a fim de gerar resultados significativos permitindo a representação de comportamentos em um formato mais completo.

6 Considerações Finais

Este trabalho abordou diversos aspectos relacionados a dados sobre a trajetória de indivíduos englobando questões relativas a coleta, processamento, análise e o entendimento sobre os movimentos. Foram apresentadas algumas abordagens para representação desse tipo de conhecimento utilizando o conceito de trajetórias que contempla algumas características que definem principalmente os movimentos e paradas realizadas pelos indivíduos. Também foi abordado questões relativas ao processamento desse conteúdo e de técnicas utilizadas para adicionar significado as características da trajetória em um processo denominado de enriquecimento semântico. Por fim foram apresentadas abordagens de aprendizado de máquina que permite a extração de conhecimento. Esse conhecimento descoberto define o comportamento do indivíduo sendo composto por um conjunto de características que representam o modo como o indivíduo se comporta durante o movimento.

Neste contexto foi desenvolvido um processo que possibilita a descoberta de padrões sequenciais dos movimentos realizados. Este processo permite a extração de comportamentos de trajetórias que são apresentados em um formato mais rico de significado facilitando a compreensão e permitindo que análises mais profundas possam ser realizadas para o entendimento sobre o movimento.

O processo é baseado na descoberta de padrões sequenciais utilizando agrupamentos de paradas que representam os locais de alta relevância para o indivíduo. A frequência de movimentos entres os agrupamentos descobertos determinam os comportamentos sequenciais mais comuns do indivíduo no período observado. Comportamentos são características peculiares ao indivíduo e por essa razão, para obtenção desses comportamentos, é recomendado a utilização desse processo em conjunto com trajetórias específicas a cada indivíduo.

A proposta de utilizar agrupamento de paradas para identificação dos locais mais relevantes para um indivíduo permite resolver alguns tipos de incertezas existentes para esse tipo de conteúdo composta por uma série de imprecisões técnicas, operacionais e contextual. A estratégia de utilizar o movimento entre esses agrupamentos descobertos aumenta a possibilidade de identificação de padrões de repetição. O conjunto de padrões identificados definem o perfil de movimento do indivíduo. Esses perfis podem ser utilizados para diferentes tipos de abordagem como análises preditivas, avaliações para confirmar ou desmentir um comportamento e também comparações coletivas.

A flexibilidade do processo é a maior contribuição que este trabalho apresenta. Este processo pode ser utilizado durante o processo de enriquecimento semântico ou posteriormente com o objetivo de obtenção dos padrões de comportamento. O processo também possui flexibilidade na seleção dos algoritmos de agrupamento e sequenciamento da mesma forma que permite diferentes estratégias para manipulação do conjunto de dados facilitando a identificação dos padrões. Além disso o processo pode ser executado em fases, como apresentado no estudo de caso, possibilitando a comparação entre os resultados das diferentes execuções e permitindo a avaliação da variação ou manutenção dos comportamentos identificados.

Uma implementação do processo foi desenvolvida com o objetivo de validar o aspecto conceitual do processo como um todo, das etapas e dos componentes que o definem. Através da implementação também foi possível observar algumas dificuldades comuns da utilização desse processo tanto relativo a conexão com outras bases de dados com o objetivo de adicionar mais significado aos resultados como também da necessidade de adição de filtros para eliminação de imprecisões contextuais. Por fim, a implementação permitiu a aplicação do processo utilizando um conjunto de dados reais, obtido no contexto de trajetórias de veículos, a fim de observar os resultados e realizar análises para confirmação das características que definem o perfil e comportamentos identificados.

Este trabalho serviu de inspiração para que melhorias e desenvolvimentos futuros possam ser realizados com o intuito de aprimorar os resultados obtidos. Uma das maiores dificuldades observadas está relacionada a utilização de fontes externas para enriquecimento semântico das paradas. Embora as fontes contemplem um grande conjunto de dados observou-se que esse conjunto é incompleto e parte dos dados é obsoleto. Dessa forma é preciso a utilização de alguma outra abordagem em conjunto para minimizar esse tipo de imprecisão. A incerteza semântica, ocasionada em função dessa limitação, acaba restringindo análises que possam utilizar essas características para obtenção de novos conhecimentos sobre os comportamentos identificados.

Os resultados apresentados neste trabalho são relativos aos comportamentos mais comuns identificados para os indivíduos. Esses comportamentos precisam ser persistidos para que consultas utilizem esse conteúdo.

Da mesma forma, o processo pode ser evoluído para utilização de outras características relativas ao movimento para obtenção de comportamentos que contemple aspectos temporais

e outros aspectos contextuais do domínio de aplicação. Os principais comportamentos que definem um perfil podem ser utilizados de forma comparativa. Para isso, é necessário a utilização de abordagens que permitam avaliar a consistência desses perfis e evolução dos mesmos ao longo do tempo. Essa abordagem precisa levar em conta as mudanças de comportamento que naturalmente podem ocorrer.

Por fim, como trabalhos futuros, espera-se aprimorar as etapas do processo proposto e utilizá-lo em diferentes contextos para a extração de conhecimento que represente os principais comportamentos relativo ao movimento dos indivíduos analisados.

Referências

ANDRIENKO, N.; ANDRIENKO, G. Designing visual analytics methods for massive collections of movement data. **Cartographica**, v. 42, n. 2, p. 117–138, 2007.

ALVARES, L. O.; LOY, A. M.; RENSO, C.; BOGORNY, V. An algorithm to identify avoidance behavior in moving object trajectories. **Journal of the Brazilian Computer Society**, v. 17, n. 3, p. 193–203, 2011.

BAGLIONI, M.; DE MACÊDO, J. A.; RENSO, C.; TRASARTI, R.; WACHOWICZ, M. Towards Semantic Interpretation of Movement Behavior. In: SESTER, M.; BERNARD, L.; PAELKE, V. (Ed.). Advances in GIScience: Proceedings of the 12th AGILE Conference. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 271–288, 2009.

BISHOP, C. M. **Pattern Recognition and Machine Learning.** Springer, 2007.

BOGORNY, V.; RENSO, C.; AQUINO, A. R.; SIQUEIRA, F. L. S.; ALVARES, L. O. CONSTAnT - A Conceptual Data Model for Semantic Trajectories of Moving Objects. **Transactions in GIS**, v. 18, n. 295179, p. 66-88, 2013.

BOGORNY, V. Análise de dados de Movimento: você já pensou que está sendo monitorado?. ERBD (Escola Regional de Banco de Dados), 2016. Disponível em: <http://cross.dc.uel.br/erbd2016/anais-e-slides/>. Acesso em: 17 jul. 2016.

BREITMAN, K. **Web Semântica.** A internet do futuro. Rio de Janeiro, Brasil: LTC, 2005. 190 p.

DENTLER, K.; CORNET, R.; TEIJE, A. T.; KEIZER, N. Comparison of reasoners for large ontologies in the OWL 2 EL profile. **Semantic Web**. v. 42, n. 2, p. 71–87, 2011.

DODGE, S.; WEIBEL, R.; LAUTENSCHUTZ, A. K. Taking a systematic look at movement: Developing a taxonomy of movement patterns. **AGILE Workshop on GeoVisualization of Dynamics, Movement and Change**. p. 1–10, 2008.

ESTER, M.; KRIEGEL, H.; XU, X.; MIINCHEN, D.-. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, 1996.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. **Advances in Knowledge Discovery & Data Mining**. 1 ed. American Association for Artificial Intelligence, Menlo Park, Califórnia, 1996. 611 folhas.

FILETO, R.; KRÜGER, M.; PELEKIS, N.; THEODORIDIS, Y.; RENSO, C. Baquara: A Holistic Ontological Framework for Movement Analysis Using Linked Data. In: NG, W.; STOREY, V. C.; TRUJILLO, J. C. (Ed.). **Conceptual Modeling: 32th International Conference, ER 2013, Hong-Kong, China, November 11-13, 2013. Proceedings**. Berlin, Heidelberg: Springer Berlin Heidelberg, p. 342–355, 2013.

FILETO, R.; MAY, C.; RENSO, C.; PELEKIS, N.; KLEIN, D.; THEODORIDIS, Y. The Baquara2 knowledge-based framework for semantic enrichment and analysis of movement data. **Data & Knowledge Engineering**, v. 98, p. 104–122, 2015.

FOURNIER-VIGER, P.; GOMARIZ, A.; CAMPOS, M.; THOMAS, R. Fast Vertical Mining of Sequential Patterns Using Co-occurrence Information. p. 40–52, 2014.

FOURNIER-VIGER, P.; LIN, J. C. A Survey of Sequential Pattern Mining. v. 1, n. 1, p. 54–77, 2017.

GIANNOTTI, F.; NANNI, M.; PINELLI, F.; PEDRESCHI, D. Trajectory pattern mining. **Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. ACM Press, p. 330–339, 2007.

GRUBER, T. R. A translation approach to portable ontology specifications. **Knowledge Acquisition**, v. 5, p. 199–220, 1993.

HADAJ, S. El. Modeling, Mining, and Analyzing Semantic Trajectories: The Process to Extract Meaningful Behaviors of Moving Objects. v. 124, n. 8, p. 15–21, 2015.

ILARRI, S.; STOJANOVIC, D.; RAY, C. Semantic management of moving objects: A vision towards smart mobility. **Expert Systems with Applications**, v. 42, n. 3, p. 1418–1435, 2015.

LACY, L. **OWL: Representing information using the Web Ontology Language**. Victoria, Canada: Trafford, p. 285, 2005.

LAUBE, P; IMFELD, S. Analyzing relative motion within groups of trackable moving point objects. **Proceedings of the 2nd International Conference on Geographic Information Science (GIScience'02)**. Springer, v. 2478, p. 132–144, 2002.

LAUBE, P.; IMFELD, S.; WEIBEL, R. Discovering relative motion patterns in groups of moving point objects. **International Journal of Geographical Information Science**, v. 19, n. 6, p. 639–668, 2005.

LAUBE, P. **Progress in movement analysis.** In Behaviour Monitoring and Interpretation. IOS Press, v. 3, p. 43–71, 2009.

LAUSCH, A.; SCHMIDT, A.; TISCHENDORF, L. Data mining and linked open data - New perspectives for data analysis in environmental research. **Ecological Modelling**, v. 295, p. 5–17, 2015. Disponível em: <http://dx.doi.org/10.1016/j.ecolmodel.2014.09.018>.

LI, Q.; ZHENG, Y.; XIE, X.; CHEN, Y.; LIU, W.; MA, W.-Y. Mining User Similarity Based on Location History. In: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, New York, NY, USA. **Anais**... New York, NY, USA: ACM, 2008.

MITCHELL, T. **Machine learning**. McGraw Hill, 1997.

MONREALE, A.; PINELLI, F.; TRASARTI, R. WhereNext: a Location Predictor on Trajectory Pattern Mining. p. 637–645, 2009.

MONREALE, A.; TRASARTI, R.; PEDRESCHI, D.; RENSO, C.; BOGORNY, V. C-safety: a Framework for the Anonymization of Semantic Trajectories. **Transactions on Data Privacy**, v. 4, n. 2, p. 73–101, 2011.

NANNI, M.; KUIJPERS, B.; KÖRNER, C.; MAY, M.; PEDRESCHI, D. Spatiotemporal Data Mining. In: GIANNOTTI, F.; PEDRESCHI, D. (Ed.). **Mobility, Data Mining and Privacy: Geographic Knowledge Discovery**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. p. 267–296.

PARENT, C.; SPACCAPIETRA, S.; RENSO, C.; ANDRIENKO, G.; ANDRIENKO, N.; BOGORNY, V.; DAMIANI, M. L.; GKOULALAS-DIVANIS, A.; MACEDO, J.; PELEKIS, N.; THEODORIDIS, Y.; YAN, Z. Semantic Trajectories Modeling and Analysis. **ACM Computing Surveys**, v. 45, n. 4, p. 42:1–42:32, 2013.

SRIKANT, R.; AGRAWAL, R. Mining Sequential Patterns: Generalizations and Performance Improvements. p. 1-17, 1996.

SPACCAPIETRA, S.; PARENT, C.; DAMIANI, M. L.; MACEDO, J. A.; PORTO, F.; VANGENOT, C. A conceptual view on trajectories. **Data & Knowledge Engineering**, v. 65, n. 1, p. 126–146, 2008.

SONG, R.; SUN, W.; ZHENG, B.; ZHENG, Y. PRESS: A Novel Framework of Trajectory Compression in Road Networks. **Proc. VLDB Endow.** v. 7, n. 9, p. 661–672, maio 2014. Disponível em: <http://dx.doi.org/10.14778/2732939.2732940>.

THÉRIAULT, M.; CLARAMUNT, C.; VILLENEUVE, P. A spatio-temporal taxonomy for the representation of spatial set behaviours. **Proceedings of the International Workshop on Spatio-Temporal Database Management (STDBM'99)**. Springer, v. 1678, p. 1–18, 1999.

WOOD, Z.; GALTON, A. A taxonomy of collective phenomena. **Applied Ontology**, v. 4, n. 3, p. 267–292, 2009a.

WOOD, Z.; GALTON, A. Classifying collective motion. In: **Behavior Monitoring and Interpretation – BMI**. IOS Press, v. 3, p. 129–155, 2009b.

YAN, Z.; PARENT, C.; MACEDO, J.; SPACCAPIETRA, S. Trajectory ontologies and queries. **Transactions in GIS**, v. 12, n. 1, p. 75–91, 2008.

YAN, Z. Towards semantic trajectory data analysis: A conceptual and computational approach. **Proceedings of the VLDB 2009. PhD Workshop**, p. 1-6, 2009.

YAN, Z.; PARENT, C.; SPACCAPIETRA, S.; CHAKRABORTY, D. A hybrid model and computing platform for spatio-semantic trajectories. **Proceedings of the 7th Extended Semantic Web Conference (ESWC'10**), v. 1, p. 60–75, 2010.

YAN, Z.; CHAKRABORTY, D.; PARENT, C.; SPACCAPIETRA, S.; ABERER, K. SeMiTri: A framework for semantic annotation of heterogeneous trajectories. **Proceedings of the 14th International Conference on Extending Database Technology**, p. 259–270, 2011.

ZHANG, C.; HAN, J.; SHOU, L.; LU, J.; LA PORTA, T. Splitter: Mining Fine-grained Sequential Patterns in Semantic Trajectories. Proc. VLDB Endow., v. 7, n. 9, p. 769–780, maio 2014. Disponível em: <http://dx.doi.org/10.14778/2732939.2732949>.

ZHENG, Y.; XIE, X. Learning travel recommendations from user-generated GPS traces. **ACM Transaction on Intelligent Systems and Technology,** jan. 2011.

ZHENG, Y. U. Trajectory Data Mining: An Overview. **ACM Trans. Intell. Syst. Technol.**, v. 6, n. 3, p. 1–41, 2015.

Apêndice A

Resultados da Fase 1.

Individuo	Paradas	Trajetórias	Grupos	Outliers	Sequências
1	346	79	17	73	15
2	125	49	9	30	6
3	409	87	23	105	8
4	520	88	29	106	65
5	361	83	19	99	7
6	430	79	24	166	8
7	172	55	8	52	10
8	477	82	13	279	7
9	611	91	21	89	44
10	520	91	25	57	47
11	546	90	23	70	27
12	569	92	28	113	32
13	472	87	29	142	30
14	448	87	19	114	20
15	187	42	14	57	73
16	410	81	21	147	12
17	26	15	1	13	41
18	472	91	18	214	8
19	459	87	18	81	23
20	102	62	3	38	2
21	295	66	10	68	20
22	333	89	17	61	15
23	205	49	16	46	16
24	206	82	9	29	8
25	371	86	22	117	11
26	213	57	7	106	6

27	161	51	11	44	9
28	339	76	12	116	16
29	294	79	14	89	9
30	193	58	15	41	10
31	409	73	32	73	20
32	774	91	39	101	5
33	354	92	10	57	20
34	545	92	32	175	15
35	459	92	19	85	21
36	253	48	16	65	26
37	103	70	4	53	2
38	268	81	13	50	4
39	330	88	12	60	14
40	315	70	18	133	9
41	233	49	14	57	13
42	180	44	11	75	17
43	454	74	24	133	33
44	301	35	17	85	115
45	486	78	23	124	23
46	756	92	38	86	187
47	405	86	22	118	19
48	401	76	16	86	19
49	489	71	20	122	31
50	277	78	18	69	11
Total	**18064**	**3691**	**893**	**4569**	**1209**
Média	**361**	**74**	**18**	**91**	**24**

Apêndice B

Resultados da Fase 2.

Individuo	Paradas	Trajetórias	Grupos	Outliers	Sequências
1	292	79	9	66	10
2	81	35	6	28	3
3	375	72	32	123	5
4	411	83	27	96	17
5	415	88	25	95	8
6	266	74	16	114	2
7	183	53	13	44	10
8	183	52	10	95	3
9	572	89	26	92	24
10	461	89	24	76	34
11	531	89	24	52	15
12	519	88	26	112	16
13	356	74	25	122	12
14	463	86	29	121	14
15	160	32	10	54	9
16	328	82	20	98	14
17	70	20	5	27	16
18	381	90	21	148	5
19	444	89	20	67	15
20	230	63	16	57	3
21	289	72	12	49	19
22	349	88	27	99	4
23	122	45	10	41	4
24	257	74	19	46	4
25	362	82	33	116	3

26	198	69	8	97	1
27	237	57	25	72	3
28	247	62	13	87	6
29	287	85	13	87	9
30	84	30	8	32	3
31	319	63	30	81	13
32	600	87	27	124	5
33	353	89	11	51	16
34	457	87	39	148	9
35	442	85	27	95	16
36	191	44	15	55	8
37	83	63	4	39	1
38	351	87	15	56	4
39	389	89	24	80	6
40	258	61	14	107	3
41	577	89	35	184	7
42	217	53	15	98	3
43	242	62	15	108	7
44	0	0	0	0	0
45	576	84	28	91	24
46	733	90	45	98	80
47	486	87	36	129	9
48	339	66	24	71	20
49	378	77	25	132	10
50	237	63	15	74	8
Total	**16381**	**3517**	**996**	**4234**	**540**
Média	**328**	**70**	**20**	**85**	**11**

Apêndice C

Exemplo de pesquisa de endereço para obtenção da latitude e longitude usando a API do GoogleMaps.

Requisição:
http://maps.google.com/maps/api/geocode/json?address=RUA%20DEZENOVE,%20450,%20Consolacao,%20RIO%20CLARO,%20SP&sensor=false

Retorno:


```
{
   "results" : [
      {
         "address_components" : [
            {
               "long_name" : "450",
               "short_name" : "450",
               "types" : [ "street_number" ]
            },
            {
               "long_name" : "Rua 19",
               "short_name" : "R. 19",
               "types" : [ "route" ]
            },
            {
               "long_name" : "Consolação",
               "short_name" : "Consolação",
               "types" : [ "political", "sublocality", "sublocality_level_1" ]
            },
            {
               "long_name" : "Rio Claro",
               "short_name" : "Rio Claro",
               "types" : [ "administrative_area_level_2", "political" ]
            },
            {
               "long_name" : "São Paulo",
               "short_name" : "SP",
               "types" : [ "administrative_area_level_1", "political" ]
            },
            {
               "long_name" : "Brasil",
               "short_name" : "BR",
               "types" : [ "country", "political" ]
            },
            {
               "long_name" : "13503-300",
               "short_name" : "13503-300",
               "types" : [ "postal_code" ]
            }
         ],
         "formatted_address" : "R. 19, 450 - Consolação, Rio Claro - SP, 13503-300, Brasil",
         "geometry" : {
            "location" : {
               "lat" : -22.4200539,
               "lng" : -47.5721826
            },
            "location_type" : "ROOFTOP",
            "viewport" : {
               "northeast" : {
                  "lat" : -22.4187049197085,
                  "lng" : -47.5708336197085
               },
               "southwest" : {
                  "lat" : -22.42140288029151,
                  "lng" : -47.5735315802915
               }
            }
         },
         "place_id" : "ChIJBQAAOPHax5QR2GRWJl79FXc",
         "types" : [ "establishment", "point_of_interest" ]
      }
   ],
   "status" : "OK"
}
```