MÉTODO DE SENSORIAMENTO SOCIAL PARA CARACTERIZAÇÃO E DETECÇÃO DE EVENTOS URBANOS: UMA APLICAÇÃO EM ACIDENTES DE TRÂNSITO

ALICE ADATIVA FERREIRA MENEZES

MÉTODO DE SENSORIAMENTO SOCIAL PARA CARACTERIZAÇÃO E DETECÇÃO DE EVENTOS URBANOS: UMA APLICAÇÃO EM ACIDENTES DE TRÂNSITO

Dissertação apresentada ao Programa de Pós-Graduação em Informática do Instituto de Computação da Universidade Federal do Amazonas como requisito parcial para a obtenção do grau de Mestre em Informática.

Orientador: Carlos Maurício Seródio Figueiredo

Manaus

Março de 2017

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

M543m

Menezes, Alice Adativa Ferreira
Método de sensoriamento social para caracterização e detecção de eventos urbanos : Uma aplicação em acidentes de trânsito / Alice Adativa Ferreira Menezes. 2017
69 f.: il. color; 31 cm.

Orientador: Carlos Maurício Seródio Figueiredo
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. sensoriamento social. 2. redes sociais baseadas em localização. 3. acidentes de trânsito. 4. processamento de linguagem natural. I. Figueiredo, Carlos Maurício Seródio II. Universidade Federal do Amazonas III. Título

FOLHA DE APROVAÇÃO

"Método de Sensoriamento Social para Caracterização e Detecção de Eventos Urbanos: Uma aplicação em Acidentes de Trânsito"

ALICE ADATIVA FERREIRA MENEZES

Dissertação de Mestrado defendida e aprovada pela banca examinadora constituída pelos Professores:

Prof. Carlos Maurício Seródio Figueiredo - PRESIDENTE

Profa. Fabiola Guerra Nakamura - MEMBRO INTERNO

Profa. Elloá Barreto Guedes da Costa - MEMBRO EXTERNO

Manaus, 24 de Março de 2017

Resumo

Acidentes de trânsito são um problema recorrente nas áreas urbanas, causando prejuízos, danos físicos e materiais. Atualmente, existem diversos órgãos públicos e privados que incentivam a criação de soluções que ajudem a minimizar a ocorrência destes acidentes em áreas urbanas. Neste sentido, apresentamos uma solução que utiliza os conceitos de Sensoriamento Social para o monitoramento e a caracterização de acidentes de trânsito. Sensoriamento Social é um novo paradigma no qual é realizado um processo distribuído de coleta de dados sociais, através de pessoas que compartilham dados contextuais voluntariamente. Como estudo de caso, aplicamos a solução para o monitoramento e a caracterização do trânsito em áreas urbanas, pois os habitantes ali presentes compartilham um grande número de informações em redes sociais. Além disso, em algumas áreas existem bases de dados oficiais, disponibilizadas pelo governo, as quais podem ser utilizadas para validação da solução proposta. A solução considera as limitações do Sensoriamento Social e os experimentos utilizam tanto dados públicos oficiais quanto dados sociais provenientes do Twitter e do Foursquare. Os resultados obtidos mostram que, para os cenários avaliados, torna-se possível a utilização de redes sociais como um meio alternativo de monitoramento e caracterização de acidentes de trânsito.

Palavras-chave: Sensoriamento social, redes sociais, redes sociais baseadas em localização, acidentes de trânsito, processamento de linguagem natural.

Abstract

Traffic accidents are a recurrent problem in urban areas, causing damages and injuries. Currently, there are several public and private entities that encourage the creation of solutions that help to minimize the occurrence of these accidents in urban areas. In this way, we present a solution, which uses the concepts of Social Sensing for monitoring and characterization of traffic accidents. Social sensing is a new paradigm in which is performed a distributed process of collecting social data, through people sharing contextual data voluntarily. As a case study, we applied the solution for monitoring and characterization of traffic in urban areas, as their inhabitants share a large number of information on social networks. Furthermore, there are official databases related to the city, made available by the government, which can be used for validation of the proposed solution. The solution considers the limitations of Social Sensing and the experiments use both official public data and social data from Twitter and Foursquare. The results show that, for the scenarios evaluated, it becomes possible to use social networks as an alternative of traffic accidents monitoring and characterizing.

Keywords: Social sensing, social networks, location-based social networks, traffic accidents, natural language processing.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Um dos conceitos que vem ganhando destaque recentemente para o monitoramento do trânsito é o Sensoriamento Social [Silva et al., 2014a]. No Sensoriamento Social, os usuários de redes sociais online, como Waze, Foursquare e Instagram podem compartilhar dados a respeito do ambiente ou do contexto em que se encontram por meio dos *smartphones*. Neste tipo de sensoriamento, podemos citar como desafio a modelagem, visualização e análise dos dados coletados. Esses dados, podem ser de grande importância para entender os padrões de comportamento urbano dos habitantes de uma determinada localidade de forma rápida e com baixo custo.

Dentre as principais aplicações do Sensoriamento Social, podemos citar as de monitoramento de tráfego através de redes sociais para auxiliar na segurança dos condutores, passageiros e pedestres. Essas aplicações tornaram-se de grande importância devido aos danos causados por acidentes de trânsito ao longo dos últimos anos.

A Organização das Nações Unidas (ONU), chamou a atenção para os números alarmantes relacionados aos acidentes de trânsito no mundo e, em 2010, definiu que entre o período de 2011 a 2020 diversos países deveriam atingir uma meta de estabilizar e reduzir as mortes causadas pelo trânsito [de Morais Neto et al., 2012]. Já o Departamento de Transporte dos Estados Unidos criou um programa para incentivar tanto a indústria quanto a academia a desenvolver sistemas seguros com o objetivo de prevenir acidentes de trânsito. Este programa envolve montadoras, como a General Motors (GM), e instituições de ensino, como a Universidade da Califórnia [Figueiredo et al., 2001].

O objetivo deste trabalho surge da necessidade de monitorar o trânsito e de caracterizar áreas urbanas, possibilitando a detecção de acidentes e de vias não seguras. A partir disto, torna-se possível a definição de soluções para minimizar a ocorrência de acidentes desta natureza.

1.1 Problema

Para prover soluções para os acidentes de trânsito, é preciso primeiramente, analisar dados históricos e estudar os fatores relacionados a estes acidentes. Porém, em alguns perímetros urbanos, como nas cidades brasileiras, a falta de informações necessárias ou até mesmo a disparidade de dados, podem dificultar este processo [CNM, 2009]. Na Tabela 1.1, podemos observar este aspecto entre os anos de 2002 a 2008, nos quais notamos que fontes de informação distintas fornecem dados estatísticos diferentes para o mesmo dado.

O monitoramento por tecnologias tradicionais ou manuais, faz com que a coleta de dados relevantes torne-se custosa, o que pode dificultar a busca de soluções. Neste sentido, o Sensoriamento Social pode fornecer as informações necessárias para um problema estudado, com baixo custo de coleta de dados [Karamshuk et al., 2013].

Dentre os pontos positivos de buscar a solução para este problema com a utilização de Sensoriamento Social, podemos citar as dinâmicas semelhantes entre cidades [Silva et al., 2014a]. Neste caso, estudando e provendo soluções para cidades cujas fontes de dados são mais precisas, é possível aplicar estas soluções na demais cidades, as quais possuem dinâmicas semelhantes.

1.2 Motivação Social

Anualmente ocorrem cerca de 1,3 milhões de mortes por acidentes de trânsito nos países de média e baixa renda, que abrigam 84,5% da população mundial. Além disso, o número de pessoas lesionadas está entre 20 e 50 milhões [de Morais Neto et al., 2012]. Outro dado importante vem da Organização Pan-Americana de Saúde (OPS), que aponta que 6% das deficiências físicas no mundo são causadas por acidentes de trânsito [de Oliveira, 2010].

O Brasil é um dos países com o maior número de mortes no trânsito no mundo, ocupando o 5º lugar e sendo precedido pela Índia, China, EUA e Rússia [de Morais Neto et al., 2012]. Apesar disso, o fato do país investir pouco na criação de órgãos especializados em coleta e análise de dados relativos à segurança no trânsito vem sendo criticado, pois isto contribui para que não existam dados estatísticos e fiéis o suficiente para o desenvolvimento de estratégias de intervenção adequadas e concretas [CNM, 2009]. Além disso, diversos estudos mostram a importância da criação de políticas e mecanismos para prover a segurança no trânsito em território nacional.

1.3 Justificativa

A Confederação Nacional de Municípios [CNM, 2009] apresenta uma análise dos dados sobre mortes no trânsito em estados e municípios brasileiros. Os dados do estudo foram coletados das bases de dados do DENATRAN[1], SUS[2] e DPVAT[3], e mostram um pico histórico nas mortes por acidentes de trânsito em 2007, com 66.836 mortes (Tabela 1.1).

O estudo também aponta que 80% das mortes ocorrem com pessoas do sexo masculino entre 20 e 39 anos e que moram em cidades de pequeno e médio porte, o que pode ser justificado pelo fato de que muitos jovens ainda conduzem seus veículos sob o efeito de álcool ou drogas.

Ano	DENATRAN	SUS	DPVAT
2002	18.877	32.753	-
2003	22.629	33.139	-
2004	25.526	35.105	-
2005	26.409	35.994	55.024
2006	19.910	36.367	63.776
2007	-	37.407	66.836
2008	-	-	57.116

Tabela 1.1: Registro de mortes no trânsito na Brasil [CNM, 2009].

O estudo apresentado por de Morais Neto et al. [2012] tem como objetivo analisar a tendência temporal da mortalidade por acidentes de transportes terrestres e identificar a existência e a localização de aglomerados de alto risco de mortes. Todos os acidentes cujos dados foram coletados, ocorreram entre 2000 e 2010 e resultaram em uma taxa de mortalidade por acidente de trânsito entre 18 e 22,5 óbitos/100 mil habitantes. Vale ressaltar que as taxas calculadas foram padronizadas por idade, para Unidades Federadas (UF) e municípios por parte populacional.

O estudo também chama a atenção para o fato de que é necessária uma atuação do governo, da sociedade civil e dos cidadãos para que o número de acidentes de trânsito diminua com o passar dos anos, tendo em vista que países de média e baixa renda possuem 91,5% das mortes e que os países de alta renda respondem apenas por 8,5%.

Por fim, o estudo apresentado por de Oliveira [2010] discorre a respeito dos custos que os acidentes de trânsito geram para o Brasil. São cerca de 20 mil dólares por ferido

[1]Departamento Nacional de Trânsito
[2]Sistema Único de Saúde
[3]Danos Pessoais Causados por Veículos Automotores de Via Terrestre

grave com uma média de internação de 20 dias, tendo em vista que dois terços dos leitos hospitalares dos setores de ortopedia e traumatologia são ocupados por essas vítimas.

O estudo também mostra outros problemas causados pelo trânsito brasileiro, como os problemas de saúde da população que podem vir a ocorrer devido à poluição causada pelos veículos. No estado de São Paulo, por exemplo, a Secretaria de Estado dos Transportes Metropolitanos estima que as perdas financeiras com acidentes de trânsito, poluição e engarrafamentos sejam de 4,1 bilhões de reais por ano.

A partir desses relatos, percebemos a necessidade da criação de políticas de segurança no trânsito, assim como a criação de mecanismos que possam auxiliar o monitoramento de vias a fim de diminuir os acidentes de trânsito, responsáveis por mortes e lesões no Brasil e no mundo.

1.4 Objetivo Geral

O objetivo desta dissertação é projetar e avaliar um método que utiliza Sensoriamento Social para análise do perfil social urbano, frente ao padrão de mobilidade, para a monitoramento e caracterização de acidentes de trânsito.

1.4.1 Objetivos Específicos

Os objetivos específicos são descritos a seguir:

1. Verificar a viabilidade do uso de redes sociais para o monitoramento e a caracterização de acidentes de trânsito, considerando a possibilidade de auxílio às ferramentas de monitoramento tradicionais;

2. Propor e avaliar um método de correlação do perfil social de uma área urbana com características de acidentes de trânsito;

3. Propor e avaliar um método de extração de dados em redes sociais que permita monitorar ocorrência de acidentes de trânsito;

4. Demonstrar, através de estudos de caso, os diversos cenários em que as soluções baseadas em redes sociais podem ser utilizadas para monitorar e caracterizar dinamicamente os acidentes de trânsito.

1.5 Organização da Dissertação

Esta dissertação está organizada da seguinte forma: no Capítulo 2, apresentamos os fundamentos teóricos necessários para o entendimento dos métodos adotados. No Capítulo 3, apresentamos uma síntese dos trabalhos relacionados, expondo as características e particularidades dos métodos existentes. No Capítulo 4 é apresentado o método proposto. Nos Capítulos 5 e 6, são apresentados os resultados obtidos. Por fim, no Capítulo 7, apresentamos as conclusões para este trabalho.

Capítulo 2

Fundamentação Teórica

Neste capítulo são apresentados os conceitos prévios necessários para o entendimento do trabalho. A fundamentação teórica é dividida em quatro seções: na Seção 2.1, são abordados os conceitos relacionados ao Sensoriamento Social, assim como as redes sociais que serão utilizadas no desenvolvimento deste trabalho. Na Seção 2.2 é feita uma descrição de como o compartilhamento de informações, através de redes sociais, pode influenciar na detecção da rotina dos usuários. A Seção 2.3, apresenta um resumo dos conceitos de aprendizagem de máquina utilizados neste trabalho. Por fim, na Seção 2.4, encontram-se as considerações finais deste capítulo.

2.1 Sensoriamento Social

Com a evolução dos celulares, hoje chamados de *smartphones*, tornou-se possível a detecção de dados relacionados aos usuários, como as suas opiniões e sua localização [Silva et al., 2014a]. A introdução de sensores como GPS, acelerômetro, microfone, câmera e giroscópio nestes dispositivos, permitiram o sensoriamento de diversas áreas, em sua maioria urbanas [Lane et al., 2010]. Este tipo de sensoriamento, que ocorre através da coleta de dados de redes sociais, é denominado Sensoriamento Social e consiste em um usuário e seu dispositivo móvel compartilhando informações do contexto em que se encontra [Silva et al., 2014a].

A Figura 2.1, apresenta uma visão geral deste sensoriamento, mostrando que os usuários são entidades móveis autônomas capazes de sensoriar todo o ambiente no qual estão inseridos [Silva et al., 2014a]. Assim, é possível adquirir informações relacionadas a uma determinada localidade através de dados de redes sociais e de dispositivos móveis utilizados por estes usuários. Desta forma, a coleta de dados através de redes sociais

permite identificar condições diversas como o trânsito, o clima, os eventos naturais, os aspectos econômicos, entre outros fatores.

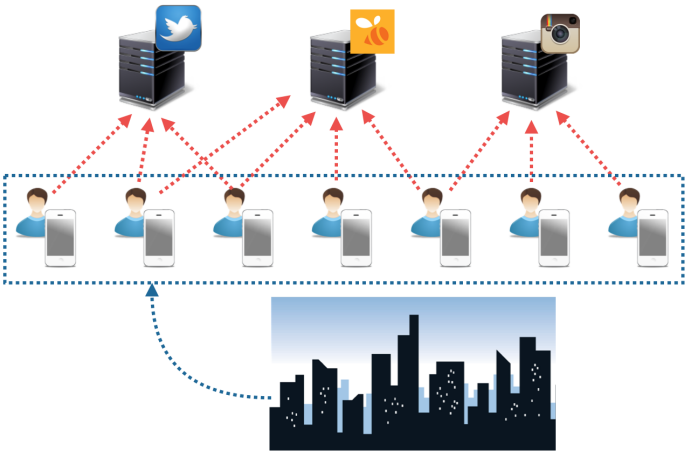


Figura 2.1: Visão geral do Sensoriamento Social.

Outro aspecto importante do Sensoriamento Social é que os dados podem ser coletados para análise assim que o usuário os disponibiliza nas redes sociais. Isto permite a criação de soluções que provêem informações em tempo hábil para a tomada de decisões.

Podemos citar como desafio deste tipo de sensoriamento, a modelagem, visualização e análise dos dados coletados. Esses dados, provenientes de sensores sociais, podem ser de grande importância para entender os padrões de comportamento urbano dos habitantes de uma determinada localidade de forma rápida e com baixo custo [Ribeiro et al., 2014]. Os dados relacionados a alertas de trânsito tem sido amplamente estudados, com o objetivo de diminuir os danos e os prejuízos causados por fatores como os engarrafamentos e acidentes.

2.1.1 Redes Sociais

Nesta subseção, são apresentadas as redes sociais utilizadas nos estudos de caso deste trabalho. Estas redes, são amplamente utilizadas na literatura para o sensoriamento social, conforme será apresentado no capítulo 3.

2.1.1.1 Twitter

O Twitter[1] é uma rede social para *microblogging*, na qual os usuários podem enviar e receber atualizações através de mensagens de até 140 caracteres, chamadas de *tweets*.

[1]https://www.twitter.com

Esta rede social, lançada em Julho de 2006, atualmente possui cerca de 320 millhões de usuários ativos mensalmente[2].

A Figura 2.2, apresenta um exemplo de *tweet*, que é composto de informações como o nome e a identificação do usuário, uma mensagem, e a localização do usuário no momento em que a mensagem foi publicada na rede social. Vale ressaltar que a localização do usuário não é uma informação obrigatória e pode estar ausente em alguns *tweets*.

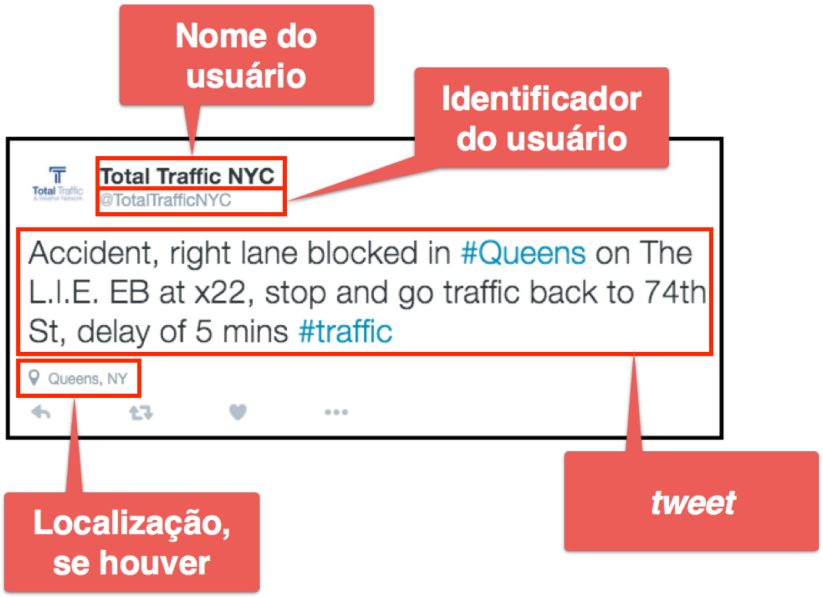


Figura 2.2: Exemplo de *tweet*.

O Twitter é a principal fonte de dados de sensoriamento, pois é através dele que é feita a coleta de dados de *tweets* e de outras redes sociais, como o Foursquare e o Instagram. A coleta de dados pode ser realizada através de duas APIs: *Search* e *Streaming*, que são explicadas em detalhes no Capítulo 4.

2.1.1.2 Foursquare

O Foursquare[3] é uma rede social baseada em localização, cujo principal serviço é o *check-in*, um evento que é sensoriado quando a hora e o local de um usuário são coletados [Silva et al., 2014a].

A Figura 2.3, apresenta um exemplo de *check-in*, que é composto de informações como o nome e a localização do usuário, o local e a data em que o *check-in* foi realizado. Diferentemente do Twitter, a localização é obrigatória nos dados provenientes do Foursquare.

[2]https://about.twitter.com/pt/company
[3]https://www.foursquare.com

Figura 2.3: Exemplo de *check-in*.

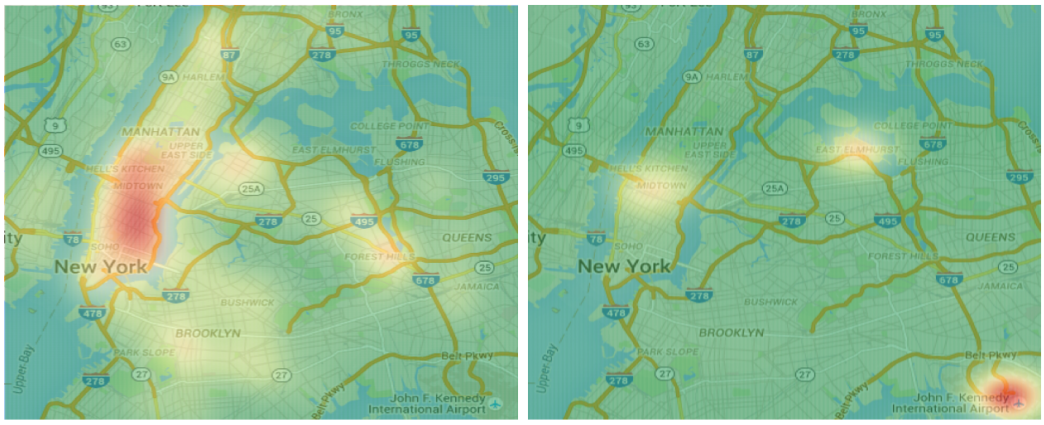
Os locais cadastrados no Foursquare são associados a uma das 10 categorias atualmente existentes na rede social. Essas categorias[4] são: *Arts and Entertainment*, *College and University*, *Event*, *Food*, *Nightlife Spot*, *Outdoors and Recreation*, *Professional and Other Places*, *Residence*, *Shop and Service* e *Travel and Transport*. Alguns exemplos de locais associados a estas categorias são apresentados na Tabela 2.1.

Categoria do Foursquare	Exemplos de locais
Arts and Entertainment	Cinemas, museus e casinos.
College and University	Escolas, laboratórios universitários e centros de estudo.
Food	Restaurantes, cafeterias e padarias.
Nightlife Spot	Bares, clubes de *rock* e boates.
Outdoors and Recreation	Parques, academias e praias.
Professional and Other Places	Fábricas, auditórios e centros médicos.
Shop and Service	Lojas, salões de beleza e supermercados.
Travel and Transport	Aeroportos, estradas e hotéis.
Residence	Propriedades privadas, prédios e reboques residenciais.
Events	Conferências, convenções e festivais.

Tabela 2.1: Exemplos de locais existentes nas categorias do Foursquare.

A importância destas categorias está relacionada a detecção de pontos específicos em uma determinada localidade (Figura 2.4), assim como o reconhecimento da dinâmica das cidades e da mobilidade urbana.

[4]https://developer.foursquare.com/categorytree



(a) Categoria: *Residence* (b) Categoria: *Travel and Transport*

Figura 2.4: Concentração de dados de acordo com as categorias do Foursquare.

2.1.1.3 Waze

O Waze[5] é um aplicativo de trânsito e navegação baseado em uma comunidade. Com o aplicativo, é possível compartilhar informações de trânsito das vias em tempo real através de alertas, com o objetivo de fazer todos economizarem tempo e combustível em seus deslocamentos diários. Em 2013, foi anunciado que o Waze possuia 50 milhões de usuários.

Assim como outras redes sociais, os dados do Waze são coletados através do Twitter, tendo em vista que o mesmo não possui uma API própria de coleta de dados. Apesar dos dados do Waze serem geolocalizados, os dados provenientes do Twitter relacionados ao Waze não apresentam dados de geolocalização. Desta forma, para saber a localização de um alerta, é preciso extrair os dados do texto do *tweet* e utilizar uma API de geolocalização para obter a latitude e a longitude.

2.1.2 Camadas de sensoriamento

O conceito de Camadas de Sensoriamento, foi apresentado formalmente por Silva et al. [2014b] como dados que descrevem aspectos específicos de uma determinada localização geográfica. Os dados que devem ser utilizados nas Camadas de Sensoriamento devem ser coletados, processados, analisados e padronizados.

Cada Camada de Sensoriamento (Figura 2.5) possui os seguintes atributos: data e hora em que o dado foi criado, localização, dados de especialização (foto, local, alerta de trânsito, condição climática, etc.), identificação do usuário e estado da camada (ativa ou inativa).

[5]https://www.waze.com

As Camadas de Sensoriamento permitem que os estudos relacionados à dinâmica das cidades possam conter dados mais robustos, tendo em vista que os dados das diferentes camadas se complementam.

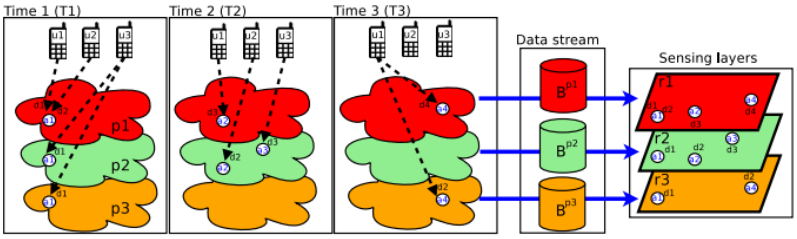


Figura 2.5: Ilustração do compartilhamento de dados em três fontes de sensoriamento ao longo do tempo, resultando nas camadas de dados [Silva et al., 2014b].

Nos estudos de caso apresentados neste trabalho, as camadas de sensoriamento são utilizadas de forma que os dados de diferentes fontes de sensoriamento estejam separados para uma análise posterior.

2.2 Rotinas e o Compartilhamento de Dados

Utilizando os dados do Foursquare, é possível fazer uma análise da rotina dos usuários. Neste cenário, vemos que algumas categorias possuem uma quantidade relevante de *check-ins*, enquanto outras categorias não possuem (Figura 2.6). Isto pode ocorrer devido a fatores como a privacidade. Uma das categorias, por exemplo, é referente ao local de moradia dos usuários. Em decorrência disso, a frequência de dados relacionados a esta categoria é baixa. Outro fator importante diz respeito a popularidade de determinados lugares, que é o caso da categoria *Events*, cuja quantidade de *check-ins* também é baixa.

Através das categorias do Foursquare é possível identificar a rotina e o comportamento humano em uma determinada localidade. Na Figura 2.7, por exemplo, fazemos a comparação entre duas categorias: *Nightlife Spot* e *Professional and Other Places*. Verificando a quantidade de *check-ins* ao longo de uma semana, percebe-se um padrão na rotina dos usuários, visto que dados do Foursquare relacionados à categoria *Nightlife Spot* são mais frequentes durante o final de semana. O oposto ocorre com a categoria *Professional and Other Places*, que é mais frequente durante a semana.

O comportamento dos usuários pode mudar de uma localidade para outra, alterando as informações provenientes dos dados de redes sociais. Isto ocorre devido a um fator cultural [Silva et al., 2014b].

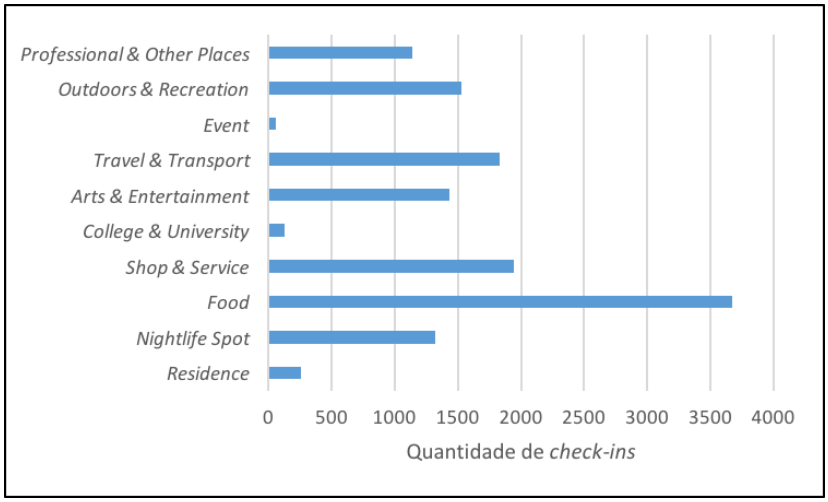


Figura 2.6: Quantidade de *check-ins* durante o período de uma semana.

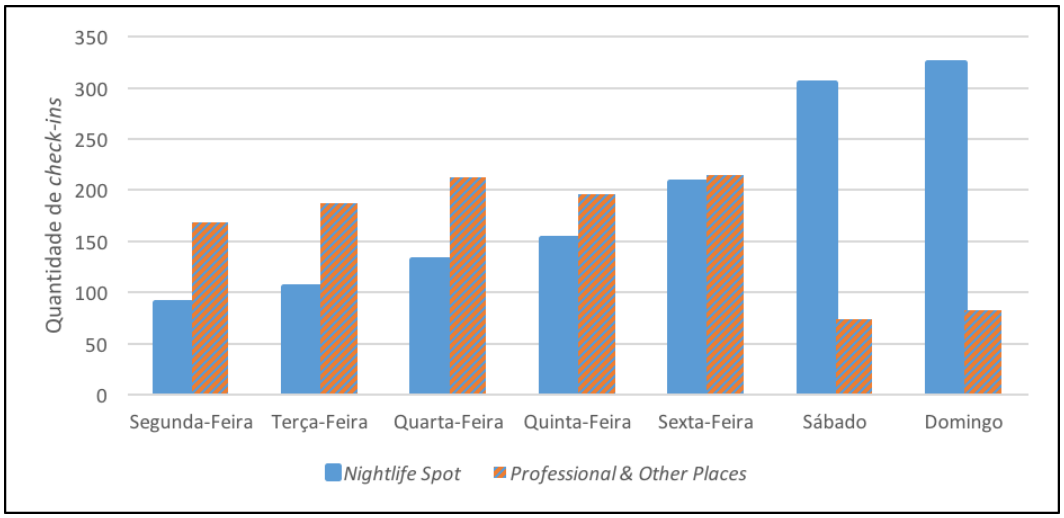


Figura 2.7: Quantidade de *check-ins* por dia da semana.

2.3 Monitoramento do Trânsito

Na literatura existem diversos modelos que lidam com eventos relacionados ao trânsito, sendo que estes modelos utilizam conceitos de aprendizagem de máquina [Silva et al., 2015]. Aprendizagem de máquina diz respeito a fazer computadores modificarem ou adaptarem suas ações a fim de torná-las mais precisas [Marsland, 2009]. A precisão das ações tomadas são medidas através da comparação destas com as ações consideradas corretas.

Algoritmos de aprendizagem de máquina podem ser classificados como supervisionados (o algoritmo aprende através de um conjunto de dados de entrada que contém a resposta correta) e não-supervisionados (não existem dados de entrada, o algoritmo

apenas organiza os dados através da sua similaridade) [Marsland, 2009].

Nesta seção, apresentamos o DBScan, algoritmo não-supervisionado utilizado em um dos estudos de caso da solução proposta neste trabalho. Além disso, também apresentamos um algoritmo de extração de características de textos, utilizado para prover os dados necessários para o DBScan.

2.3.1 Extração de Características de Textos

No Sensoriamento Social, alguns dados são provenientes de textos, como os dados dos *tweets*. Desta forma, é necessário utilizar algoritmos de extração de características em textos, de forma que dados textuais possam ser utilizados posteriormente por outros algoritmos, como os algoritmos de aprendizagem de máquina, por exemplo.

Neste trabalho, utilizamos o algoritmo *Term Frequency–Inverse Document Frequency* (TF-IDF) para a extração de características em textos, que significa Frequência do Termo–Inverso da Frequência nos Documentos. O TF-IDF, verifica a importância de um documento em relação a um conjunto de documentos, através de medidas estatísticas. Para isto, o algoritmo computa dois valores para cada palavra:

- ***Term Frequency* (TF)**: calcula a frequência de um termo em um documento, dividindo o número de vezes que uma palavra aparece em um documento pelo número total de palavras do documento;
- ***Inverse Document Frequency* (IDF)**: obtém a importância de um termo em relação a toda a amostra, através do cálculo logarítmico da razão do número total de documentos, pelo número de documentos que contém o termo.

Como exemplo, podemos considerar que, em um documento de 1.000 palavras, a palavra "acidente" ocorre 5 vezes. Assim:

$$TF(acidente) = \frac{5}{1.000} = 0.005$$

Agora, supondo que existem 500 mil documentos na amostra e que a palavra "acidente" aparece em 5 mil documentos, temos:

$$IDF(acidente) = \log \frac{500.000}{5.000} = 2$$

Logo, o TF-IDF para a palavra "acidente" é:

$$TF - IDF(acidente) = TF(acidente) * IDF(acidente) = 0.005 * 2 = 0.01$$

O resultado final do algoritmo é uma matriz esparsa, que contém os valores de TF-IDF para cada palavra da amostra.

2.3.2 DBScan: *Density Based Spatial Clustering of Applications with Noise*

O DBScan foi apresentado por Ester et al. [1996] como um algoritmo de agrupamento de dados, projetado para descobrir grupos de dados a partir de uma amostra, separando o ruído das mesmas. Como este algoritmo é baseado na teoria do vizinho mais próximo, é necessário informar como parâmetro a distância mínima que o algoritmo deve considerar para agrupar os dados, além da quantidade mínima de dados em um grupo.

A teoria do vizinho mais próximo, baseia-se no fato de que instâncias dentro de um conjunto de dados, geralmente existem em estreita proximidade com outros casos que possuem propriedades similares [Cover and Hart, 1967]. Neste contexto, o algoritmo do DBScan, busca agrupar dados que possuem alta densidade, ao passo que trata como ruído os dados cuja densidade é baixa.

Na Figura 2.8, apresentamos um exemplo do funcionamento deste algoritmo. Podemos verificar que o ponto R foi considerado um ruído por não possuir pontos próximos. Os pontos A e B são considerados pontos de borda, pois estão com uma densidade moderada em relação aos demais pontos. Os demais pontos são considerados pontos centrais, pois neles há uma densidade considerável.

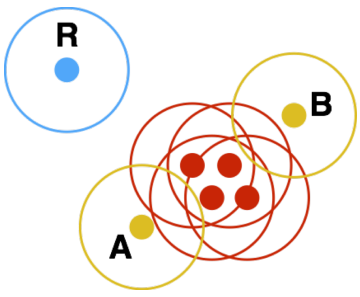


Figura 2.8: Funcionamento do DBScan, no qual o ponto R é um ruído, os pontos A e B são pontos de borda e os demais pontos são centrais.

2.4 Considerações Finais

Neste capítulo foram apresentados os fundamentos teóricos necessários para o desenvolvimento deste trabalho. Os métodos apresentados nos próximos capítulos, foram definidos a partir destes conceitos. No próximo capítulo, são apresentados os trabalhos relacionados, cuja teoria também está relacionada aos conceitos apresentados neste capítulo.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, é apresentado um resumo das soluções recentes relacionadas ao Sensoriamento Social e ao monitoramento do trânsito. Os trabalhos abordam a detecção de padrões em áreas urbanas, pontos de interesse e eventos. O monitoramento mediante a utilização de Sensoriamento Social enfrenta alguns desafios, como a limitação na etapa de coleta de dados, que em sua maioria é realizada indiretamente por meio do Twitter.

3.1 Pontos de Interesse

No trabalho proposto por Lee and Sumiya [2010], os autores apresentam um modelo de pesquisa para detecção de eventos locais, utilizando experiências coletivas de uma população através de dados compartilhados no Twitter.

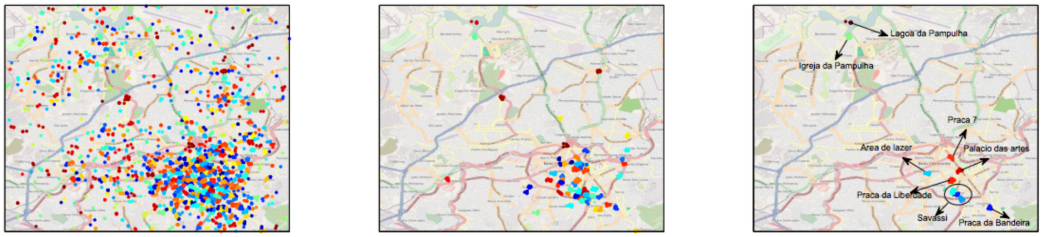
O modelo proposto consiste em 3 etapas:

1. Coleta de dados geo-localizados através de um sistema de monitoramento;
2. Identificação de regiões de interesse (RoIs) dos usuários do Twitter, através de medidas de regularidades do comportamento populacional;
3. Detecção de eventos geo-sociais, através da comparação entre medidas de regularidades.

Como experimento, os autores fizeram uma lista de 15 eventos festivos que iriam ocorrer no Japão, para verificar se o modelo proposto conseguiria determinar a existência desses eventos. Como resultado, foram coletados 9 eventos festivos (60% de acurácia). Porém, através do modelo, também foi possível detectar a ocorrência de eventos naturais que ocorreram no mesmo período.

Já no trabalho proposto por Silva et al. [2013b], os autores utilizaram o Instagram, visando mostrar os desafios e as oportunidades que emergem do Sensoriamento Social realizados através dos usuários. Para os experimentos, foram coletadas 2,3 milhões de fotos do Instagram utilizando o Twitter, o que mostra a abrangência da rede. Entre 30 de junho e 31 de julho de 2012, foram coletados 2.272.556 *tweets* contendo fotos geo-referenciadas, postadas por 482.629 usuários. Cada *tweet* é composto de coordenadas (latitude e longitude) e o horário de compartilhamento da foto.

Através dos dados coletados, foi possível construir uma aplicação que identifica os pontos de interesse em uma determinada região (Figura 3.1).



(a) Todos os clusters (b) Pontos de interesse (c) Pontos turísticos

Figura 3.1: Pontos de Interesse de Belo Horizonte [Silva et al., 2013b].

Em outro trabalho, a abordagem apresentada por Frias-Martinez et al. [2012], utiliza dados geo-localizados que levam em consideração a privacidade do usuário. O método proposto identifica automaticamente as atividades mais comuns entre os usuários de uma área específica, através da padronização dos dados provenientes do Twitter. O método também identifica Pontos de Interesse automaticamente, sendo que estes possuem uma grande quantidade de *tweets* relacionados a eles. Para estudo de caso, foram utilizados dados da cidade de Nova Iorque em um período de 49 dias.

3.2 Regiões Funcionais e Análise da Dinâmica das Cidades

Os dados de redes sociais tornaram possível a análise da dinâmica das cidades para a caracterização de usuários e detecção de regiões funcionais, que determinam as características de uma área [Cranshaw et al., 2012]. Neste sentido, no trabalho de Noulas et al. [2011a], é apresentado o primeiro estudo sobre comportamento humano com dados provenientes do Foursquare. Os autores coletaram aproximadamente 12.000.00 *check-ins* de usuários da rede social. É apresentada uma análise da dinâmica geo-temporal de atividades coletivas de usuários, a fim de demonstrar como *check-ins* provêem meios

para descoberta de padrões de comportamento humano diário e semanal, propriedades urbanas de vizinhanças e transições recorrentes entre diferentes atividades.

Em outro trabalho, Noulas et al. [2011b] demonstra como informações semânticas sobre localidades e atividades sociais podem ser exploradas no contexto de aplicações móveis e pesquisa científica. Os autores propõe o uso da categoria de lugares para criação de "impressões digitais" de usuários e localidades. Desta forma, pessoas podem ser caracterizadas a partir dos tipos de lugares que elas visitam, enquanto regiões podem ser modeladas de acordo com seus locais. Para tanto, os autores utilizam um algoritmo de agrupamento espectral [Shi and Malik, 1997] para agrupar localidades e usuários.

No trabalho de Cranshaw et al. [2012], os autores apresentam uma nova metodologia para o estudo da dinâmica, estrutura e caraterísticas de uma cidade em larga escala. Para isso, eles utilizaram uma aborgagem na qual, a partir de uma quantidade massiva de dados geolocalizados provenientes do Foursquare, foi desenvolvido um algoritmo que mapeia áreas geográficas distintas de uma cidade fornecendo um visão em tempo real de qualquer mundaça no padrão de atividade da população. Tal algoritmo utiliza técnicas de agrupamento espectral e leva em consideração tanto a proximidade espacial dos usuários, quanto a proximidade social.

No trabalho de Silva et al. [2012], os autores introduzem um nova técnica de visualização de dados, chamada *City Image*, cujo objetivo consiste em auxiliar o entendimento sobre a dinâmica de cidades. Para validar sua técnica, os autores utilizaram como base de dados 4,7 milhões de *check-ins* do Foursquare de usuários de oito cidades diferentes. O resultado de tal técnica é uma matriz quadrada que sumariza a dinâmica de cidades. Tal matriz é gerada com base em grafos de transição, modelados a partir da mobilidade de usuários de acordo com as categorias de localidades da rede social Foursquare.

Em Zhang et al. [2013], os autores argumentam que a informação temporal referente à mobilidade humana é de extrema importância para análises urbanas, e portanto, deve ser considerada em conjunto com informações espaciais. Os autores utilizaram dados do Foursquare, provenientes das cidades de Nova Iorque e São Franciso, para analizar a influência da dinâmica temporal de atividades no estudo de mobilidade humana.

Em comparação com o presente trabalho, os trabalhos citados anteriormente não apresentam um estudo que relaciona as regiões funcionais e as atividades de seus habitantes, apesar de utilizarem dados socias.

No artigo de Zhi et al. [2016], os autores introduziram um novo modelo para detecar regiões funcionais baseado no modelo LRA (*Low Rank Approximation*). Para

tanto, utilizaram uma base de dados de aproximadamente 15 milhões de *check-ins* provenientes da cidade de Shangai. Foram identificadas estruturas espaço-temporais latentes que, quando analisadas, reveleram uma série de associações entre as atividades espaciais e temporais de cidadãos. Por meio do algoritmo K-means, os autores obtiveram 5 tipos de clusters que estavam diretamente relacionados com a combinação temporal de atividades dos cidadãos. Desta forma, os autores destacaram a correlação direta entre grupos de regiões e diferentes atividades durantes períodos variados no decorrer no dia.

O trabalho citado determina funções para as regiões diante 5 aspectos, enquanto que este trabalho baseia-se nas categorias do Foursquare. Além disso, nosso método permite que as funções das regiões possam ser analisadas tanto em conjunto quanto de forma individual. Por fim, ressaltamos que este trabalho apresenta possíveis aplicações para a utilização das regiões funcionais.

Existem ainda trabalhos que utilizam dados de dispositivos móveis para determinar a característica das regiões. Este processo é mais custoso e a mobilidade dos usuários não está explicitamente ligada a uma atividade. No trabalho de Xiang et al. [2013], os autores realizam partições de regiões por meio da análise da trajetória de cidadãos utilizando um modelo de tópicos latentes (LDA), no qual são consideradas distribuições de probabilidades ao particionar regiões. A base de dados utiliza contém 500 milhões de chamadas de celulares provenientes de cidadãos da China. Ao invés de utilizar algoritmos de agrupamento baseados em similaridade espacial, os autores consideram informações temporais a fim de determinar regiões funcionais.

O trabalho aqui apresentado diferencia-se dos anteriores por levar em consideração o cotidiano dos habitantes da região estudada, mostrando que as regiões funcionais podem alterar sua função ao longo do tempo, permitindo que os aspectos da dinâmica das cidades seja analisado de forma espaço-temporal.

Autor	Rede social utilizada	Regiões Funcionais
Xiang et al. 2013	Não utiliza	Sim
Noulas et al. 2011a	Foursquare	Não
Noulas et al. 2011b	Foursquare	Não
Silva et al. 2012	Foursquare	Não
Zhang et al. 2013	Foursquare	Não
Cranshaw et al. 2012	Foursquare	Sim
Zhi et al. 2016	Foursquare	Sim

Tabela 3.1: Trabalhos relacionados às regiões funcionais, descritos nesta seção.

3.3 Detecção de Eventos em Redes Sociais

Uma das áreas relacionadas ao Sensoriamento Social é a detecção de eventos, que estuda os fenômenos do mundo real reportados por meio de usuários de redes sociais e que contém informação espaço-temporal [Valkanas and Gunopulos, 2013].

Nesta área, alguns estudos utilizam agrupamento de dados para detectar eventos não especificados, como notícias recentes [Sankaranarayanan et al., 2009] ou situações de desastre natural [Toriumi and Baba, 2016]. Outros estudos utilizam a abordagem de agrupamento para diferenciar eventos do mundo real de mensagens não relacionadas a eventos [Becker et al., 2011].

Sankaranarayanan et al. [2009] propôs um sistema chamado *TwitterStand*, que captura *tweets* relacionados a notícias recentes. Na solução, um classificador baseado em Naive Bayes é responsável por separar notícias de outras informações. Em seguida, um algoritmo de agrupamento forma *clusters* de notícias utilizando *hashtags* para reduzir erros de agrupamento.

Phuvipadawat and Murata [2010] apresentam um método para o rastreamento de notícias recentes do Twitter utilizando palavras-chave em na coleta dos dados. Em seguida, mensagens com termos similares, *hashtags* e nomes de usuários são agrupados. A solução considera o número de seguidores e o número de *retweetes* para ranqueamento dos *clusters*.

Becker et al. [2011] propuseram uma técnica de agrupamento em redes sociais que assimila *tweets* através do TF-IDF. Posteriormente, os *clusters* são classificados em eventos do mundo real e não-eventos através de um algoritmo de *Support Vector Machine* (SVM).

Existem também estudos que analisam apenas um tipo de evento utilizando aprendizagem supervisionada. Nestes estudos, é necessário que informações preliminares sobre os eventos sejam disponibilizadas. Além disso, se o tipo de evento mudar, novas informações devem ser adquiridas para a solução.

Sakaki et al. [2010] propuseram um modelo para detectar um tipo específico de evento. Os autores treinaram um algoritmo de SVM através de dados do Twitter rotulados manualmente com informações de terremotos e tufões. Em seguida, estimaram a trajetória dos tufões aplicando Filtros de Kalman e Filtros de Partícula.

Nguyen et al. [2016] apresentaram um sistema chamado *TrafficWatch*, que coleta, filtra e analisa *tweets* relacionados a incidentes na Austrália. O objetivo era utilizar as redes sociais como um canal adicional de monitoramento do trânsito e gerenciador de incidentes. Para isto, os autores aplicaram processamento de linguagem natural para extrair informações dos *tweets*. Em seguida, foram utilizados os algoritmos de SVM e

Árvores de Decisão para definir eventos relevantes e não relevantes.

Autor	Rede social utilizada	Tipo de aprendizagem	Evento estudado
Sankaranarayanan et al. (2009)	Twitter	Não-Supervisionada	Notícias
Phuvipadawat e Murata (2010)	Twitter	Não-Supervisionada	Notícias
Becker et al. (2011)	Twitter	Supervisionada	Eventos do mundo real
Sakaki et al. (2010)	Twitter	Supervisionada	Terremotos e Tufões
Nguyen et al. (2016)	Twitter	Supervisionada	Trânsito

Tabela 3.2: Trabalhos relacionados à detecção de eventos, descritos nesta seção.

3.4 Considerações Finais

O Sensoriamento Social está mudando a forma como vemos as cidades, a sociedade e a interação entre as pessoas. Nesse tipo de rede, os usuários de redes sociais, como Waze, Foursquare e Instagram podem compartilhar dados a respeito do ambiente ou do contexto em que se encontram através dos *smartphones*.

Podemos citar como desafio deste tipo de sensoriamento a coleta, a modelagem, visualização e análise dos dados coletados. Esses dados, provenientes de sensores sociais, podem ser de grande importância para entender os padrões de comportamento urbano dos habitantes de uma determinada localidade, de forma rápida e com baixo custo. Os dados relacionados a alertas de trânsito tem sido amplamente estudados, com o objetivo de diminuir os danos e os prejuízos causados por fatores como os engarrafamentos e acidentes.

Capítulo 4

Método Proposto

Neste capítulo é descrito o método proposto, que envolve a caracterização e detecção de eventos urbanos. Além disso, a descrição da arquitetura é feita desde a coleta de dados até a etapa de extração de características.

A visão geral do método proposto, apresentada na Figura 4.1, propõe uma coleta e filtragem de dados de redes sociais, adquiridos por intermédio do Twitter. Em seguida serão formadas as camadas de dados, para que posteriormente os dados possam ser utilizados em conjunto ou individualmente para gerar informações e resultados. Por fim, serão utilizados métodos para a extração de características de eventos em áreas urbanas.

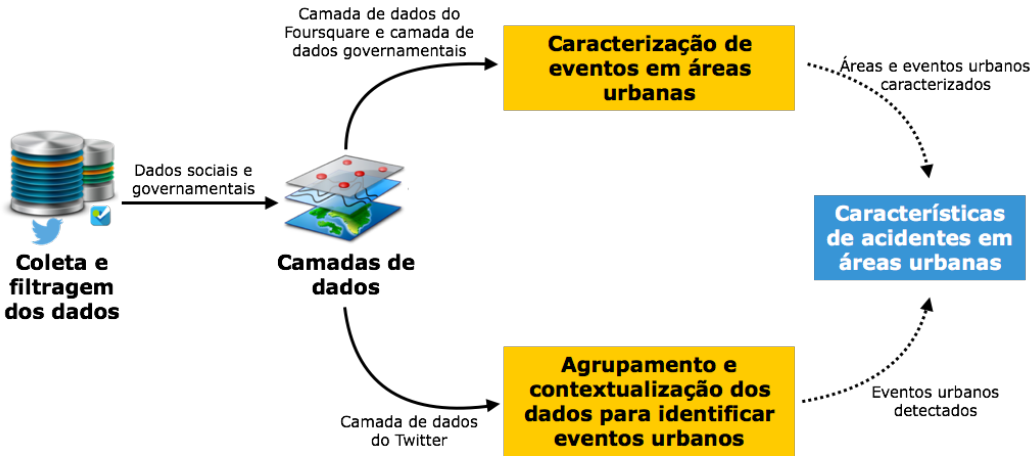


Figura 4.1: Visão geral da arquitetura da solução proposta.

4.1 Coleta e Filtragem de Dados

Os dados que serão utilizados neste método podem vir de diversas fontes, sendo que as principais são as redes sociais online. Outras fontes de dados podem ser as bases de departamentos do governo ou APIs, como as APIs de dados climáticos, por exemplo.

Em relação ao Sensoriamento Social, todos os dados são coletados pela API do Twitter utilizando *querys* de busca, mesmo que estes dados pertençam a outras redes sociais. Para buscar por dados do Foursquare, por exemplo, é utilizada a *query* de busca "swarmapp.com". Já para a coleta de dados do Waze é utilizada a *query* de busca "@waze". Vale ressaltar que nesta API apenas dados públicos podem ser coletados.

Atualmente, o Twitter possui duas APIs de coleta de dados, sendo elas[1]:

- ***Streaming* API**: coleta de dados em tempo real, sem limite de *tweets* na coleta;
- ***Search* API**: coleta de dados históricos de até uma semana, com um limite de 2.700 *tweets* a cada 15 minutos.

Além da coleta de dados gerais através das APIs apresentadas acima, também é possível coletar *tweets* de um usuário específico, utilizando o método *user timeline*[2], que coleta os 3.200 *tweets* mais recentes de um usuário.

Para a utilização destas APIs, é necessário cadastrar uma aplicação no Twitter para obter a chave de acesso OAuth[3], composta por 4 subchaves: *consumer key*, *consumer secret*, *access token* e *access token secret*. O mesmo procedimento é válido para a utilização das APIs de outras fontes de dados sociais, como o Foursquare e o Instagram.

Ressaltamos que todos os dados coletados do Twitter estão com o horário padronizado na zona UTC. Sendo assim, caso o horário do *tweet* seja necessário para o método, é necessário converter o horário para a região que será analisada. A cidade de Nova Iorque, por exemplo, encontra-se em uma região cuja zona é UTC-5. Assim, é necessário reduzir em 5 horas o horário de todos os dados coletados.

Após a coleta dos dados, será realizada a filtragem ou limpeza dos mesmos. Esta etapa se faz necessária devido ao fato de que os dados podem possuir algum tipo de anomalia que invalide a sua utilização pelos métodos presentes na solução deste trabalho. Dentre as anomalias, podemos destacar:

- *Check-ins* cujas URLs foram coletadas, mas durante a extração de dados viu-se que os mesmos foram excluídos;

[1] https://dev.twitter.com/
[2] https://dev.twitter.com/rest/reference/get/statuses/user_timeline
[3] https://dev.twitter.com/oauth/overview/application-owner-access-tokens

- *Check-ins* cujas contas dos usuários foram excluídas ou tornaram-se privadas;
- *Tweets* cujos conteúdos não correspondem ao problema estudado;
- Alertas do Waze que não pertencem as áreas urbanas estudadas.

Por fim, os dados serão agrupados em camadas, para que possam ser combinados de forma que gerem informações que ajudem na solução do problema abordado.

4.2 Caracterização de Acidentes em Áreas Urbanas

Para a caracterização de acidentes de trânsito, propomos um método genérico que detecta regiões funcionais em uma área urbana da seguinte forma (figura 4.2):

1. Definição de dados sociais geolocalizados, com funções relacionadas a eles;
2. Definição de uma grade com um limite definido sobre uma determinada área urbana;
3. Associação de dados sociais com as células da grade, de forma que cada célula possua diversas funções;
4. Ranqueamento das funções em cada célula, fazendo com que apenas a principal função prevaleça;
5. Verificação e agrupamento de regiões adjacentes que possuam a mesma função.

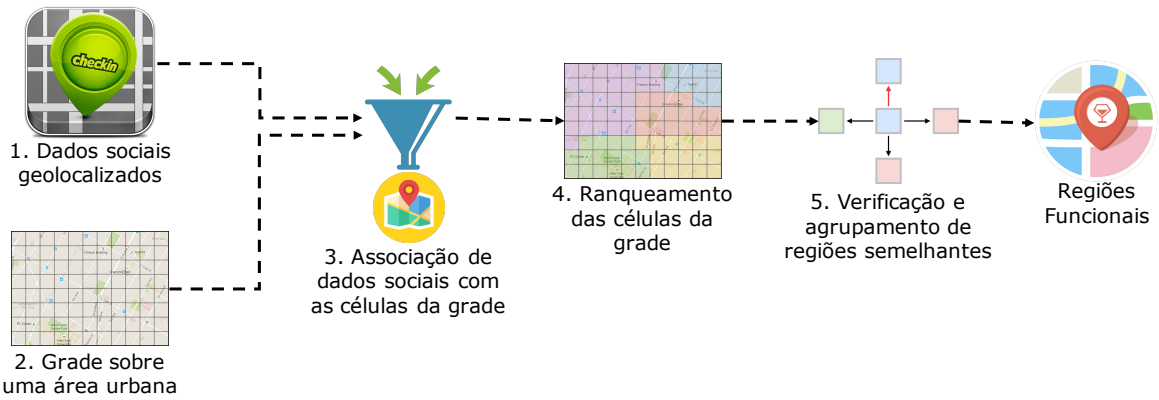


Figura 4.2: Método proposto.

Como resultado, obtemos regiões compostas por polígonos, formados por coordenadas, e funções extraídas de dados sociais compartilhados por usuários da área

urbana estudada. As funções das regiões definidas podem variar em dias e meses por exemplo, devido às mudanças dos dados das redes sociais ao longo do tempo. Desta forma, podemos dizer que as regiões funcionais são estabelecidas através de atividades humanas.

De acordo com o método proposto, definimos que as regiões funcionais possuem as seguintes variáveis:

- Áreas formadas por uma lista de polígonos P, de forma que $P = \{p_1, p_2, p_3, ..., p_i\}$ e i é o total de células da grade. O valor de i é definido no passo 2, pois ao estabelecermos a área em que a grade será formada, informamos quatro pontos, NLa, SLa, LLo e OLo referentes as latitudes norte e sul e as longitudes leste e oeste, respectivamente. Em seguida, esta área é dividida em subregiões pertencentes a lista P.

- Funções predominantes formadas por dados sociais geolocalizados F, em que $F = \{f_1, f_2, f_3, ..., f_j\}$ e j é a quantidade de funções referentes a cada polígono pertencente a lista P. No passo 3, diversos dados sociais são associados a cada célula da grade, fazendo com que as mesmas possuam diversas funções. Desta forma, as funções predominantes são determinadas após o ranqueamento das funções das células da grade, no passo 4, em que apenas a função de maior ocorrência permanece.

- Intervalos de tempo T, de forma que $T = \{t_1, t_2, t_3, ..., t_k\}$ e k é a quantidade de intervalos de tempo. Estes intervalos de tempo podem ser estabelecidos em horas, dias, semanas, meses e etc.

Assim, ao final do passo 4, possuímos uma matriz que associa P a F para cada intervalo de tempo em T, de forma que $P \rightarrow \{F \times T\}$ e $p \mapsto (f, t_i)$, $f \in F$ e $t_i \in T$, para $i = \{1, ..., k\}$.

Por fim, para os intervalos de tempo em T, verificamos para cada região em P se as regiões adjacentes possuem a mesma função predominante contida em F, de forma que as regiões semelhantes sejam interpoladas. Assim, quando p_1 é adjacente a p_2 e $p_1 \mapsto (f, t_k)$ e $p_2 \mapsto (f, t_k)$ possuem funções predominantes iguais, ocorre a interpolação entre p_1 e p_2.

O método proposto também permite que possamos analisar as regiões funcionais tanto em conjunto (todos os valores em F) quanto de forma individual. Assim, também é possível formar os polígonos em P, baseando-se apenas em uma função de estudo.

Desta forma, após a obtenção das regiões funcionais, podemos caracterizar os acidentes contidos nas mesmas, conforme é apresentado na Seção 5.4.

4.3 Agrupamento e Contextualização dos Dados para Identificar Eventos Urbanos

Para a identificação de eventos urbanos, propomos um método que utiliza um algoritmo de agrupamento incremental para processamento de *streams* de dados provenientes de redes sociais (Figura 4.3). Esta abordagem é viável em relação à tradicional [Ester et al., 1996], pois apenas uma parte específica da base de dados é processada de acordo com a publicação de novas informações nas redes sociais e com o funcionamento do modelo *Sliding Window*. Além disso, esta abordagem permite a detecção de diferentes ocorrências relacionadas ao evento em estudo.

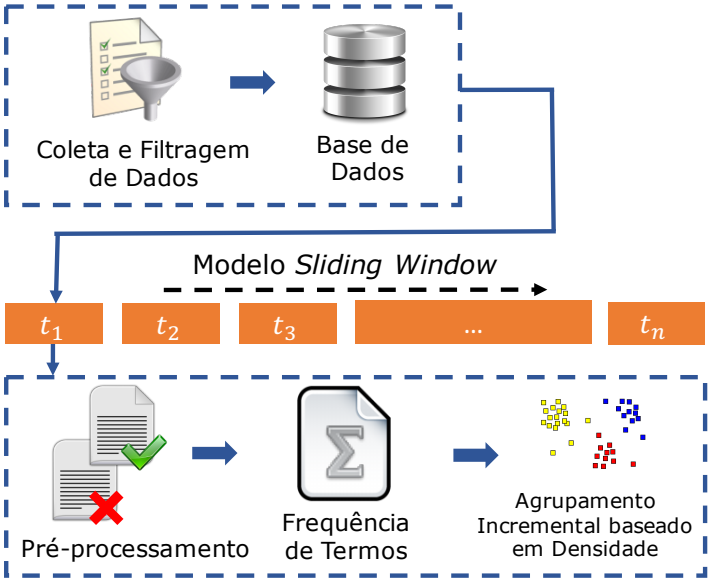


Figura 4.3: Método proposto para agrupar dados similares em *streams* de redes sociais.

Primeiramente, é realizada a coleta e filtragem dos dados através de palavras-chave relacionados ao tópico em análise. No caso deste trabalho, palavras-chave relacionadas são "*accident*", "*car*", "*injury*" e assim por diante. A coleta de dados é realizada através do mecanismo de busca da API da rede social Twitter ou *Web Crawlers*, resultando em dados massivos de texto que aumentam constantemente. A seguir, a filtragem dos dados é realizada, o que é importante devido os dados de redes sociais possuírem bastante ruídos [Sankaranarayanan et al., 2009, Sakaki et al., 2010].

Após a remoção dos ruídos, é realizada a extração de catacterísticas dos *tweets*. Apesar de ser possível coletar dados como imagens, fotos de usuários e geolocalização de locais, a maioria das informações provenientes de redes sociais são em forma de texto. Assim, nós extraímos apenas os dados de texto das ocorrências coletadas e aplicamos um algoritmo para determinar a frequência dos termos na extração de características.

Por fim, aplicamos o algoritmo de agrupamento incremental baseado no DBSCAN [Ester et al., 1996] para processar apenas novos dados que são adicionados à base, além dos dados que são afetados por esta atualização. Desta forma, o tempo de agrupamento de novas informações reduz, tornando possível a utilização do método para *streams* de dados em constante mudança.

4.4 Considerações Finais

Neste capítulo, foi apresentado o método proposto para este trabalho. Através da Figura 4.1, é possível obter uma visão geral do método, que possui as etapas de coleta de dados, formação de camadas e caracterização e detecção de eventos urbanos.

Nos capítulos 5 e 6, serão apresentados a descrição dos problemas que buscamos resolver e os resultados e avaliações relacionados a este método, sendo um deles utilizando redes sociais baseadas em localização e outro utilizando processamento de linguagem natural.

Capítulo 5

Detecção de Regiões Funcionais Utilizando Dados de Redes Sociais Baseadas em Localização

Como foi descrito anteriormente, nossa abordagem consiste tanto em agrupar e contextualizar o conteúdo de *tweets* para a identificação de acidentes de trânsito, quanto em caracterizar áreas urbanas. O estudo da caracterização destas áreas, permite que sejam identificados os fatores que tornam um acidente de trânsito propício a acontecer em um perímetro urbano.

Para a obtenção dos resultados a serem apresentados, foi utilizado um conjunto de dados de 157.054 *check-ins* do Foursquare, com o objetivo de determinar regiões funcionais. Comparando os resultados com dados governamentais, verificamos que foi obtida uma acurácia de 86% para o método proposto. Além disso, também apresentamos dois estudos de caso para o mesmo: um relacionado à análise de ocorrências de trânsito, e outro relacionado à dinâmica das áreas urbanas.

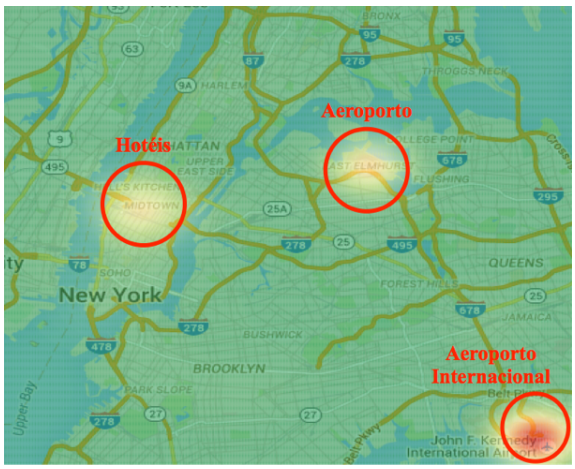
5.1 Definição do problema

A caracterização de áreas urbanas, realizada por meio de dados sociais, pode ser utilizada para diversos fins, dentre os quais podemos citar o estudo da dinâmica das cidades e da rotina dos usuários [Silva et al., 2014a], a detecção de pontos de interesse [Silva et al., 2013b] e a escolha do melhor local para iniciar um empreendimento [Karamshuk et al., 2013].

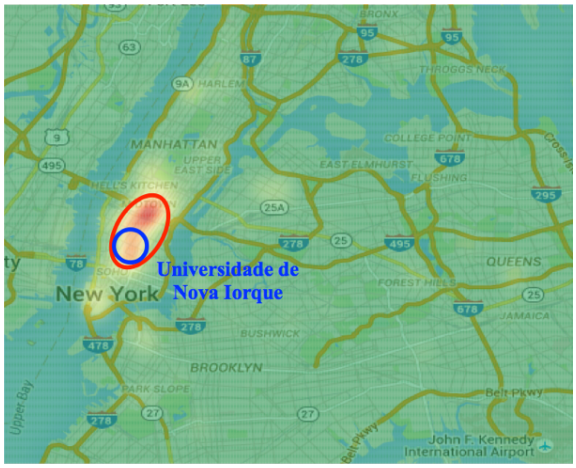
Os fatores socioeconômicos de um determinado perímetro urbano podem carac-

terizar e influenciar diretamente na rotina de seus habitantes e nos acontecimentos em torno de uma região, como a ocorrência de crimes e de acidentes de trânsito. Para estes estudos, redes sociais como o Twitter, o Instagram e o Foursquare são amplamente utilizadas [Silva et al., 2013a,b].

O Foursquare, por exemplo, permite que sejam identificadas características de uma região a partir de suas categorias. Como exemplo, apresentamos a Figura 5.1, na qual é possível identificar, na cidade de Nova Iorque, áreas residenciais, cinemas, aeroportos, universidades, entre outros, por meio das categorias do anteriormente citadas.



(a) Categoria *Travel & Transport*



(b) Categoria *College & University*

(c) Categoria *Arts & Entertainment*

(d) Categoria *Residence*

Figura 5.1: Caracterização do perímetro urbano da cidade de Nova Iorque através de categorias do Foursquare.

As particularidades do Foursquare, se somadas a outras fontes de dados, podem auxiliar na extração de características de perímetros urbanos, para a identificação de regiões funcionais, além dos fatores que influenciam na ocorrência de acidentes de trân-

sito. Através desta caracterização, torna-se possível a identificação dos melhores locais para a implantação de melhorias que proporcionem a redução de mortes e acidentes no trânsito, além de tornar possível um melhor gerenciamento de recursos.

Neste contexto, este trabalho apresenta um método de detecção de regiões funcionais que utiliza dados do Foursquare. O método associa dados sociais com regiões em uma determinada área urbana. Assim, o cotidiano das pessoas que habitam a área irá definir as funções das regiões.

5.2 Experimentos e Resultados Preliminares

Nesta seção são apresentados os resultados preliminares do método proposto, que utilizam como fonte de dados o Foursquare e dados governamentais.

No primeiro resultado preliminar, descrito na Seção 5.2.1, foi utilizado apenas uma fonte de dados, ou seja, apenas uma camada de dados. Além disso, o algoritmo de inserção de pontos em polígonos ainda não fazia parte do método, e as áreas estudadas eram cidades, e não pequenas unidades federativas ou pequenas regiões. Estas técnicas passaram a fazer parte do método a partir do segundo resultado preliminar, apresentado na Seção 5.2.2.

5.2.1 Detecção de Pontos de Interesse

A detecção de Pontos de Interesse pode ser feita por meio de Sensoriamento Social para diversos fins, como a identificação de pontos turísticos de um cidade [Silva et al., 2013b] e a detecção de eventos em uma região específica [Lee and Sumiya, 2010].

Neste sentido, a principal contribuição deste resultado preliminar foi: um método de detecção de Pontos de Interesse, através dos dados da rede social do Foursquare, para identificar os locais mais frequentados em uma cidade. Como prova de conceito, foi criada a aplicação *Points of Interests* (PoI), cujo o público-alvo é composto por empresários e órgãos governamentais que desejem identificar locais frequentados por um público específico, através de dados como a categoria do local e o sexo dos frequentadores, tornando possível a realização de campanhas e divulgação de produtos.

A arquitetura proposta para a aplicação PoI é composta essencialmente por um servidor e uma aplicação *web*, além de um gerenciador de banco de dados, como ilustrado na Figura 5.2. Desta forma, as requisições dos usuários são enviadas ao servidor para que este se comunique com as camadas da aplicação. Logo em seguida, a aplicação solicita dados do gerenciador de banco de dados. Este, por sua vez, retorna os

dados solicitados para a aplicação, a fim de que a mesma direcione o resultado para o servidor. Por fim, o servidor envia uma resposta para o *browser* dos usuários.

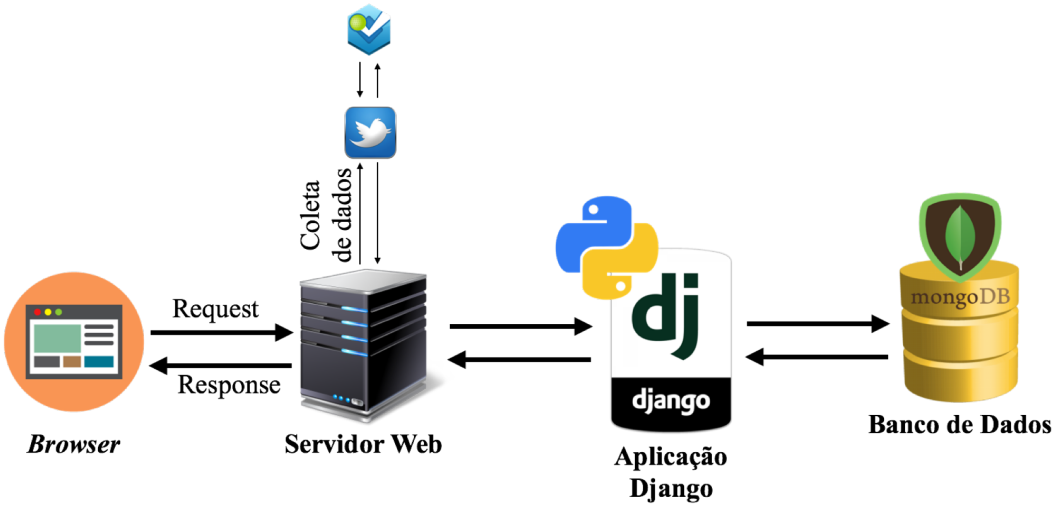


Figura 5.2: Visão Geral da arquitetura da aplicação PoI.

Os dados da aplicação são provenientes do Foursquare e foram coletados entre 14/12/2015 e 29/12/2015. A coleta dos dados foi realizada através do Twitter, utilizando a *query* de busca "swarmapp.com". Para este período foram coletados 333.286 *tweets* provenientes de capitais brasileiras, contendo URLs do Swarm[1], aplicação responsável por registrar os *check-ins* do Foursquare.

Após a filtragem e a extração dos dados, verificamos que o Rio de Janeiro era a capital brasileira com a maior quantidade de *check-ins*, seguida por Porto Alegre e São Paulo (Figura 5.3). Além disso, verificamos que as 5 categorias do Foursquare com mais *check-ins* nesse período eram *Shop and Service*, *Food*, *Outdoors & Recreation*, *Professional & Other Places* e *Residence*, nesta ordem. Isto pode ser justificado pelo fato de que os dados são referentes a uma época festiva, o Natal.

Através dos *check-ins* também é possível recuperar informações a respeito do usuário que realizou o mesmo. Assim, verificamos que, neste período, cerca de 75% dos *check-ins* foram realizados por homens. Além disso, verificamos que cerca de 50% dos *check-ins* ocorreram no horário comercial (08:00 às 17:00), enquanto 42% ocorreram durante a noite e a madrugada.

Por fim, assumindo que os locais mais frequentados de uma região são os locais com o maior número de *check-ins*, identificamos através da aplicação PoI os locais mais frequentados nas capitais brasileiras, de acordo com cada categoria do Foursquare.

[1] `https://www.swarmapp.com/`

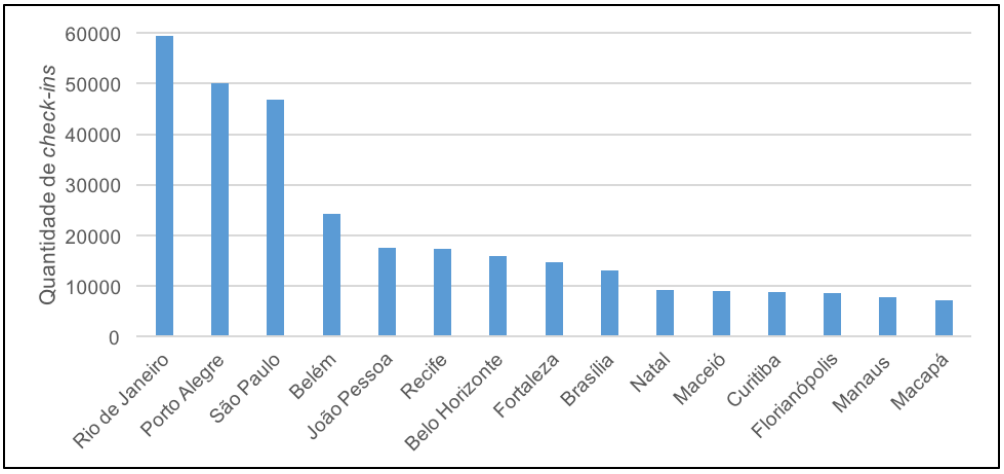


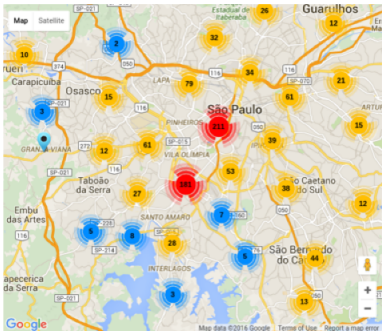
Figura 5.3: Quantidade de *check-ins* do Foursquare em 15 capitais brasileiras, extraídos no período de 14/12/2015 a 29/12/2015.

Como exemplo, apresentamos a Figura 5.4, na qual é possível identificar Pontos de Interesse da categoria *College & University*, na cidade de Manaus. Nesta categoria, os locais com o maior número de *check-ins* foram universidades particulares e públicas da cidade.

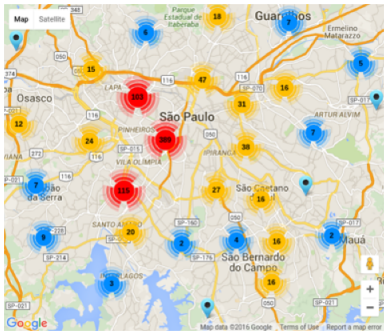
Figura 5.4: Detecção de Pontos de Interesse da categoria *College & University*, na cidade de Manaus.

Em outro exemplo, apresentamos a Figura 5.5, na qual é possível identificar Pontos de Interesse em diferentes categorias na cidade de São Paulo e outras regiões próximas, nos dias 20/12/2015 e 21/12/2015. Para este exemplo, apresentamos a Tabela 5.1, na qual é possível verificar os 3 locais mais frequentados em 4 categorias, ou seja, os 3 principais Pontos de Interesse destas categorias.

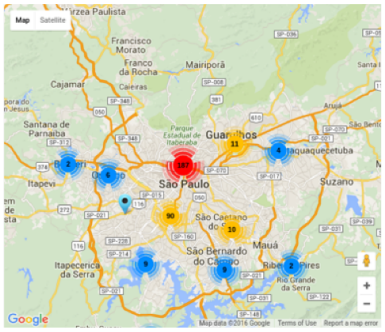
Além das categorias do Foursquare, também é possível determinar Pontos de Interesse utilizando os dados dos usuários. Um exemplo disso, seria verificar os locais mais frequentados por homens ou mulheres em uma região.



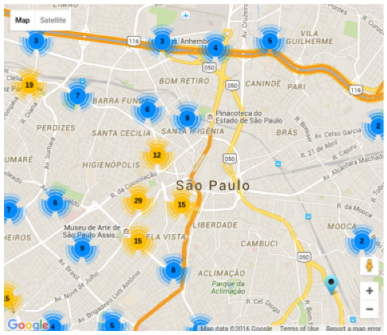
(a) Categoria *Shopping & Service*



(b) Categoria *Food*



(c) Categoria *Nightlife Spot*



(d) Categoria *Arts & Entertainment*

Figura 5.5: Pontos de Interesse em diferentes categorias na cidade de São Paulo e regiões próximas, nos dias 20/12/2015 e 21/12/2015.

Categoria do Foursquare	Quantidade de *check-ins*	Lugares mais frequentados
Shop & Service	1331	Shopping Metrô Boulevard Tatuapé, Linda Unha, Havaianas.
Food	1249	Cris Doces e Salgados, Central de Massas, Lanches Escadão.
Nightlife Spot	386	Ilhabela Irish Pub, Barcearia, Coconut Brasil.
Arts & Entertainment	363	Estúdio Iquiririm, Cinemark, Cinépolis.

Tabela 5.1: Lista dos 3 locais mais frequentados nos dias 20/12/2015 e 21/12/2015, divididos por categoria referentes à cidade de São Paulo.

Com os exemplos apresentados nesta subseção, verificamos que a aplicação PoI, através das redes sociais, pode prover informações necessárias para diversos fins, dentre os quais podemos destacar:

- Verificação de pontos de interesse de acordo com parâmetros previamente estabelecidos (categoria do local, idade e sexo do usuário, etc.). Isto pode auxiliar em campanhas publicitárias que queiram atingir um determinado público-alvo, por exemplo;
- Listagem dos locais mais frequentados em um região para o auxílio de turistas.

5.2.2 Caracterização de Áreas em um Perímetro Urbano

O método de caracterização de acidentes em áreas urbanas foi utilizado na análise de pequenas unidades federativas da cidade de Nova Iorque, as *tracts*. Para este estudo, foram utilizados dados coletados da rede social Foursquare e do Censo de Nova Iorque, assim como dados de acidentes de trânsito do Departamento de Polícia da cidade.

Em relação ao Foursquare, foram coletados 157.054 *check-ins* no período de 24/04/2014 a 01/08/2014. Para este mesmo período, foram coletadas 49.163 ocorrências de acidentes de trânsito do Departamento de Polícia. Os acidentes de trânsito foram rotulados automaticamente de acordo com a sua gravidade. Acidentes que possuíam registros de vítimas fatais ou lesionadas, eram rotulados como graves. Já os acidentes no qual foram registrados apenas danos materiais foram rotulados como leves.

Após a definição dos polígonos, representados por uma lista de coordenadas geográficas (Figura 5.6), e do agrupamento dos dados nas *tracts*, foi possível realizar uma análise mais detalhada de cada região, com ênfase nas regiões com o maior e o menor número de acidentes de trânsito.

Assim, verificamos os aspectos de cada região que podem caracterizar os acidentes de trânsito. Na Figura 5.7, por exemplo, apresentamos a *tract* com o menor número de acidentes e a *tract* com o maior número de acidentes para o período descrito anteriormente. Na primeira, podemos identificar a presença de vias residenciais. Já na segunda, existe uma avenida. Isto pode influenciar em fatores como a velocidade limite de tráfego dos veículos na região. Vale ressaltar que as *tracts* foram ordenadas de acordo com a quantidade de ocorrências de acidentes de trânsito registradas. Em caso de empate, era levada em consideração a quantidade de ocorrências de acidentes graves.

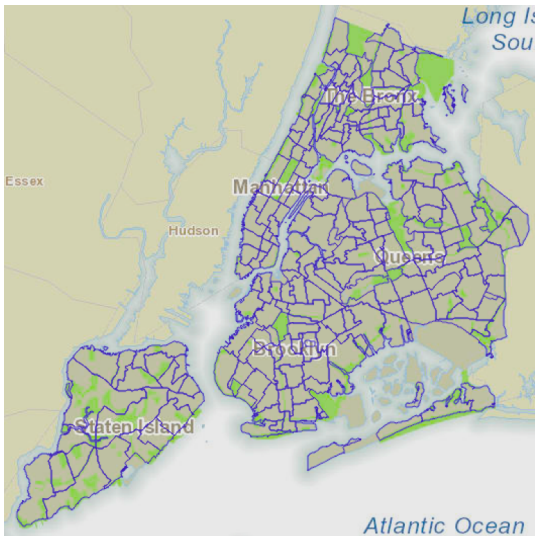
A Tabela 5.2 apresenta outras características extraídas de cada camada de dados, as quais representam as fontes de informação anteriormente citadas. Nesta tabela, é possível verificar que na região de maior risco ocorreram 275 acidentes de trânsito, sendo 19 graves. Além disso, a categoria do Foursquare predominante na região é *Nightlife Spot*, o que indica uma presença relevante de bares e casas noturnas. Também

```
<td>36001000100</td>
   <Polygon>
      <coordinates>
         -73.741265,42.678681,0 -73.742978,42.679275,0 -73.743117,42.679323,0
         -73.74314,42.679235,0 -73.743737,42.677112,0 -73.745404,42.671092,0
         -73.745096,42.669948,0 -73.743362,42.665549,0 -73.744124,42.663822,0
         -73.74166,42.660655,0 -73.74283,42.658936,0 -73.739429,42.656803,0
         -73.730864,42.663279,0 -73.7291,42.665081,0 -73.724964,42.670179,0
         -73.723263,42.672879,0 -73.72619,42.673994,0 -73.732619,42.67597,0
         -73.735153,42.676873,0 -73.736974,42.677355,0 -73.741265,42.678681,0
      </coordinates>
   </Polygon>

<td>36001000200</td>
   <Polygon>
      <coordinates>
         -73.750223,42.653544,0 -73.750127,42.65373,0 -73.750045,42.654009,0
         -73.750031,42.654244,0 -73.749944,42.654679,0 -73.749727,42.656209,0
         -73.749412,42.657812,0 -73.749265,42.658246,0 -73.747405,42.660192,0
         -73.744764,42.662751,0 -73.744335,42.663338,0 -73.744124,42.663822,0
         -73.743362,42.665549,0 -73.745096,42.669948,0 -73.745404,42.671092,0
         -73.743737,42.677112,0 -73.747608,42.675794,0 -73.748927,42.674681,0
         -73.750436,42.67351,0 -73.753777,42.670737,0 -73.754544,42.670112,0
         -73.755915,42.669012,0 -73.757285,42.667655,0 -73.757322,42.667004,0
         -73.757106,42.666758,0 -73.756117,42.666067,0 -73.757012,42.665246,0
         -73.757871,42.664443,0 -73.758196,42.664125,0 -73.758664,42.663708,0
         -73.760179,42.662273,0 -73.760827,42.661618,0 -73.761101,42.661338,0
         -73.761562,42.660941,0 -73.759932,42.659968,0 -73.7616,42.658335,0
         -73.761745,42.658177,0 -73.756806,42.656376,0 -73.754756,42.655187,0
         -73.750875,42.652557,0 -73.750223,42.653544,0
      </coordinates>
   </Polygon>

.
.
.
```

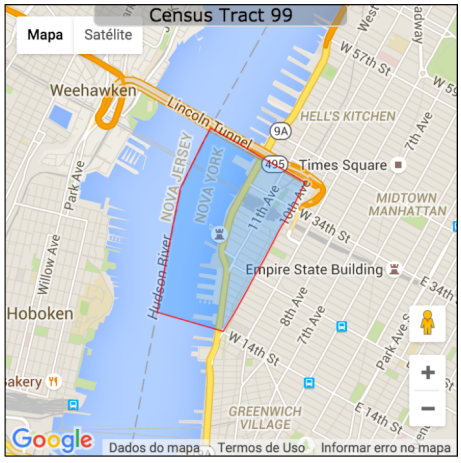
(a) Exemplo do documento com a lista de coordenadas geográficas para a formação das *Tracts*



(b) Representação dos polígonos das *Tracts* presentes no documento de coordenadas geográficas

Figura 5.6: Exemplo do documento de coordenadas geográficas para a definição dos polígonos das *Tracts*.

(a) *Tract* com o menor número de acidentes de trânsito



(b) *Tract* com o maior número de acidentes de trânsito

Figura 5.7: *Tracts* de Nova Iorque com diferentes características urbanas e sociais, sendo a Figura (a) a *tract* com o menor número de acidentes de trânsito, e a Figura (b) a *tract* com o maior número de acidentes de trânsito.

verificamos que nesta região a idade média da população é menor do que na outra região apresentada, a qual é definida como uma região de menor risco. Na região com menor número de acidentes, a categoria do Foursquare predominante é *Residence*, o que indica uma presença relevante de casas, prédios e condomínios residenciais. Além disso, verificamos que a velocidade permitida nas vias é menor e que a quantidade de mulheres é maior que a quantidade de homens, diferentemente do que pode ser observado na região de risco.

Tract	Categoria predominante do Foursquare	Total de acidentes de trânsito	Velocidade nas vias	Total da população	Idade média da população
99 (Manhattan)	*Nightlife Spot*	275 (19 graves; 256 leves)	35–55 mph (56–89 km/h)	3.617 (2.020 Homens; 1.597 Mulheres)	34,4 anos
1020 (Brooklyn)	*Residence*	1 (0 graves; 1 leve)	15–45 mph (24–72 km/h)	2.054 (938 Homens; 1.116 Mulheres)	43,3 anos

Tabela 5.2: Características das *tracts* com o maior e o menor número de acidentes de trânsito identificados.

Com este método de extração de características e de rotulação de dados automática, é possível gerar informações necessárias para algoritmos de classificação ou de mineração de dados, que por sua vez podem ser utilizados para gerar ou inferir novas

informações. Como exemplo, podemos inferir um local para implantar uma solução preventiva para acidentes de trânsito.

5.3 Experimentos e Resultados Finais: Regiões Funcionais

Após os resultados preliminares, substituímos as *tracts* por pequenas regiões, com o objetivo de realizar um estudo mais detalhado. Chamamos estas regiões de funcionais, pois cada uma irá obter uma função específica F de acordo com as categorias do Foursquare (vide Seção 4.3). Escolhemos esta rede social pois a mesma é baseada em localização. Além disso, suas categorias podem estabelecer a função de uma região de acordo com as atividades dos habitantes da mesma.

Desta forma, as regiões funcionais geradas pelo método proposto estarão sempre de acordo com a atualidade, facilitando estudos relacionados, que em sua maioria costumam ser custosos se realizados por meios tradicionais.

A área de estudo foi definida na cidade de Nova Iorque e regiões adjacentes, gerando uma lista P com 2.208 polígonos com lados de 1 quilômetro de comprimento. O estudo foi feito com 6 períodos de tempo em T, sendo eles: 00:00 a 03:59, 04:00 a 07:59, 08:00 a 11:59, 12:00 a 15:59, 16:00 a 19:59 e 20:00 a 23:59.

Escolhemos polígonos de 1 quilômetro de comprimento para uma análise mais minuciosa das regiões, dado que os trabalhos relacionados analisam toda a área urbana para obtenção de resultados. Já as seis faixas de horários foram escolhidas para que fosse possível verificar como as regiões mudam sua função ao longo do dia de acordo com a rotina dos habitantes. Ressaltamos que tanto o tamanho das células quanto os horários são configuráveis.

Para o desenvolvimento e avaliação do método proposto, utilizamos o Python e suas APIs e bibliotecas disponíveis para a coleta e análise de dados. Para a coleta dos dados utilizamos a API de busca do Twitter, já que o Foursquare não permite uma coleta direta por sua API. Essa coleta é possível porque sempre que um usuário do Foursquare faz um *check-in* e tem uma conta do Twitter associada, os dados do Foursquare vão para o Twitter também. Para o mapa presente nos resultados, utilizamos a API do Google Maps e para a análise dos dados utilizamos bibliotecas como pandas, matplotlib e numpy.

5.3.1 Descrição dos Dados

Para a realização dos experimentos, foram coletados 157.054 *check-ins* do Foursquare entre abril de 2014 e agosto de 2014. Estes dados foram coletados no respectivo ano para posterior validação do método, que foi comparado com dados governamentais do mesmo período. Na Figura 5.8, apresentamos a quantidade de *check-ins* por categoria do Foursquare com os mesmos períodos de tempo contidos em T.

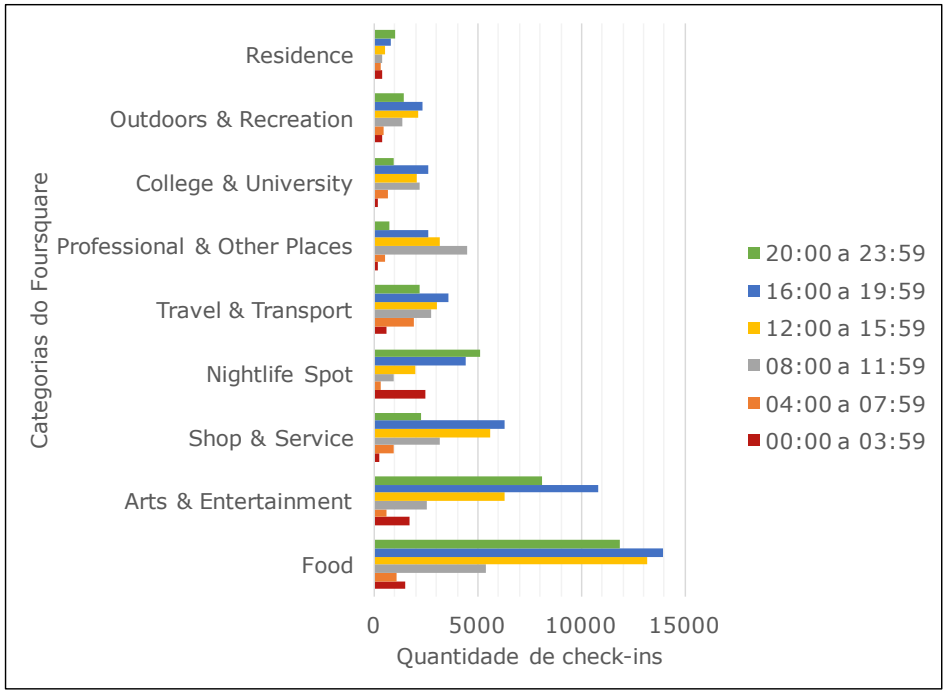


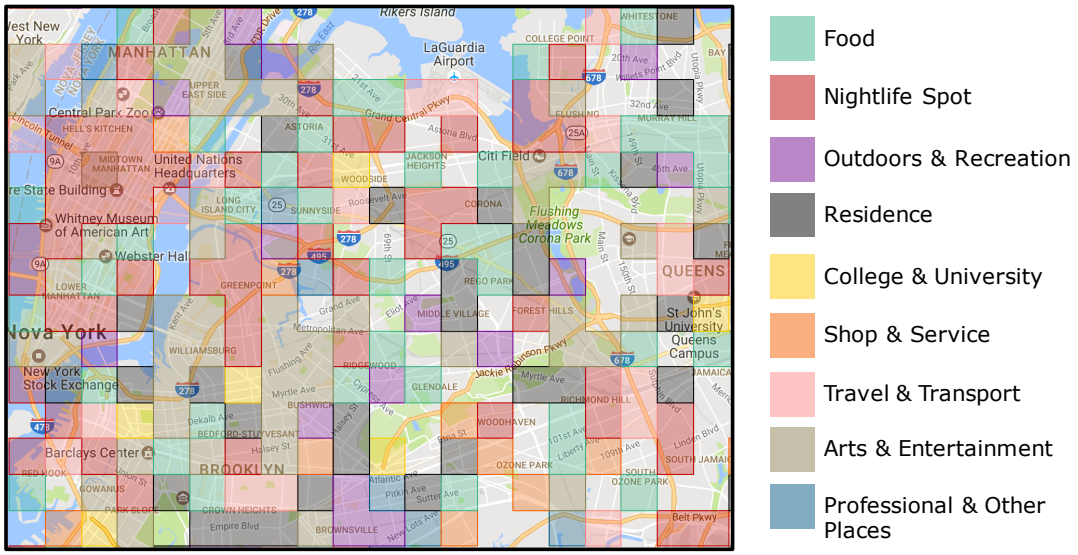
Figura 5.8: Quantidade de *check-ins* por categoria do Foursquare ao longo do dia.

Podemos observar que a categoria com o maior número de *check-ins* é *Food*, enquanto que a menor é *Residence*. Isto ocorre porque muitos usuários evitam registrar suas propriedades privadas em redes sociais por motivos de segurança. Apesar disso, podemos observar nos resultados que algumas regiões funcionais são referentes à áreas residenciais.

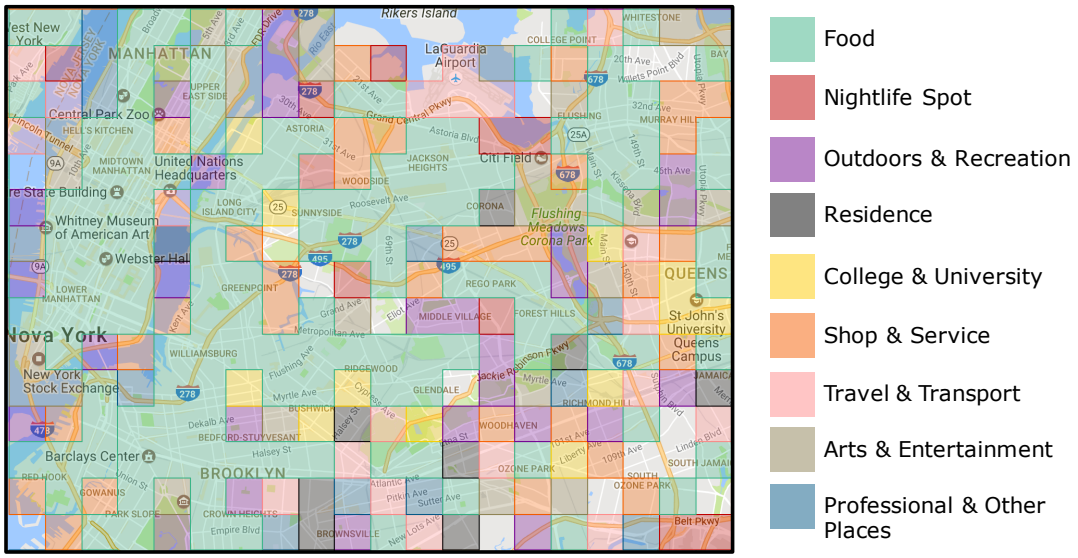
5.3.2 Comparação com Dados Governamentais

Após a utilização dos dados sociais, o método proposto gerou regiões funcionais que mudam suas funções ao longo do tempo, de acordo com as atividades dos habitantes das mesmas. Como exemplo, podemos verificar na Figura 5.9, as regiões funcionais estabelecidas nos períodos 00:00 a 03:59 e 12:00 a 15:59.

No primeiro período (Figura 5.9a), as categorias com mais *check-ins* foram *Nightlife Spot*, *Arts & Entertainment* e *Food*, nesta ordem. Isto é refletido nas regiões



(a) 00:00 a 03:59



(b) 12:00 a 15:59

Figura 5.9: Resultado do método proposto para diferentes períodos de tempo, apresentando a mudança nas funções das regiões.

funcionais estabelecidas. Da mesma forma, no segundo período (Figura 5.9b), as categorias com mais *check-ins* foram *Food* e *Shop & Service*, nesta ordem. Além disso, podemos verificar que algumas áreas estão vazias. Isto ocorre porque, nestes períodos, não houveram dados sociais que refletissem a função da região. Isto pode ser resolvido aumentando-se a quantidade de dados sociais para estudo. Também é possível verificar o resultado da interpolação das regiões, que apresenta novos polígonos em diferentes formas.

Através destes resultados, podemos verificar que o método proposto permite que

possamos realizar diferentes estudos de uma região, pois a função da mesma se altera de acordo com a época do ano e o período do dia, de forma que reflete a rotina dos usuários que nela residem.

Para a validação do método proposto, foram utilizados dados do Departamento de Planejamento Urbano da cidade de Nova Iorque, denominados *City Owned and Leased Properties*[2]. Estes dados podem ser visualizados na aplicação *Zoning and Land Use Application*[3].

Os dados foram fornecidos em forma de uma planilha e estavam divididos por categorias que podem ser mapeadas para as categorias do Foursquare. Porém, os dados não eram geolocalizados, tornando necessária a utilização de uma API de mapa para transformar os endereços dos locais em coordenadas.

Em seguida, estes dados governamentais foram inseridos nas regiões funcionais cujas coordenadas geográficas coincidiam. Por fim, mapeamos as categorias de locais nos dados governamentais para as categorias macro do Foursquare, de acordo com a hierarquia de categorias desta rede social[4]. A partir disso, criamos e comparamos as regiões funcionais de ambas as fontes de dados. Regiões idênticas eram contadas como acerto, enquanto regiões diferentes eram contadas como erro. Desta forma, apuramos que o nosso método possui uma acurácia de 86% para a caracterização de uma área urbana.

Com a similaridade encontrada, temos a contribuição de que o método proposto pode ser muito útil para entender o comportamento de cidades mesmo quando não se tem recursos necessários para realização de censos. Além disso, verificamos que os dados sociais podem ser utilizados como meio alternativo para o auxílio no planejamento urbano de uma cidade.

5.4 Utilização das Regiões Funcionais

Nesta seção, apresentamos duas possíveis aplicações para o método proposto, a análise de acidentes de trânsito e o monitoramento dinâmico das cidades. Estas aplicações mostram como a utilização de regiões funcionais pode facilitar a tomada de decisões e ações preventivas em uma determinada região, de forma dinâmica e com baixo custo. Apesar do foco nestas duas aplicações, o método proposto também pode ser empregado para a avaliação econômica, social e de mobilidade da área urbana estudada.

[2] `http://www1.nyc.gov/site/planning/about/publications/colp.page`
[3] `http://maps.nyc.gov/doitt/nycitymap/template?applicationName=ZOLA`
[4] `https://developer.foursquare.com/categorytree`

5.4.1 Aplicação I: Análise de Acidentes de Trânsito

Com o objetivo de criar soluções eficientes para a prevenção de acidentes de trânsito, órgãos governamentais buscam analisar os fatores relacionados aos mesmos em uma determinada região. Aspectos como a localização e o horário em que os acidentes ocorreram são levados em consideração, assim como as condições das vias, condutores e veículos.

Desta forma, uma possível aplicação para as regiões funcionais, além do auxílio ao planejamento urbano, é a verificação dos fatores relacionados aos acidentes de trânsito em cada região detectada que subdivide uma área urbana. Para isto, definimos as regiões funcionais através dos dados do Foursquare (conforme apresentado na Seção 5.3). Em seguida, utilizamos dados de acidentes de trânsito fornecidos pelo departamento de polícia da cidade de Nova Iorque[5], com o objetivo de verificarmos a relação entre as regiões funcionais e os aspectos dos acidentes de trânsito.

Nestes dados do departamento de polícia, obtivemos informações sobre a data e a hora em que o acidente ocorreu, além da latitude, longitude, envolvidos no acidente e fatores que contribuíram para o mesmo. Após uma análise, divimos os fatores em 4 categorias:

- **Direção imprópria**: ações como dirigir enquanto utiliza algum dispositivo móvel, fazer manobras proibidas na via e alta velocidade estão inclusas nesta categoria;
- **Motorista não condicionado**: condutores que estavam sob o efeito de bebidas alcoólicas ou drogas, estavam cansados ou perderam a consciência no momento do acidente estão inseridos nesta categoria;
- **Externo**: vias em condições ruins, animais e distrações provocadas por passageiros estão nesta categoria;
- **Veículo não condicionado**: falhas no acelerador, freios ou outras partes do veículo estão inseridas nesta categoria.

Além dos fatores descritos acima, também verificamos se o acidente ocasionou apenas danos materiais ou envolveu a morte do condutor ou de terceiros, como pedestres e ciclistas. No primeiro caso, o acidente foi classificado com **leve**. No segundo, como **grave**.

[5] `https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95`

Como as regiões funcionais são representadas por polígonos com dados sociais relacionados a eles, aplicamos um algoritmo para verificar em qual região funcional os dados geolocalizados dos acidentes de trânsito (latitude e longitude) estavam inseridos. Assim, tornou-se possível associar os dados do departamento de polícia com os dados do Foursquare.

Desta forma, analisamos 73.182 acidentes de trânsito entre abril de 2014 e agosto de 2014, o mesmo período dos dados do Foursquare apresentados na Seção 5.3. O resultado da análise foi dividido em 6 períodos de tempo para a verificação da mudança nas associações dos dados ao longo do dia (Figura 5.10).

Através das associações representadas na Figura 5.10 (sendo que quanto maior a incidência da associação, maior é a espessura da aresta), fizemos as seguintes observações relacionadas aos acidentes de trânsito:

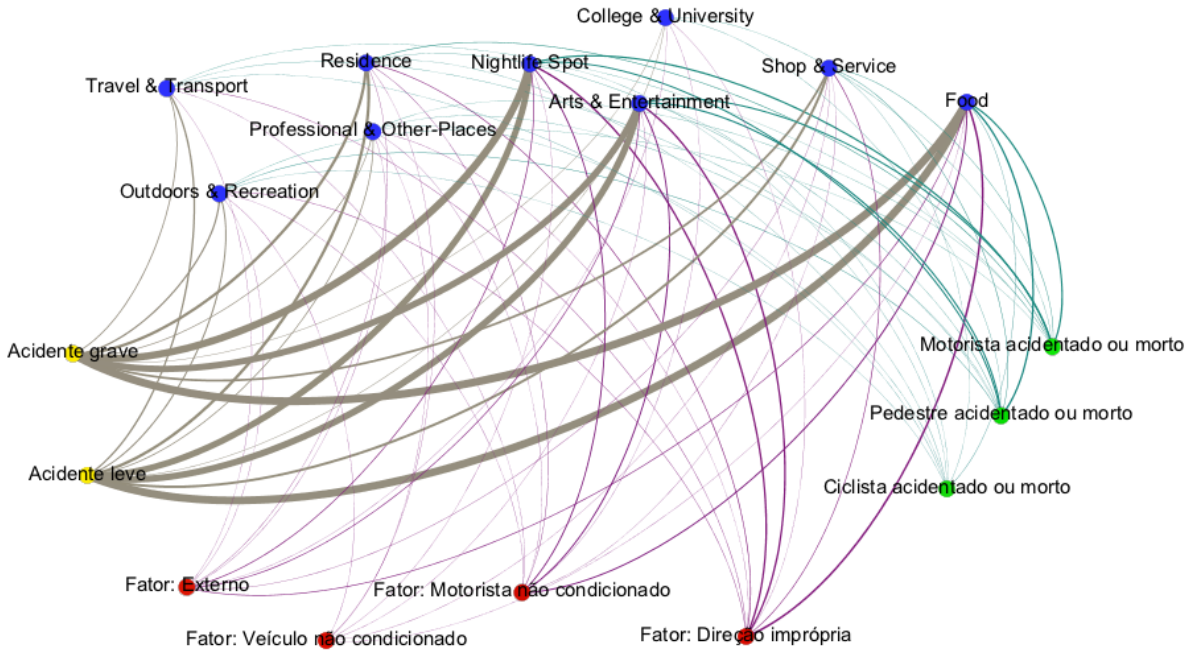
- No período de 00:00 a 03:59 (Figura 5.10a), percebemos que a maioria dos acidentes, tanto graves quanto leves, estão associados às categorias *Nightlife Spot*, *Arts & Entertainment* e *Food*, assim como os fatores *Motorista não condicionado* e *Direção imprópria*. Isto sugere que os acidentes nestas regiões estão relacionados ao uso de substâncias não permitidas, como o álcool, por exemplo. Além disso, verificamos que estes mesmos acidentes ocasionaram a morte de motoristas e pedestres.

- No período de 04:00 a 07:59 (Figura 5.10b), percebemos que os fatores relacionados aos acidentes começam a ser direcionados para outras categorias do Foursquare, como *Professional & Other Places*, *College & University* e *Shop & Service*. Isto está diretamente relacionado à rotina das pessoas na região estudada, que passam a realizar suas atividades cotidianas neste período de tempo.

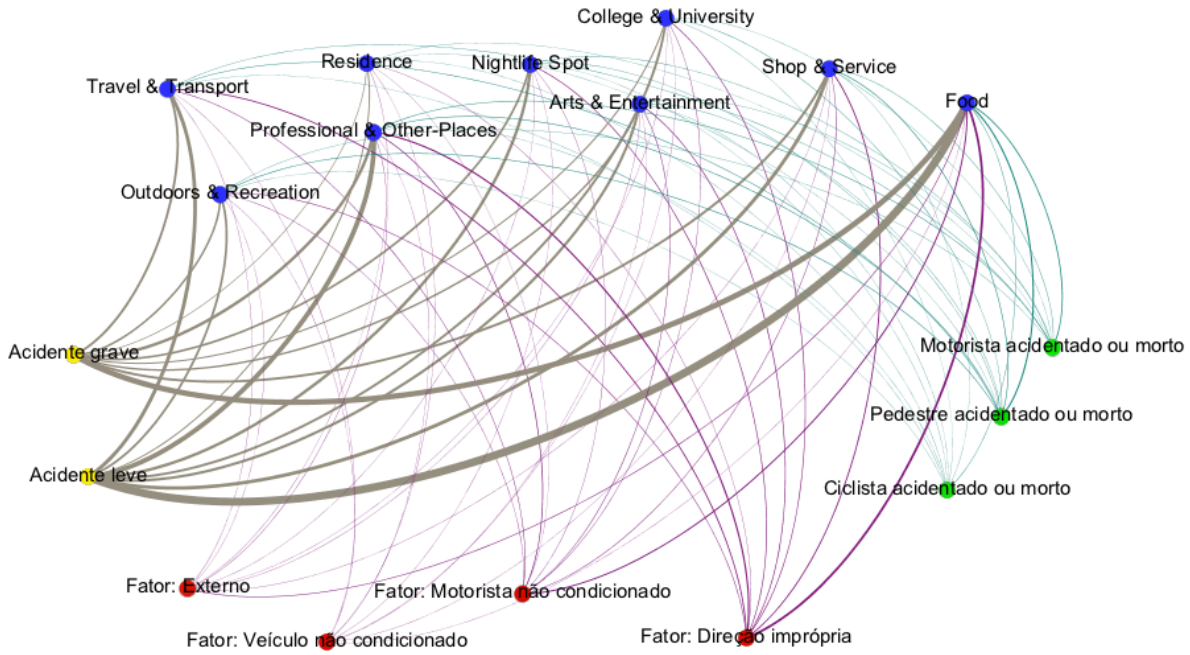
- No período de 08:00 a 11:59 (Figura 5.10c), percebemos que os acidentes passam a se concentrar nas categorias *Professional & Other Places* e *Food*, devido ao horário do *rush*, e que a maioria dos acidentes está relacionado a direção imprópria. Neste período, o índice de acidentes leves é maior do que o de acidentes graves, o que indica que, na maioria dos acidentes, ocorrem apenas danos materiais. Um padrão semelhante é apresentado na Figura 5.10d.

- No período de 16:00 a 19:59 (Figura 5.10e), percebemos que a maioria dos acidentes estão concentrados nas categorias *Food* e *Shop & Service*, e estão relacionados aos fatores *Motorista não condicionado* e *Direção imprópria*. Além disso, o número de mortes de motoristas e pedestres é maior em relação aos períodos

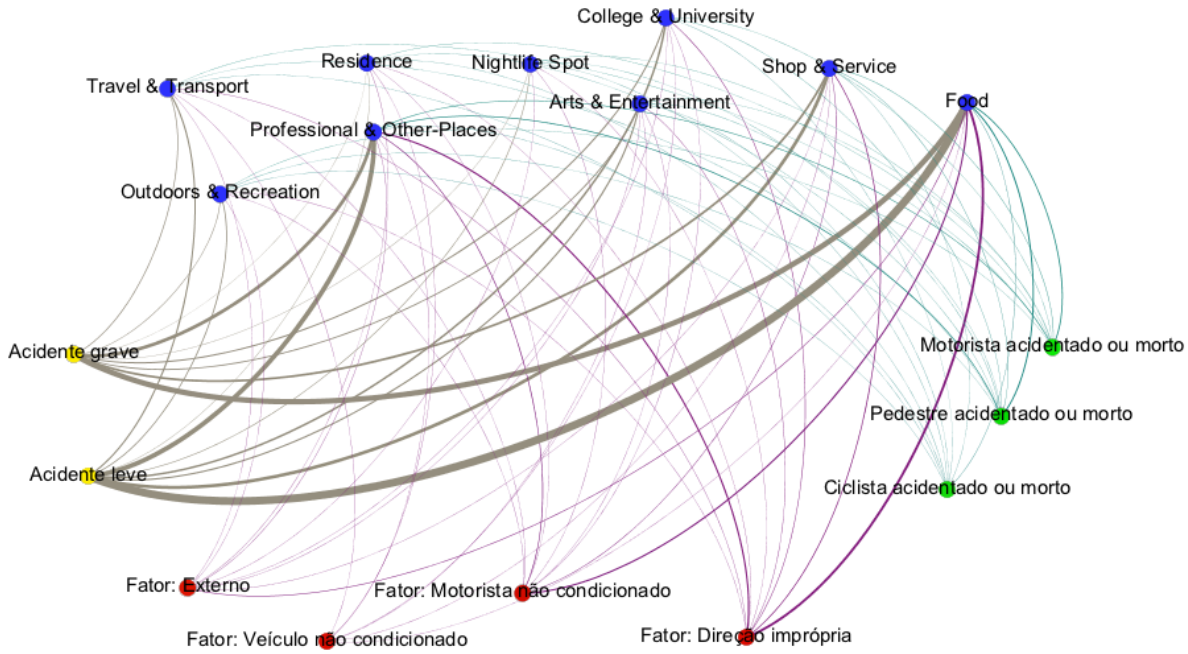
anteriores. Um padrão semelhante é apresentado na Figura 5.10f, porém apenas com a categoria *Food*.



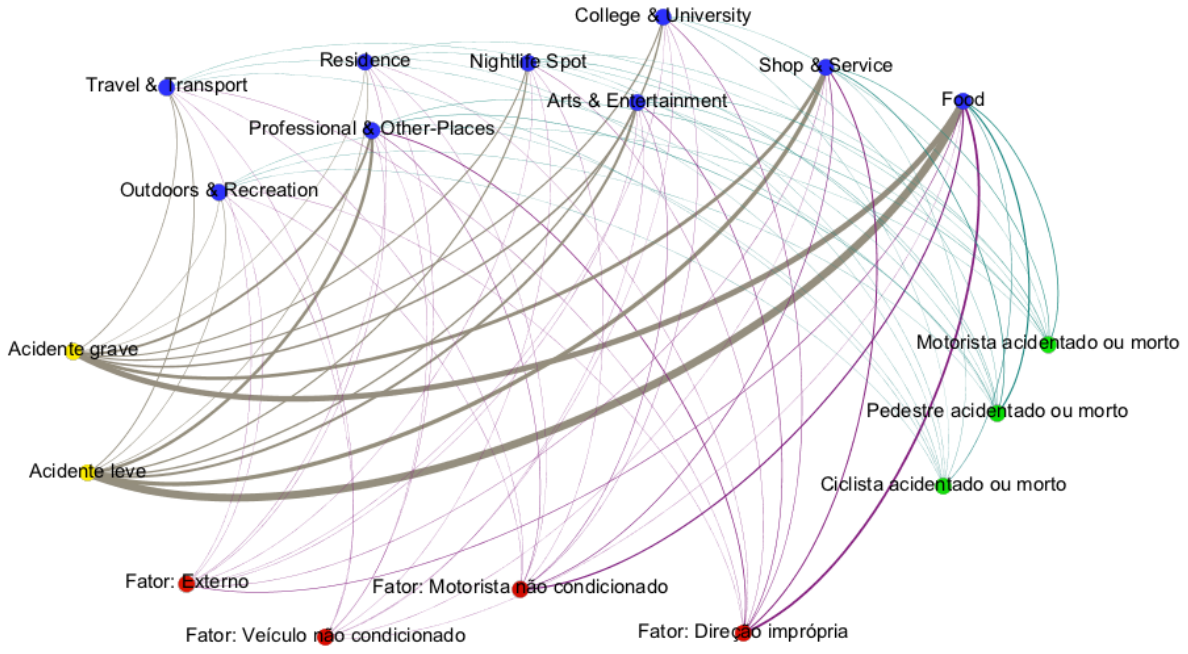
(a) 00:00 a 03:59



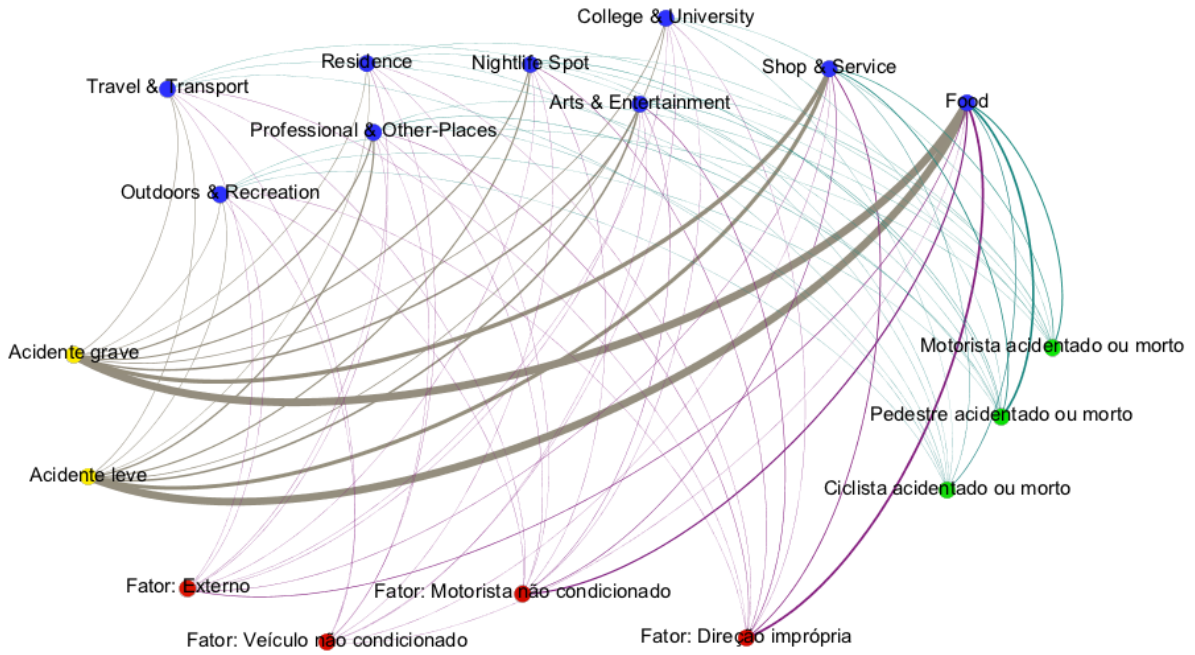
(b) 04:00 a 07:59



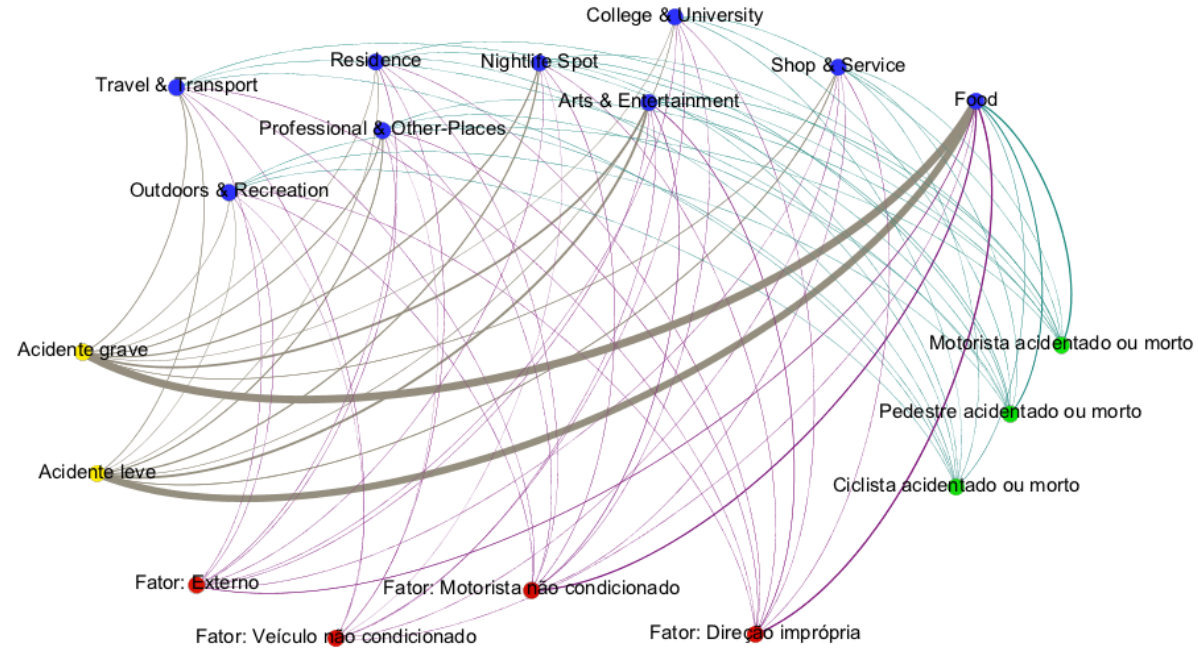
(c) 08:00 a 11:59



(d) 12:00 a 15:59



(e) 16:00 a 19:59



(f) 20:00 a 23:59

Figura 5.10: Representação da associação entre as regiões funcionais e os aspectos relacionados a acidentes de trânsito ao longo do dia.

Esta aplicação é capaz de identificar os aspectos relacionados aos acidentes de trânsito e o que eles ocasionam, e pode ser utilizada para auxiliar na definição de

soluções que minimizem os mesmos, além de possibilitar um melhor gerenciamento de recursos.

5.4.2 Aplicação II: Monitoramento dinâmico da cidade

Com a detecção de regiões funcionais por meio de redes sociais, tornou-se possível que o levantamento de informações necessárias para o planejamento urbano fosse realizado dinamicamente, conforme a evolução das cidades. Além disso, este método facilita a execução desta tarefa, uma vez que os dados dos censos não são estabelecidos com baixo custo. Outro fator importante é que os levantamentos necessários são realizados entre longos intervalos de tempo, de 2 a 4 anos, por exemplo. Com os dados provenientes de redes sociais, as regiões funcionais podem ser definidas e alteradas conforme os usuários que habitam a região compartilham suas informações.

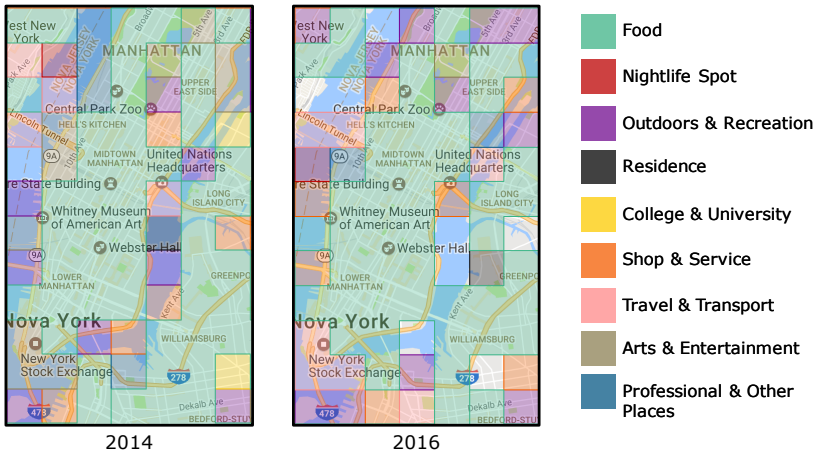
Para mostrar como este monitoramento dinâmico é viável, apresentamos a Figura 5.11, na qual é feita uma comparação da formação das regiões funcionais entre os anos de 2014 e 2016. Para isto, foram coletados 70.339 *check-ins* do Foursquare entre julho de 2016 e agosto de 2016 na mesma área urbana.

Podemos perceber, na Figura 5.11a, que a categoria *Food* prevalece sobre as demais tanto no ano de 2014, quanto no ano de 2016. Isto é ocasionado devido ao horário em que as regiões funcionais foram formadas, de 12:00 a 15:59. Além disso, a presença de regiões funcionais com a categoria *Outdoors & Recreation* aumentou no ano de 2016 em relação ao ano de 2014. Também verificamos que, nas duas regiões funcionais em que anteriormente prevalecia a categoria *College & University*, agora prevalecem as categorias *Shop & Service* e *Outdoors & Recreation*. Estas mudanças podem ocorrer devido a expansão da cidade, construção de novos conjuntos residenciais e a criação de novos empreendimentos, por exemplo.

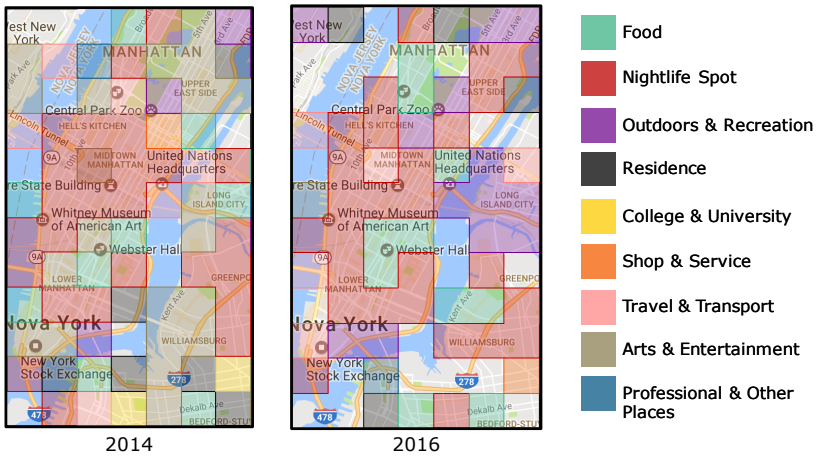
Na Figura 5.11b, a comparação é realizada em um horário noturno, de 00:00 a 03:59. Neste horário, também verificamos o aumento na aparição de regiões funcionais com a categoria *Outdoors & Recreation*, apesar da categoria *Nightlife Spot* prevalecer sobre as demais. A categoria *College & University* também foi eliminada.

Apesar das categorias *Food* (Figura 5.11a) e *Nightlife Spot* (Figura 5.11b) prevalecerem como as principais em seus respectivos horários, é possível verificar que algumas regiões mudaram suas funções ao longo do tempo.

Desta forma, com esta aplicação, verificamos uma forma alternativa, de baixo custo, para obter as informações necessárias para o auxílio ao planejamento de uma área urbana. Isto torna possível não só que este planejamento seja realizado de forma



(a) 12:00 a 15:59



(b) 00:00 a 03:59

Figura 5.11: Comparação entre as regiões funcionais dos anos de 2014 e 2016 em diferentes períodos de tempo.

dinâmica, como viabiliza que áreas urbanas com poucos recursos também consigam estas informações.

5.5 Considerações Finais

Neste capítulo, apresentamos um método para a detecção de regiões funcionais através de redes sociais baseadas em localização. O método consiste na utilização de polígonos para a formação das regiões, além dos dados sociais e de intervalos de tempo. Desta forma, as regiões funcionais estão diretamente relacionadas a rotina dos habitantes da área urbana estudada, capturando inclusive sua dinâmica.

Como experimento, o método foi aplicado à cidade de Nova Iorque, e os resul-

tados obtidos foram comparados com dados governamentais oficiais disponíveis. Com isso, apuramos que o método proposto obteve uma acurácia de 86% na definição das regiões funcionais, o que o torna uma alternativa viável para o auxílio no planejamento urbano em outras cidades onde não há tal monitoramento de forma sistemática. Suas principais vantagens são o pequeno custo de monitoramento e a possibilidade de observar o comportamento dinâmico das cidades e seus habitantes em intervalos de tempos muito menores que os métodos tradicionais.

Por fim, apresentamos duas aplicações para a qual as regiões funcionais podem ser utilizadas, além do planejamento urbano. Em uma das aplicações, é apresentado um estudo sobre os aspectos de acidentes de trânsito, relacionando-os com as regiões funcionais encontradas. Na outra, é apresentado como o monitoramento dinâmico de uma área urbana pode ser realizado através do método proposto.

Ressaltamos que o método proposto foi desenvolvido de forma genérica e com parâmetros configuráveis, para que possa ser utilizado em diversos estudos e não apenas para o estudo de caso de acidentes de trânsito, que é o foco deste trabalho.

Capítulo 6

Agrupamento Incremental para a Detecção de Eventos em Redes Sociais

O Sensoriamento Social utiliza informações disponíveis em redes sociais que podem ser coletadas e analisadas para a detecção de eventos relevantes em uma área habitacional [Silva et al., 2014a]. Neste sentido, trabalhos recentes buscam processar as informações de *stream* do Twitter para a detecção de notícias [Sankaranarayanan et al., 2009, Phuvipadawat and Murata, 2010], descoberta de eventos desconhecidos [Becker et al., 2011, Li et al., 2012] e detecção de eventos específicos, como terremotos e trânsito [Sakaki et al., 2010, Nguyen et al., 2016].

As soluções nesta área devem considerar limitações como confiabilidade das informações e dados não estruturados, com ruídos ou redundantes [Sakaki et al., 2010, Boettcher and Lee, 2012]. Desta forma, é comum utilizar técnicas de aprendizagem de máquina tanto no segmento supervisionado quanto não supervisionado para a análise das informações. Particularmente, técnicas de agrupamento são comumente aplicadas para a detecção de eventos, pois resolvem o problema de redundância dos dados através do agrupamento de informações semelhantes relacionadas ao mesmo evento.

Trabalhos na literatura normalmente utilizam métodos para o processamento e análise de uma grande massa de dados de redes sociais. O problema é que os dados destas redes são dinâmicos, de forma que as informações podem ser publicadas por diversos usuários a qualquer momento e em diferentes formatos. Por exemplo, na detecção de acidentes de trânsito, os usuários podem reportar os eventos conforme passam na mesma área da ocorrência e de acordo com o impacto do mesmo. Assim, com o objetivo de capturar eventos, estes trabalhos vão sacrificar o tempo de detecção do

evento para esperar até que uma quantidade significativa de dados possa ser processada. Outra possibilidade é que ocorra alta demanda de processamento devido à utilização sucessiva dos algoritmos, conforme a atualização da base de dados.

Neste contexto, a principal contribuição do método proposto neste capítulo é a utilização de um algoritmo de agrupamento incremental, que é mais viável de ser utilizado no cenário de tempo real das redes sociais, em contraposição aos métodos estáticos tradicionais. Além disso, o método proposto diferencia-se de trabalhos anteriores por permitir a detecção de diferentes ocorrências de um evento relacionado ao mesmo assunto, em vez de agrupar os dados sociais em diferentes categorias (exemplo: notícias e tendências de tópicos).

Neste trabalho, realizamos um estudo de caso relacionado a detecção de ocorrências de acidentes de trânsito e apresentamos uma similaridade de 90% nas mesmas, enquanto reduzimos o tempo de processamento dos dados.

6.1 Definição do problema

A detecção de eventos por meio de redes sociais tornou-se importante devido à disseminação de informações em tempo real, diferentemente das fontes de informações tradicionais [Sankaranarayanan et al., 2009]. Isso ocorre porque estas informações são reportadas assim que acontecem, através de usuários que são afetados pelo evento, ou através dos usuários que replicam informações por meio de mecanismos como o *retweet*. Apesar desta vantagem, os dados de redes sociais possuem ruídos (erros de gramática, aglutinação de palavras, uso de gírias, etc.) e seu processamento é trabalhoso em relação às fontes de informações oficiais.

Alinhado a estes fatores, os seguintes aspectos devem ser considerados para a detecção de eventos em redes sociais:

- Ocorrências relacionadas devem ser agrupadas, de forma que apenas um evento seja detectado. Isto deve ser considerado devido às informações relacionadas, que são reportadas e replicadas em redes sociais por múltiplos usuários, semelhante a uma rede de sensores sem fio que coleta dados de um mesmo objeto de estudo;

- O algoritmo de detecção de eventos deve ser incremental, de forma que a adição de novos dados não influencie na performance do algoritmo ao longo do tempo, tornando sua utilização viável em bases de dados massivas, que são atualizadas continuamente.

6.2 Agrupamento Incremental

O processo do algoritmo de agrupamento incremental (Algoritmo 1) consiste em verificar a distância entre os dados e unificar dados próximos. A distância máxima entre dois dados está presente na variável *eps-neighborhood*. Quando os *clusters* são formados, os dados que não pertencem a nenhum *cluster* são considerados ruídos pelo algoritmo. Em outras palavras, o algoritmo verifica a densidade dos *clusters*. Os de alta densidade são considerados informações de um mesmo evento, enquanto os de baixa densidade são considerados ruídos. Esta característica do algoritmo faz com que o mesmo seja viável para ser utilizado na abordagem de Sensoriamento Social, dado que as informações podem conter palavras-chave relacionadas ao estudo de caso, mas podem não pertencer a um evento.

No algoritmo, o *eps-neighborhood* de um dado p, denotado por $N_{Eps}(p)$ é definido $N_{Eps}(p) = q \in D|dist(p,q) \leq Eps$, onde D é a base de dados. Além disso, o algoritmo precisa que para cada dado p em um *cluster* C exista um dado q em C, de forma que a distância entre p e q seja menor que o *eps-neighborhood* e que $N_{Eps}(q)$ contenha no mínimo $MinPts$ dados.

Basicamente, nós inserimos um conjunto de novos dados em uma base de dados previamente agrupada. Devido à natureza de densidade do algoritmo, a atualização influencia apenas dados próximos. Neste processo, dados classificados como ruídos em um agrupamento prévio, podem passar a ser considerados informações, devido às mudanças nos centróides dos *clusters*.

Em nossa solução, aplicamos os conceitos de algoritmo incremental baseado em densidade na biblioteca *Scikit-Learn*, modificando a implementação do algoritmo DBSCAN[1] [Ester et al., 1996] seguindo os passos apresentados no Algoritmo 1. Basicamente, inserimos um conjunto de novos dados em uma base previamente agrupada. Em seguida, atualizamos os *clusters* mais próximos aos novos dados. Neste processo, dados classificados como ruídos podem ser classificados como informações de um mesmo evento devido à mudança dos centróides.

Ressaltamos que, em nossa abordagem, o uso de outros algoritmos de agrupamento podem ser utilizados, como o BIRCH [Zhang et al., 1996], desde que sejam adaptados para serem incrementais.

No Sensoriamento Social, uma abordagem incremental é necessária devido a quantidade de dados massiva provenientes das redes sociais. Utilizando uma abordagem estática, todos os dados são processados novamente mesmo que apenas poucos dados sejam inseridos na base, o que não é viável devido às atualizações em tempo real.

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html`

Algoritmo 1 Algoritmo de Agrupamento Incremental

```
Entrada: Conjunto de novos dados, eps, MinPts
Saída: Lista de clusters atualizada
D ← Novos dados
C ← Lista de clusters
para cada novo dado d_i em D faça
    Insere d_i na base
    se para cada centróide dist(centróide, d_i) ≤ eps então
        Atualiza centróide
    fim se
    Atualiza clusters em C
fim para
para cada cluster atualizado c_i em C faça
    se |N_Eps(dado central em c_i)| ≤ MinPts então
        Define c_i como um cluster válido
    senão
        Define c_i como ruído
    fim se
fim para
```

6.3 Experimentos e Resultados

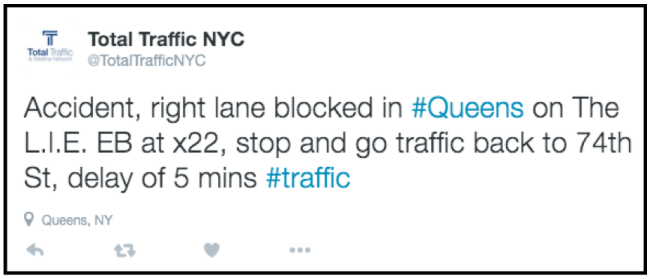
Nesta seção, descreventos os experimentos, resultados e avaliações do algoritmo de agrupamento incremental para a detecção de eventos. Também apresentamos um estudo de caso em que aplicamos a solução proposta para a identificação e contagem de eventos relacionados a acidentes de trânsito.

6.3.1 Descrição dos Dados

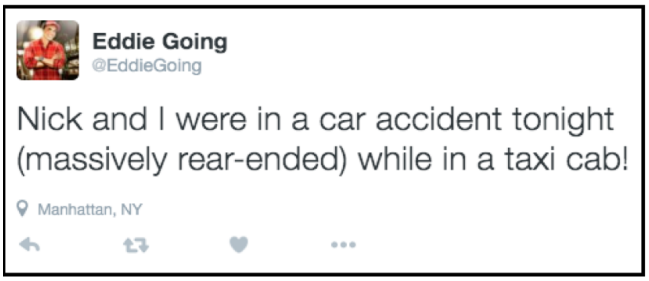
Para estudo de caso e validação do método, coletamos 135.078 *tweets* de março de 2015 a setembro de 2016, relacionados a acidentes de trânsito na cidade de Nova Iorque.

Conforme apresentado na Figura 6.1, informações coletadas de redes sociais possuem diferentes formatos e características. No Twitter, enquanto os *tweets* de usuários comuns não possuem um padrão, *tweets* de canais de notícia possuem sempre a mesma formatação, na qual são reportados o local (regiões e rotas) e o tipo de ocorrência. Em decorrência disso, diversas soluções utilizam apenas os dados provenientes de canais de notícia para a detecção de eventos [Albuquerque et al., 2015].

Para a solução proposta, *tweets* de canais de notícia são mais relevantes do que os *tweets* de usuários comuns, devido à etapa de cálculo da frequência dos termos. Na Tabela 6.1, apresentamos um exemplo do grau de importância de alguns termos. De maneira geral, termos relacionados às vias são considerados os mais relevantes,



(a) *Tweet* de canal de notícia



(b) *Tweet* de usuário comum

Figura 6.1: Exemplos de *tweets* relacionados a acidentes de trânsito.

enquanto termos comuns são considerados menos relevantes. Desta forma, sabemos que as vias são as características mais importantes para o algoritmo de agrupamento incremental.

Termos relevantes	Frequência de Termos	Termos não relevantes	Frequência de Termos
I-287	4.9512	EB (Eastbound)	2.1180
I-495	4.9512	Blocked	2.3486
8th (Ave)	4.9512	Street	2.3486
Route-24	4.9512	Lane	2.4663
14th (Ave)	4.9512	SB (Southbound)	2.5089

Tabela 6.1: Relevância dos termos presentes nos dados de acidentes de trânsito do Twitter por meio de frequência de termos.

Antes de aplicar o algoritmo que calcula a frequência dos termos, realizamos algumas modificações nos dados, como a remoção de *stop words*.

6.3.2 Avaliação de similaridade

Com o objetivo de avaliar a similaridade do algoritmo de agrupamento, realizamos o agrupamento manual de uma amostra da base de dados com 3.410 *tweets* de 27/01/2016 a 21/02/2016). Desta forma, foi possível comparar os agrupamento por dia e hora.

Na Tabela 6.2, apresentamos um exemplo de ocorrência de acidente de trânsito em que os dados foram agrupados manualmente no mesmo *cluster*. No processo de agrupamento manual, consideramos certas características, como as regiões, vias e horário da ocorrência reportada.

Hora	Ocorrências agrupadas
15:20	Accident in #TheBronx:OnTheDeeganExpwy on I-87 NB at W 230th St
16:30	Major accident on I-87 N #NYCTraffic
16:41	Accident in #TheBronx:OnTheDeeganExpwy on I-87 NB between W 230th St and Van Cortlandt Park S
17:01	Major accident on I-87 N #NYCTraffic
17:23	Accident in #TheBronx:OnTheDeeganExpwy on I-87 NB between W 230th St and Van Cortlandt Park S
17:31	Accident on I-87 N #NYCTraffic

Tabela 6.2: Exemplo de dados sociais agrupados relacionados à mesma ocoorrência de evento de acidente de trânsito.

Durante este processo, verificamos que os informes de acidentes eram relacionados com vias expressas e avenidas. Além disso, verificamos que o maior número de informes ocorria na hora do *rush*. Isto pode ser explicado pelo fato de que usuários geralmente reportam eventos em redes sociais que são importantes ou tem impacto em seu contexto. Desta forma, acidentes de trânsito que ocorrem nas vias mais utilizadas e em horários de pico, tendem a ser mais impactantes do que outros acidentes.

Na Figura 6.2, apresentamos a comparação de ocorrências de acidentes de trânsito entre o agrupamento manual e o agrupamento incremental. Com o objetivo de comparar os agrupamentos, aplicamos a métrica *V-Measure*, obtendo 90% de similaridade.

Após a comparação, analisamos em diferentes períodos do dia o tempo médio utilizado pelo algoritmo para que os *clusters* sejam formados (Figura 6.3). Basicamente, estes tempos médios são obtidos pela diferença de tempo entre a primeira notificação do acidente, geralmente considerada um ruído por ser diferente dos demais, até a formação do *cluster*. Assim, verificamos que eventos de alto impacto, que em sua maioria ocorrem em horários de pico, são formados em menor tempo que os demais eventos de acidentes de trânsito.

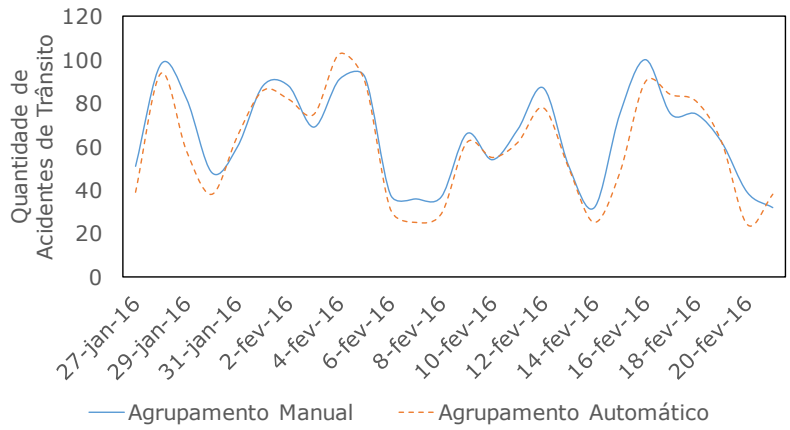


Figura 6.2: Comparação das ocorrências de acidentes de trânsito identificadas no agrupamento incremental e no agrupamento manual.

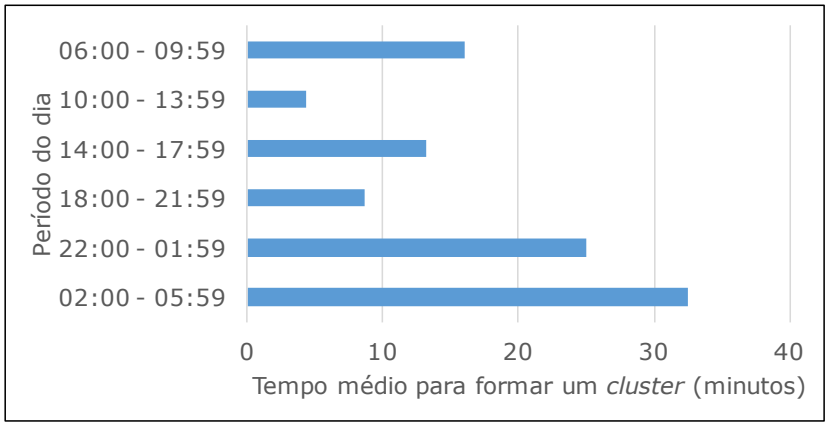


Figura 6.3: Tempo médio de formação dos *clusters* em diferentes períodos do dia.

Para exemplificar a formação de um *cluster* ao longo do tempo, apresentamos a Figura 6.4, em que dividimos a formação em 6 etapas, de forma que cada *cluster* é representado por uma cor diferente. Já os dados considerados ruídos são representados por pequenos pontos pretos. Nas Figuras 6.4b e 6.4c é possível verificar o momento em que o algoritmo converte dados considerados ruídos (círculo azul pontilhado) em um *cluster* que representa uma ocorrência de acidente de trânsito após a inserção de novos dados na base. No entanto, verificamos que em alguns casos o dado permanece como ruído mesmo com a inserção de novos dados (círculo vermelho pontilhado), o que ocorre devido à falta de similaridade destes com os demais dados da base (Figuras 6.4b a 6.4f).

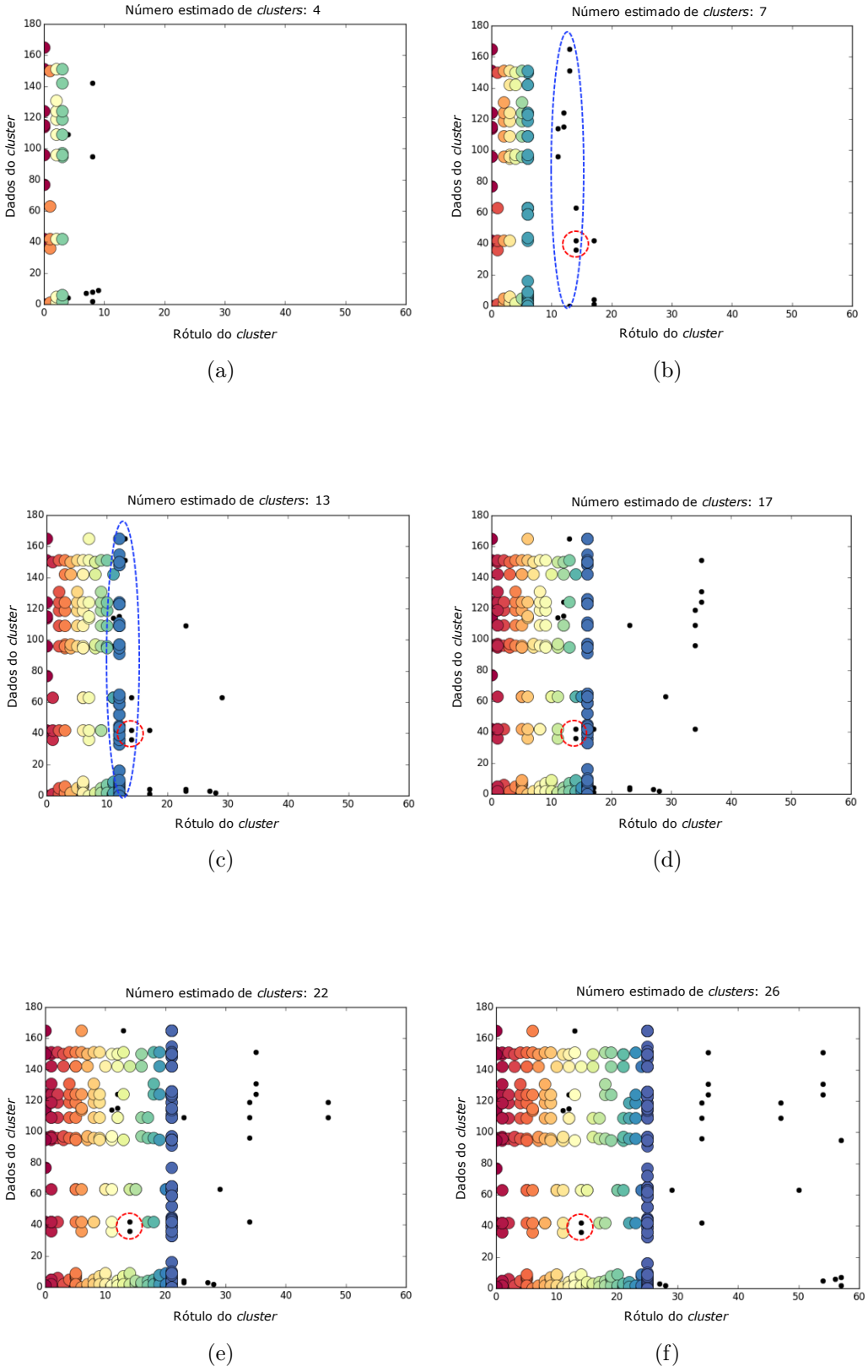


Figura 6.4: Exemplo de como ocorre a formação de *clusters* ao longo do tempo utilizando o método proposto.

6.3.3 Avaliação do tempo de execução

Para avaliar o tempo de execução do algoritmo, aplicamos tanto a abordagem incremental quanto a abordagem estática no estudo de caso de detecção de acidentes de trânsito. Cada vez que havia inserção de novos dados, o algoritmo de agrupamento estático processava todos os dados da base, enquanto que o algoritmo de agrupamento incremental processava apenas os novos dados e os dados da base afetados pela inserção.

Conforme apresentado na Figura 6.5, este comportamento impactou no tempo de execução, de forma que, com 135.078 *tweets* na base, o algoritmo de agrupamento incremental executava cerca de 4 vezes mais rápido que o estático. Desta forma, verificamos que para abordagens relacionadas às redes socais utilizar algoritmos tradicionais é custosa em termos de tempo, sendo necessária uma abordagem incremental.

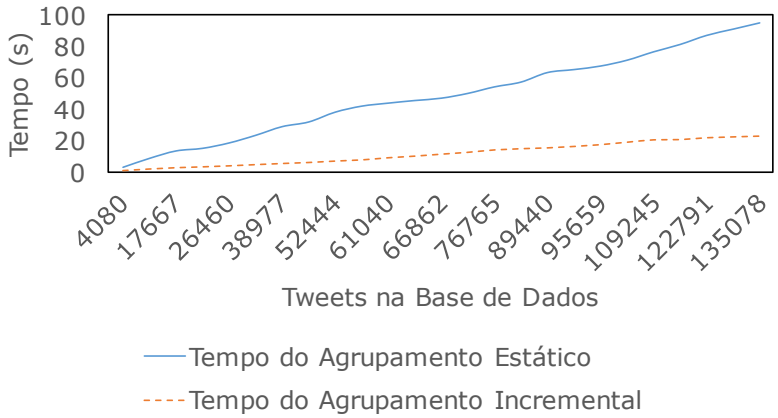


Figura 6.5: Comparação entre o tempo de agrupamento das abordagens incremental e estática.

6.4 Considerações Finais

Neste capítulo, nós investigamos o processo de detecção de eventos em tempo real nas redes sociais, através de uma abordagem de agrupamento incremental. Na solução proposta, consideramos cada usuário da rede social Twitter como um sensor que compartilha dados contextuais voluntariamente.

Todos os dados coletados foram filtrados para evitar o máximo de ruídos. As características dos textos presentes nos dados foram extraídas através de um algoritmo que calcula a frequência dos termos. Em seguida, um algoritmo de agrupamento incremental foi aplicado na base de dados de acordo com a inserção de novos dados. Como estudo de caso, executamos a solução proposta em uma base de dados do Twitter para

a detecção de acidentes de trânsito, que possuem relevância social devido aos danos e mortes causados por eles.

Dentre os principais desafios para a criação e validação do método, estão os dados ruidosos do Twitter, decorrentes do limite no número de caracteres em uma mensagem, assim como a diferenciação entre eventos que ocorrem em um mesmo horário. Apesar disso, nossa solução obteve 90% de similaridade e tempo de processamento 4 vezes mais rápido que a abordagem tradicional.

Capítulo 7

Conclusão

Neste trabalho, abordamos o problema de Sensoriamento Social, para o monitoramento e a caracterização de acidentes de trânsito. Dentre os desafios deste problema, podemos citar a limitação relacionada à proveniência dos dados necessários para a solução proposta através de redes sociais. Por este motivo, os estudos de caso apresentados neste trabalho seguiram um método no qual os dados coletados foram analisados e validados de acordo com a similaridade com o problema estudado e com as áreas urbanas as quais pertenciam.

Na solução, utilizamos as informações fornecidadas por habitantes de uma determinada região, além de dados providos pelo governo, para o monitoramento e a caracterização de acidentes. O desenvolvimento da solução foi realizado através de diversas etapas, desde a coleta de dados até a extração de características relacionadas as regiões e a acidentes de trânsito. Os resultados obtidos foram analisados e comparados com dados dados reais.

Através do Sensoriamento Social, esperamos prover uma alternativa para a obtenção de dados históricos relacionados a acidentes de trânsito, com o objetivo de identificar áreas de risco e características dos acidentes, para tornar possível a proveniência de soluções relacionadas à segurança no trânsito em uma determinada localidade. Conforme apresentado anteriormente, em algumas regiões, os dados relacionados a estes acidentes são limitados ou dispersos, dificultando a tomada de decisões ou ocasionando decisões ineficientes.

Apesar do foco deste trabalho ser os acidentes de trânsito, os métodos criados para atingir este objetivo são genéricos e configuráveis, de forma que podem ser utilizados para outras aplicações.

7.1 Publicações

Como resultado deste trabalho, publicamos o artigo *PoI: uma Aplicação de Detecção de Pontos de Interesse* no 13º Simpósio Brasileiro de Sistemas Colaborativos [Menezes et al., 2016], onde utilizamos a rede social baseada em localização Foursquare para detectar Pontos de Interesse em capitais brasileiras (vide Seção 5.2).

Além disso, temos o artigo *Um Método de Detecção de Regiões Funcionais Utilizando Dados de Redes Sociais* (vide Seções 5.3 e 5.4), aceito no XXXV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos. Já o artigo *Um Método de Agrupamento Incremental para a Detecção de Eventos em Redes Sociais* (vide Capítulo 6) foi aceito no 14º Simpósio Brasileiro de Sistemas Colaborativos.

Por fim, listamos os artigos publicados no 22º *Brazilian Symposium on Multimedia and the Web* em colaboração com membros do grupo de pesquisa de computação móvel e ubíqua (LCMU) da Universidade Federal do Amazonas (UFAM):

- *For or Against?: Polarity Analysis in Tweets About Impeachment Process of Brazil President* [Souza et al., 2016];
- *Sentiment Analysis of Portuguese Comments from Foursquare* [Almeida et al., 2016].

7.2 Limitações do Método

Através dos resultados obtidos, verificamos que o método proposto possui algumas limitações, sendo elas:

- A detecção de regiões funcionais é dependente de dados geolocalizados e das categorias do Foursquare, mas pode ser aplicada em outras redes sociais que possuem a mesma característica;
- Como a área das regiões funcionais utilizadas é menor do que a área das *tracts*, não foi possível associar os dados destas regiões para melhor análise da nossa solução;
- A deteção de ocorrências de acidentes de trânsito e das regiões funcionais, torna-se possível apenas em áreas urbanas em que há compartilhamento de dados dos usuários em redes sociais;

- O monitoramento de acidentes de trânsito nos finais de semana, por meio de redes sociais, possui poucos dados em relação a semana, o que pode dificultar a detecção de ocorrências neste período.

7.3 Trabalhos Futuros

Como trabalhos futuros para os resultados apresentados sobre regiões funcionais, pretendemos analisar outras redes sociais baseadas em localização, que podem complementar os resultados encontrados, tendo em vista que diferentes países e cidades podem utilizar mais uma rede social do que outra. Além disso, iremos demonstrar através de outras aplicações a extensão deste estudo para avaliações econômicas, sociais e de mobilidade urbana.

Em relação à detecção de eventos, planejamos analisar outras bases de dados de redes sociais com diferentes características, com o objetivo de avaliar a performance do algoritmo de agrupamento incremental.

Por fim, a partir dos resultados deste trabalho, podemos inferir as melhores regiões para que sejam implantadas soluções que auxiliem na prevenção de acidentes de trânsito, assim como no gerenciamento de recursos em um determinado perímetro urbano.

Referências Bibliográficas

F. C. Albuquerque, M. A. Casanova, H. Lopes, L. R. Redlich, J. A. F. de Macedo, M. Lemos, M. T. M. de Carvalho, and C. Renso. A methodology for traffic-related twitter messages interpretation. *Computers in Industry*, 2015.

T. G. Almeida, B. A. Souza, A. A. F. Menezes, C. Figueiredo, and E. F. Nakamura. Sentiment analysis of portuguese comments from foursquare. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, Webmedia '16, pages 355–358. ACM, 2016. ISBN 978-1-4503-4512-5. doi: 10.1145/2976796.2988180.

H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *International Conference on Weblogs and Social Media*. AAAI, 2011.

A. Boettcher and D. Lee. Eventradar: A real-time local event detection scheme using twitter stream. In *International Conference on Green Computing and Communications (GreenCom)*, pages 358–367. IEEE, 2012.

CNM. Mapeamento das mortes por acidentes de trânsito no brasil. 2009.

T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

J. Cranshaw, R. Schwartz, J. Hong, and N. Sadeh. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2012.

O. L. de Morais Neto, M. de Mesquita Silva Montenegro, R. A. Monteiro, J. ao Bosco Siqueira Júnior, M. M. A. da Silva, C. M. de Lima, L. O. M. Miranda, D. C. Malta, and J. B. da Silva Junior. Mortalidade por acidentes de transporte terrestre no brasil na última década: tendência e aglomerados de risco. In *Ciência e Saúde Coletiva*, pages 2223–2236, 2012.

M. F. de Oliveira. O impacto financeiro dos acidentes de trânsito para o sus (sistema único de saúde) no brasil, 2010.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

L. Figueiredo, I. Jesus, J. A. T. Machado, J. Ferreira, and J. L. M. de Carvalho. Towards the development of intelligent transportation systems. In *Intelligent Transportation Systems*, volume 88, pages 1206–1211, 2001.

V. Frias-Martinez, V. Soto, H. Hohwald, and E. Frias-Martinez. Characterizing urban landscapes using geolocated tweets. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 239–248. IEEE, 2012.

D. Karamshuk, A. Noulas, S. Scellato, V. Nicosia, and C. Mascolo. Geo-spotting: mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 793–801. ACM, 2013.

N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Campbell. A survey of mobile phone sensing. *Communications Magazine, IEEE*, 48(9):140–150, 2010.

R. Lee and K. Sumiya. Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks*, pages 1–10. ACM, 2010.

R. Li, K. H. Lei, R. Khadiwala, and K. C.-C. Chang. Tedas: A twitter-based event detection and analysis system. In *International Conference on Data Engineering (ICDE)*, pages 1273–1276. IEEE, 2012.

S. Marsland. *Machine Learning: An Algorithmic Perspective*. CRC Press, 2009.

A. Menezes, T. Almeida, B. Gatto, E. Santos, C. Figueiredo, and E. Nakamura. Poi: uma aplicação de detecção de pontos de interesse. In *Proceedings of the 13th Simpósio Brasileiro de Sistemas Colaborativos (SBSC)*, 2016.

H. Nguyen, W. Liu, P. Rivera, and F. Chen. Trafficwatch: Real-time traffic incident detection and monitoring using social media. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 540–551. Springer, 2016.

A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *Proceedings of the 15th International AAAI Conference on Weblogs and Social Media*, 2011a.

A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *Proceedings of the 15th Internation AAAI Conference on Weblogs and Social Media*, 2011b.

S. Phuvipadawat and T. Murata. Breaking news detection and tracking in twitter. In *International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 120–123. IEEE, 2010.

A. I. J. T. Ribeiro, T. H. Silva, F. Duarte-Figueiredo, and A. A. F. Loureiro. Studying traffic conditions by analyzing foursquare and instagram data. In *Proceedings of the 11th ACM symposium on Performance evaluation of wireless ad hoc, sensor, & ubiquitous networks*, pages 17–24. ACM, 2014.

T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th acm sigspatial international conference on advances in geographic information systems*, pages 42–51. ACM, 2009.

J. Shi and J. Malik. Normalized cuts and image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 1997.

T. H. Silva, P. O. V. de Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. Visualizing the invisible image of cities. In *Green Computing and Communications (GreenCom), 2012 IEEE International Conference on*, pages 382–389. IEEE, 2012.

T. H. Silva, P. O. S. V. de Melo, A. C. Viana, J. M. Almeida, J. Salles, and A. A. F. Loureiro. Traffic condition is more than colored lines on a map: Characterization of waze alerts. In *Social Informatics*, pages 309–318. Springer, 2013a.

T. H. Silva, P. O. S. Vaz de Melo, J. M. Almeida, and A. A. Loureiro. Uma fotografia do instagram: Caracterização e aplicação. *Proc. of XXXII SBRC13*, 2013b.

T. H. Silva, P. O. Vaz de Melo, J. M. Almeida, J. Salles, and A. A. Loureiro. Revealing the city that we cannot see. *ACM Transactions on Internet Technology (TOIT)*, 14 (4):26, 2014a.

T. H. Silva, P. O. S. Vaz De Melo, J. M. Almeida, A. C. Viana, J. Salles, A. Loureiro, et al. Participatory sensor networks as sensing layers. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*, pages 386–393. IEEE, 2014b.

T. H. Silva, P. O. S. V. de Melo, J. B. B. Neto, A. I. J. T., C. S. F. d. S. Ribeiro, V. F. S. Mota, F. D. da Cunha, A. P. G. Ferreira, K. L. d. S. Machado, R. A. d. F. Mini, et al. Redes de sensoriamento participativo: Desafios e oportunidades. *Minicurso do XXXIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC)*, pages 266–315, 2015.

B. A. Souza, T. G. Almeida, A. A. F. Menezes, F. G. Nakamura, C. M. Figueiredo, and E. F. Nakamura. For or against?: Polarity analysis in tweets about impeachment process of brazil president. In *Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web*, Webmedia '16, pages 335–338, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4512-5. doi: 10.1145/2976796.2988216.

F. Toriumi and S. Baba. Real-time tweet classification in disaster situation. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 117–118. International World Wide Web Conferences Steering Committee, 2016.

G. Valkanas and D. Gunopulos. How the live web feels about events. In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pages 639–648. ACM, 2013.

F. Xiang, L. Tu, B. Huang, and X. Yin. Region partition using user mobility patterns based on topic model. In *Proceedings of the 16th IEEE International Conference on Computational Science and Engineering (CSE)*, 2013.

K. Zhang, Q. Jin, K. Pelechrinis, and T. Lappas. On the importance of temporal dynamics in modeling urban activity. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, 2013.

T. Zhang, R. Ramakrishnan, and M. Livny. Birch: an efficient data clustering method for very large databases. In *ACM Sigmod Record*, volume 25, pages 103–114. ACM, 1996.

Y. Zhi, H. Li, D. Wang, M. Deng, S. Wang, J. Gao, Z. Duan, and Y. Liu. Latent spatio-temporal activity structures: a new approach to inferring intra-urban functional regions via social media check-in data. *Geo-spatial Information Science*, pages 1–12, 2016.