Thiago Thalison Firmino de Lima

ANÁLISE DE DADOS DO TWITTER PARA INTERPRETAÇÃO DE TRAJETÓRIAS DE OBJETOS MÓVEIS

Trabalho de Concluão de Curso submetido ao Instituto de Informática e Estatística da Universidade Federal de Santa Catarina para a obtenção do Grau de Bacharel em Sistemas de Informação.
Orientador: Prof. Dra. Vania Bogorny
Universidade Federal de Santa Catarina

Florianópolis

2016

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Lima, Thiago Thalison Firmino de
ANÁLISE DE DADOS DO TWITTER PARA INTERPRETAÇÃO DE TRAJETÓRIAS DE OBJETOS MÓVEIS / Thiago Thalison Firmino de Lima ; orientadora, Vania Bogorny - Florianópolis, SC, 2016.
75 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico. Graduação em Sistema de Informação.

Inclui referências

1. Sistema de Informação. 2. Trajetórias de Objetos Móveis. 3. Interpretação de Trajetórias. 4. Extração de Informação. 5. Twitter. I. , Vania Bogorny. II. Universidade Federal de Santa Catarina. Graduação em Sistema de Informação. III. Título.

Thiago Thalison Firmino de Lima

ANÁLISE DE DADOS DO TWITTER PARA INTERPRETAÇÃO DE TRAJETÓRIAS DE OBJETOS MÓVEIS

Este Trabalho de Conclução de Curso foi julgado aprovado para a obtenção do Título de "Bacharel em Sistemas de Informação", e aprovado em sua forma final pelo Instituto de Informática e Estatística da Universidade Federal de Santa Catarina.

Florianópolis, 01 de Dezembro 2016.

Prof. Dr. Frank Siqueira
Universidade Federal de Santa Catarina
Coordenador

Banca Examinadora:

Prof. Dra. Vania Bogorny
Universidade Federal de Santa Catarina
Orientador

Prof. Dra. Luciana de Oliveira Rech
Universidade Federal de Santa Catarina

Prof. Dr. Renato Fileto
Universidade Federal de Santa Catarina

Prof. Dr. Frank Siqueira
Universidade Federal de Santa Catarina

Dedico o meu TCC aos meus queridos pais, familiares, amigos e professores que de alguma forma contribuiram para que eu concluísse meu curso de graduação. A todos vocês um sincero muito obrigado.

AGRADECIMENTOS

Agradeço a todos aqueles que contribuíram para que a conclusão deste trabalho se tornasse possível:

Aos meus pais, Paulo e Sirlene, por sempre estarem presentes, me apoiando emocionamente e financeiramente para que eu pudesse concluir esta fase essencial da minha vida.

Às minhas irmãs, Aryane e Tatiane, por serem grandes amigas com quem tenho certeza que posso contar.

À minha avó Irene por estar sempre acreditando em mim e me inspirando a seguir de cabeça erguida mesmo diante de situações adversas. E também a todos os meus familiares, por acreditarem em mim e me motivarem a conquistar meus objetivos.

À minha professora orientadora Dra. Vania Bogorny, pela dedicação nas orientações, disponibilizando tempo e material, que foi de extrema importância para a realização deste trabalho, também a indicação de leituras de autores clássicos e contemporâneos para o embasamento teórico e troca de experiências aliando teoria e prática para a elaboração dos resultados e conclusão deste trabalho.

Aos membros da banca avaliadora do meu TCC, professores Renato Fileto, Frank Siqueira e professora Luciana Rech. Obrigado pelas críticas, contribuições e sugestões que serviram de base para aprimoramento deste trabalho.

Aos meus amigos que durante esta jornada estiveram ao meu lado, nos momentos bons e ruins.

Ao programa Ciências sem Fronteiras que promoveu intercâmbio de graduação na Laurentian University no Canadá, investindo na minha formação acadêmica e proporcionando troca de experiência e conhecimentos tecnológicos e científicos. Também ao programa MITACS que proporcionou um estágio de pesquisa na Université Du Quebec no Canadá, onde tive a oportunidade de realizar pesquisa científica com excelentes pesquisadores.

Por fim, agradeço ao corpo docente do Departamento de Informática e Estatística da UFSC por todo o ensinamento que me foi dado, permitindo que eu conclua esta etapa importante da minha vida e carreira profissional.

Não permita que ninguém destrua seus sonhos. Corra atrás deles, pois eles definirão o tamanho de suas vidas.

(Roberto Shinyashiki)

RESUMO

A Análise de Trajetórias de Objetos Móveis é um campo da Ciência da Computação que busca a criação de algoritmos capazes de descobrir e analisar padrões de movimento. Um dos maiores desafios enfrentados por este tipo de análise está em como obter dados que permitam a contextualização de padrões identificados em trajetórias utilizando-se os algoritmos existentes. Dessa maneira, acredita-se que dados oriundos de redes sociais possam ser uma fonte interessante para solucionar este tipo de problema, já que elas geram uma grande quantidade de informações diariamente. A rede social Twitter possui contas de usuário especializadas na publicação de informações sobre um assunto específico e, dentre elas, aquelas direcionadas para a publicação de informações sobre as condições de trânsito em uma região. Neste trabalho é proposto um método para extrair e organizar essas informações de trânsito e também avaliar a viabilidade do uso delas na interpretação de trajetórias de veículos.

Palavras-chave: Trajetórias de Objetos Móveis. Interpretação de Trajetórias. Redes Sociais. Twitter. Extração de Informação.

ABSTRACT

The Moving Objects Trajectory Analysis is a Computer Science field that aims to develop algorithms for extracting different patterns from these trajectories. One of the major challenges associated with this kind of analisys is the gathering of information that enables a better interpretation of the patterns identified by the existing algorithms. In this manner, the data published in social networks might be an interesting source as a solution for this problem, once these networks generate a large variety of information. There are many accounts that focus on publising information regarding a specific topic on the Twitter social network and some of them aim attention at traffic information. This project introduces a method for extracting and organizing traffic information published on Twitter and the evaluation of their use as a source of context information for the interpretation of trajectories of vehicles.

Keywords: Moving Objects Trajectory Analysis. Trajectory Interpretation. Social Networks. Twitter. Information Extraction.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API	Application Interface Programming
CDC	U.S. Centers for Disease Control and Prevention
GPS	Global Positioning System
IE	Information Extraction
NER	Named Entity Recognition
NLP	Natural Language Processing
OSM	Open Street Maps
OSN	Online Social Network
REST	Representational State Transfer
RWR	Relevants Word Recognition
SQL	Structured Query Language
SRI	Sistema de Reconhecimento de Informação
SMoT	Stops and Moves of Trajectories
TRAOD	Trajectory Outlier Detection Algorithm
WEB	World Wide Web

SUMÁRIO

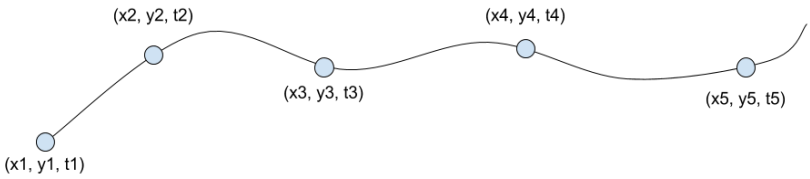
1 INTRODUÇÃO

A popularização de dispositivos com tecnologia GPS permite que indivíduos se localizem no espaço e também coletem informações acerca de seus movimentos como, por exemplo, o percurso seguido por um atleta durante uma corrida, o caminho percorrido por um taxista durante um atendimento ou até mesmo o controle da rota realizada por um caminhão durante a entrega de um produto. Muitas dessas informações encontram-se disponíveis publicamente, permitindo a realização de diferentes estudos.

A área de trajetórias de objetos móveis é o campo de pesquisa da Ciência da Computação que busca propor métodos que permitam fazer a análise desses dados através da extração de diferentes informações como o caminho mais utilizado entre duas regiões (FOSCA et al., 2007), a identificação de padrões geométricos entre diferentes movimentos (LAUBE et al., 2005), a identificação de desvios realizados em relação a um caminho padrão (LEE et al., 2008), entre outros.

A grande maioria das pesquisas realizadas na área de trajetórias são voltadas para a análise delas em sua forma bruta, ou seja, um conjunto de coordenadas geográficas coletadas em um intervalo de tempo. A figura 1 apresenta um exemplo de trajetória bruta, onde x e y representam as coordenadas geográficas e t o instante de tempo no qual elas foram coletadas.

Figura 1: Exemplo de Trajetória Bruta

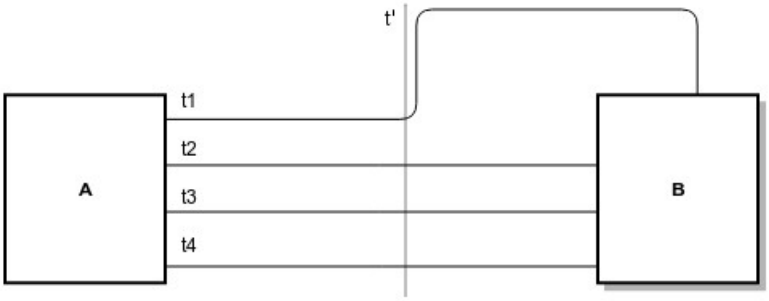


Os dados brutos de trajetórias limitam as interpretações que podem ser realizadas pelo fato de não possuírem informações do contexto no qual elas foram geradas, impossibilitando responder perguntas interessantes como: “O que teria levado um indivíduo a fazer um desvio de trajeto?”, “Quais os lugares visitados durante um percurso?”, “Qual seria o motivo para um grupo de indivíduos seguirem um trajeto comum?”, “Por que um motorista realizou uma manobra brusca?”, entre outras.

Um dos estudos pioneiros que buscam a agregação de informações de contexto na análise de trajetórias foi realizado por Alvares et al. (2007), com o método SMoT, que propõe o enriquecimento semântico das trajetórias através da identificação dos pontos de parada em um trajeto. Sendo assim, é possível identificar, por exemplo, que o ponto A e o ponto B nos quais um objeto parou por determinado período, são na verdade o restaurante X e o hotel Y, repectivamente.

A figura 2 ilustra um outro exemplo de situação em trajetórias de objetos móveis, que é a existência de um objeto que faz um desvio em relação a um caminho padrão, denominado na literatura de objeto *outlier* (FONTES et al., 2013). Nesta imagem, as áreas A e B representam regiões de uma cidade, e as linhas t1, t2, t3 e t4 representam trajetórias de veículos seguindo caminhos entre essas regiões. No instante de tempo t', o veículo da trajetória t1 faz um desvio para outra via, caracterizando-o como um *outlier*. No trabalho de Aquino (2014) foi proposto um método que estende o trabalho de Fontes et al. (2013), buscando entender o motivo para a realização deste desvio, melhorando assim a interpretação do desvio identificado.

Figura 2: Exemplo de desvio em trajetórias (*outliers*)



Um dos principais desafios relacionados à interpretação de dados de trajetórias está em como elencar fontes de informação que permitam obter dados de contexto que auxiliem nas interpretações realizadas. Neste cenário, acredita-se que dados publicados em redes sociais (OSNs - Online social Networks) possuem um potencial interessante para ser utilizado como este tipo de fonte, isto porque, apesar de serem relativamente recentes (criadas em meados dos anos 2000), são utilizadas por mais de dois terços da população que acessa a internet (N.O Report., 2013), o que gera um grande volume de informações diariamente. Um exemplo deste potencial é o trabalho de Beber (BEBER et. al, 2016),

que infere as atividades que um indivíduo realizou durante a trajetória através da análise de *tweets* postados em pontos de interesse.

O Twitter é uma rede social baseada na publicação de mensagens de até 140 caracteres criada em 2006, por Jack Dorsey, e popularizou-se entre os internautas por permitir a divulgação de informações acerca da vida pessoal deles para milhares de pessoas. Segundo uma pesquisa realizada pela empresa Monkey Business, circulam cerca de 1 bilhão de novas mensagens nesta rede por semana, o que a torna uma rica fonte para obtenção de informações.

Esta rede social tem como principal característica o fato de funcionar como um microblog, sendo comuns contas de usuários utilizadas para publicação de mensagens acerca de um tema específico, e dentre elas, existem contas utilizadas para publicação de informações sobre as condições de trânsito em uma cidade, como a conta "O dia 24 horas"no Rio de Janeiro, a "Trânsito SP"em São Paulo, a "Trânsito Zero Hora"em Porto Alegre, e a "Trânsito 24 horas"em Florianópolis. Essas informações de trânsito, se extraídas e bem estruturadas, podem ser úteis para a (*i*) anotação semântica de trajetórias, (*ii*) a interpretação do movimento das trajetórias, (*iii*) a interpretação de padrões de trajetórias, visto que muitas dessas trajetórias referem-se à trajetos percorridos por veículos pelas vias de uma cidade.

1.1 OBJETIVOS

O Objetivo deste trabalho é analisar dados do Twitter referentes à situações de trânsito em uma região e propor um método para extraí-los e estruturá-los de forma que possam ser utilizados como fonte de informação na interpretação de trajetórias de veículos que trafegam nesta mesma região. Os objetivos específicos incluem:

- Analisar como informações de trânsito são publicadas na rede social Twitter;
- Examinar as informações de trânsito encontradas e elencar os dados que precisam ser extraídos para que estas informações possam ser utilizadas com trajetórias de veículos;
- Selecionar técnicas que permitam desenvolver um método para fazer a extração e armazenamento desses dados;
- Aplicar e avaliar o método proposto;

- Avaliar a viabilidade de uso das informações de trânsito extraídas com o método proposto no contexto de trajetórias de veículos.

1.2 METODOLOGIA

A metodologia de pesquisa adotada neste trabalho foi dividida nas seguintes etapas:

- Realizar uma revisão bibliográfica sobre trabalhos relacionados ao tema desta monografia (trajetórias de objetos móveis, rede social Twitter, técnicas de extração de informações em texto);
- Estudar como informações de trânsito são publicadas na rede social Twitter;
- Definir os tipos de dados do Twitter que precisam ser extraídos, ou seja, os dados relevantes para que as informações de trânsito possam ser relacionadas com trajetórias;
- Elencar técnicas de extração de informação que permitam extrair do Twitter os dados considerados relevantes e criar um método para realizar esta extração;
- Elaborar uma estrutura que permita armazenar e recuperar as informações extraídas do Twitter de forma que possam ser posteriormente utilizadas para a interpretação de trajetórias;
- Avaliar o método proposto para extração de informações de trân-sito do Twitter através do uso das métricas de precisão e *recall*;
- Coletar um conjunto de trajetórias para analisar a viabilidade de uso das informações de trânsito extraídas do Twitter na interpretação dessas trajetórias;
- Validar um dos algoritmos propostos no trabalho de Aquino (2014) utilizando as informações de trânsito extraídas do Twit-ter;

1.3 ESCOPO E ESTRUTURA DO DOCUMENTO

O escopo deste trabalho inclui a extração, organização e armazenamento de informações sobre o trânsito em regiões de uma cidade publicadas na rede social Twitter e a análise da viabilidade de uso

dessas informações para a interpretação de trajetórias de veículos realizadas nessas regiões. O restante deste documento está organizado da seguinte forma: o capítulo 2 apresenta os conceitos básicos e trabalhos relacionados; o capítulo 3 apresenta uma proposta para extração de informações de trânsito publicadas na rede social Twitter; o capítulo 4 apresenta os experimentos realizados; e o capítulo 5 apresenta a conclusão e os trabalhos futuros.

2 CONCEITOS BÁSICOS E TRABALHOS RELACIONADOS

Neste capítulo são apresentados os principais conceitos e publicações científicas que fundamentam o desenvolvimento deste trabalho. Na seção 2.1 é apresentado o conceito de trajetórias de objetos móveis, bem como o problema de agregação de dados de contexto em trajetórias brutas. Na seção 2.2 é apresentada uma visão geral da rede social Twitter e as contas de usuário com domínio na divulgação de informações de trânsito. Na seção 2.3 são apresentados trabalhos que utilizaram informações publicadas no Twitter. Por fim, na seção 2.4 são apresentadas as técnicas de extração de informação em textos que serviram de base para este trabalho.

2.1 TRAJETÓRIAS DE OBJETOS MÓVEIS

A disseminação e o barateamento de tecnologias que permitem identificar o posicionamento de um objeto, como satélites, redes de sensores, e aparelhos que utilizam tecnologia GPS, aumentaram a possibilidade de geração de dados que registram a mobilidade desse objeto. Isto porque, eles podem registrar uma sequência de pontos relacionados ao posicionamento do objeto no tempo e no espaço. Sendo assim, uma vez que se tenha acesso a tais dados, é possível utilizá-los em pesquisas relacionadas à trajetórias de objetos móveis.

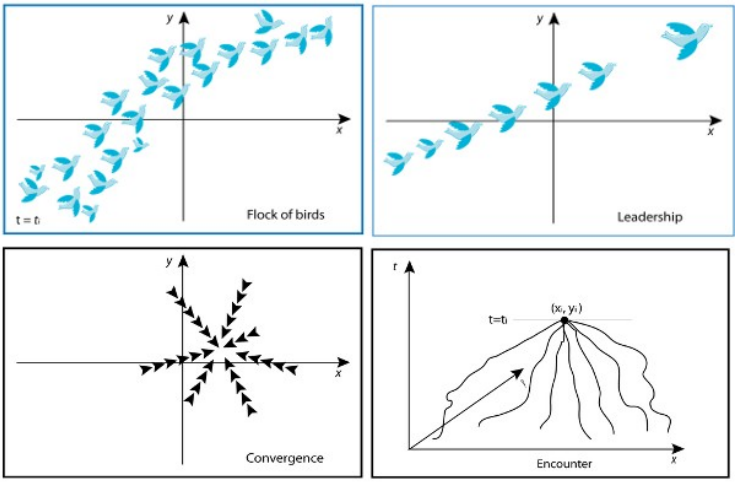
Uma trajetória é a sequência dos pontos que caracteriza a localização espacial de um objeto em um intervalo de tempo (BRAZ e BOGORNY, 2012). A representação mais simples de uma trajetória, denominada de trajetória bruta, consiste em um conjunto de pontos representados pelas seguintes informações: tid, x, y, t. O atributo tid é o identificador da trajetória; os atributos x e y são, respectivamente, as coordenadas geográficas referentes a latitude e a longitude; e o atributo t refere-se ao instante de tempo no qual o ponto foi gerado. A figura 1 apresenta um exemplo de trajetória bruta.

A análise de trajetórias brutas e a extração padrões de comportamento tem sido um tema explorado em diversos estudos científicos. No entanto, essas pesquisas ainda são relativamente recentes, e, em geral, objetivam a descoberta de padrões através da similaridade de trajetórias ou regiões densas (ALVARES et al., 2007). Uma das pesquisas pioneiras relacionada a análise de trajetórias brutas, foi o trabalho de

Laube e Imfeld (2005), que definiu uma série de padrões comportamentais. No estudo, foram identificados 4 tipos de padrões de trajetórias, conhecidos como padrões geométricos, que foram denominados: *flock*, *leadership*, *convergence* e *encounter*.

No padrão *flock*, são identificados grupos de objetos que se movem na mesma direção, num determinado raio e período de tempo, e suas trajetórias estão próximas umas das outras (exemplo: manada de bois). O padrão *leadership* refere-se à grupos de objetos que se movem na mesma direção, num determinado raio e período de tempo. No entanto, um dos objetos "lidera"o grupo (localiza-se espacialmente mais à frente dos demais), caracterizando-o como um criador de tendências. No padrão *convergence* são identificados grupos de objetos que se movem em direção à um mesmo local, num determinado período de tempo (exemplo: moradores de São Paulo indo em direção à avenida paulista). Por fim, no padrão *encounter* são identificados grupos de objetos que se deslocam para o mesmo local, em um mesmo intervalo de tempo. No entanto, é determinado um raio no qual as trajetórias devem permanecer juntas durante um período (exemplo: moradores de Florianópolis indo em direção ao mercado público, e que permanecem nele por no mínimo 30 minutos). A figura 3 apresenta um exemplo de cada um dos padrões anteriormente citados.

Figura 3: Exemplo de padrões em trajetórias



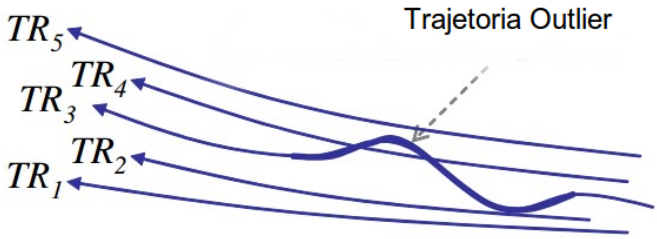
No método proposto em Carboni (2014), os dados brutos de trajetórias são utilizados para a identificação de comportamentos anômalos em trajetórias de motoristas. Tais comportamentos são identificados pelo uso de cálculos matemáticos que detectam acelerações, frenagens e mudanças bruscas de direção no decorrer da trajetória. Além disso, o método utiliza as informações detectadas para fazer a classificação dos motoristas em 4 categorias distintas: *Careful Driver*, motoristas que não realizaram nenhum comportamento anômalo durante a trajetória, *Distracted Driver*, motoristas que fizeram mudanças bruscas de movimento quando da ocorrência de algum movimento brusco na trajetória, *Dangerous Driver*, motoristas que fizeram mudanças bruscas de movimento sem motivo aparente e, por fim, *Very Dangerous Driver*, motoristas que fizeram mudanças bruscas de velocidade como acelerações acima da velocidade média da via ou frenagens repentinas, podendo também envolver mudanças bruscas de movimento. A figura 4 mostra um exemplo de aplicação do método de Carboni (2014). Nela, a imagem (a) mostra a trajetória analisada e a imagem (b) apresenta a marcação dos pontos desta trajetória onde foram identificados comportamentos anômalos.

Figura 4: Análise de comportamento anômalo em trajetórias de motoristas, adaptado de Carboni (2014)

Outro exemplo de padrão que pode ser extraído de trajetórias são os *outliers*. Um *outlier* em trajetórias é um objeto que se move de forma diferente em relação à maioria dos objetos que realizam trajetória semelhante. Um dos primeiros trabalhos de análise de *outliers* em trajetórias foi o algoritmo TRAOD (*Trajectory Outlier Detection Algorithm*) proposto por Lee et al. em 2008. Neste trabalho, as trajetórias são primeiramente divididas em sub-trajetórias (*partitioning*

phase), as quais são então comparadas (*detection phase*), através de uma medida de distância e direção, no intuito de encontrar segmentos onde a trajetória caracteriza-se como *outlier*. A figura 5 apresenta um exemplo de sub-trajetória *outlier*. Nela são apresentadas 5 trajetórias denominadas, respectivamente, de TR_1, TR_2, TR_3, TR_4 e TR_5. O segmento em destaque da trajetória TR_3 indica o momento no qual foi realizado um movimento *outlier* em relação à distância e direção dela para com as demais trajetórias.

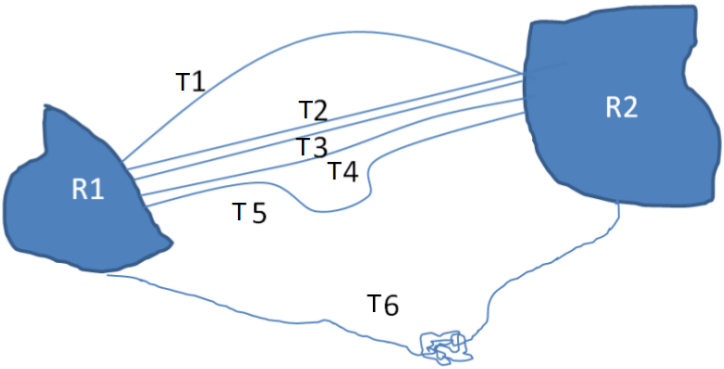
Figura 5: Identificação de Trajetórias Outliers, adaptado de Lee et al. (2008)



No trabalho de Fontes e Bogorny (2013), é proposto um outro algoritmo para detectar trajetórias *outliers*. Este difere-se do algoritmo de Lee et al. (2008) pelo fato de definir regiões de interesse entre as quais as trajetórias são analisadas, o que permite a detecção de trajetórias *outliers* em relação à um caminho padrão entre essas regiões. A figura 6 apresenta um exemplo. Nela são apresentadas 6 trajetórias denominadas T1, T2, T3, T4, T5 e T6. Supondo que as regiões R1 e R2 sejam um hotel e um restaurante respectivamente. Seria possível utilizar o algoritmo de Fontes e Bogorny (2013) para identificar que nas trajetórias T2, T3 e T4 foi seguida uma rota padrão entre os dois locais, ao passo que nas trajetórias T1, T5 e T6 foi seguida uma rota alternativa.

Uma das grandes dificuldades encontradas na análise de dados brutos de trajetórias está no fato de que este tipo de dado não dispõe de informações referentes ao contexto no qual a trajetória foi gerada. Sendo assim, é possível identificar comportamentos nessas trajetórias, mas é dificultoso elecar possíveis motivos para o mesmo. Por exemplo, no trabalho de Carboni (2014), não se tem a informação do tipo de evento que teria influenciado em um comportamento anômalo por

Figura 6: Exemplo de outliers, adaptado de Fontes et al. (2013)



parte do motorista, que pode ser um congestionamento, uma obra na pista, um alagamento, entre outros. Ainda, utilizando-se o algoritmo de Fontes (2013) para identificação de *outliers*, só é possível identificar que houve um desvio em determinado momento, não sendo este capaz de elencar um provável motivo para o mesmo.

Um dos primeiros esforços para solucionar o problema de identificação de motivos que tenham influenciado comportamentos em trajetórias foi realizado em Aquino (2014), que estende o trabalho de Fontes (2013), propondo uma solução capaz de identificar motivos de desvios em trajetórias em três situações: quando se tem a intenção de ir a um lugar fora do caminho padrão antes de se chegar ao destino (*Stop Outlier*), quando o desvio é realizado para se evitar eventos no caminho padrão (*Event Avoiding Outlier*) e quando o desvio é realizado para evitar uma situação de congestionamento (*Traffic Avoiding Outlier*). Apesar do método de Aquino (2014) fazer a identificação de desvios relacionados à eventos, não se tem a informação do tipo de evento identificado. Além disso, são necessários um número mínimo de trajetórias para identificar se o desvio foi realizado no intuito de evitar congestionamentos, visto que este método realiza cálculos matemáticos sobre dados de trajetórias brutas para identificar esta informação.

Além dos trabalhos descritos, existem uma série de outros que abordam soluções para a análise de dados de trajetórias, e que também são desprovidos de informações de contexto que permitam uma melhora das análises realizadas, pois dispõem somente de dados brutos de trajetórias como fonte de informação.

2.2 REDE SOCIAL TWITTER E CONTAS DE USUÁRIO FOCADAS NA DIVULGAÇÃO DE INFORMAÇÕES SOBRE CONDIÇÕES DE TRÂNSITO

As redes sociais são grupos de pessoas com as quais um indivíduo mantém contato ou alguma forma de vínculo social (BOWLING, 1997). Elas existem desde muito antes da criação da internet, no entanto, com a evolução desta tecnologia, elas passaram a ocorrer também através de websites, sendo conhecidas como redes sociais online. Nesta seção é apresentada a rede social Twitter e a contas de usuário do Twitter destinadas a divulgação de informações de trânsito, que foram utilizadas como fonte de informação para o desenvolvimento deste trabalho.

2.2.1 O Twitter

Dentre as principais Redes Sociais utilizadas na atualidade, a rede social Twitter destaca-se pelo número significativo de usuários. Ela foi criada em 2006 e está organizada na forma de um microblog que permite aos seus usuários a publicação de textos breves (máximo de 140 caracteres), que poderão ser visualizados publicamente, ou apenas por um grupo restrito de seguidores, conforme o usuário achar mais conveniente.

No intuito de estudar a dimensão do Twitter em escala mundial, foi realizada uma pesquisa em 2011, pela agência brasileira Monkey Business, onde foi verificado que cerca de 1 bilhão de mensagens são postadas por semana pelos usuários dessa rede. No inicio de 2012 haviam cerca de 383 milhões de contas registradas, sendo que o Brazil já ocupava o segundo lugar no ranking dos países com maior número de usuários, aproximadamente 33.3 milhões.

O Twitter pode ser comparado com um diário, onde um usuário que tenha uma conta tem a possibilidade de compartilhar informações que vão desde um texto puro até links para websites, notícias, imagens, vídeos ou outras mídias que ele achar interessante. Também é possível adicionar palavras precedidas de um símbolo (#), chamado de hashtag, que geralmente é utilizado para filtrar ou promover conteúdo.

As mensagens compartilhadas por um usuário recebem o nome de *tweet*, sendo que existe também a possibilidade de um usuário compartilhar *tweets* postados por outros usuários, o que é denominado *retweet*. Cada usuário da rede possui uma página de perfil, que contém informações pessoais (nome, fotografia, cidade de residência, etc.), al-

gumas estatísticas (número de seguidores/seguidos), e a sua *timeline*, uma lista dos *tweets* que foram postados por ele. Os *tweets* publicados sem restrições de vizualização podem ser lidos por qualquer indivíduo que acessar a página de perfil do usuário.

2.2.2 Contas Do Twitter Focadas Na Divulgação De Informações Sobre Condições De Trânsito

Com a popularização do Twitter, também se popularizaram contas de usuário utilizadas para publicação de *tweets* relacionados a um tema específico. Neste contexto, estão as contas utilizadas para publicação de informações sobre as condições de trânsito em uma cidade. A tabela 1 apresenta exemplos de *tweets* publicados por estas contas de usuário. A primeira coluna mostra a conta de usuário, e a segunda coluna um exemplo de *tweet* publicado por ela.

Tabela 1: Exemplo de tweet sobre condições de trânsito

Conta do Twitter	Exemplo de tweet publicado
O dia 24 horas	#ZonaNorte Trânsito intenso na Av.Maracanã, sentido #Centro, na alturado estádio.
Sul América Trânsito	Marginal do Tietê: motorista gasta, emmédia, 27 min na direção do Cebolão, entre Ayrton Senna e Anhanguera.
Trânsito Zero Hora	BR116 parada no sentido capital em Esteio.
Trânsito 24 horas	Beira-Mar Norte: Com trânsito lento no sentido Centro desde o CIC.

2.3 EXTRAÇÃO DE INFORMAÇÃO EM DADOS DO TWITER

As redes sociais são uma importante plataforma para a disseminação de informações, visto que permitem que seus usuários compartilhem conteúdo entre si de forma quase instantânea. O Twitter possui uma grande quantidade de usuários, e, por consequência, muitos dados compartilhados. Além disso, ele disponibiliza uma API (*Application Program Interface*) que permite a recuperação de parte desses dados, possibilitando sua utilização em diferentes análises.

Uma das formas de analisar os dados do Twitter é através da extração de informações presentes nos textos dos *tweets*. Um exemplo é o trabalho de Aron Culotta (2010), no qual foi realizado um estudo

de dados do Twitter para detectar surtos de gripe nos Estados Unidos pela análise de cerca de 570 milhões de tweets coletados durante os meses de setembro de 2009 a maio de 2010. No método proposto, foram utilizadas técnicas simples de *string matching* entre o *tweet* e palavras-chave relacionadas à gripe. Após esta etapa, foi realizada a contagem de *tweets* para os quais houve *matching*, e então foi calculada a proporção deles em relação ao total de *tweets* coletados. Como resultado da pesquisa foi possível detectar uma forte correlação entre a proporção de *tweets* para os quais houve *matching*, e as estatísticas semanais do *U.S. Centers for Disease Control and Prevention (CDC)* acerca dos surtos de gripe nos Estados Unidos. Por exemplo, utilizando apenas a palavra-chave *flu* (gripe em Inglês), foi possível detectar uma correlação de 81% entre a proporção de *tweets* e as estatísticas do CDC.

Além do texto do *tweet*, a capacidade de agregar dados geográficos nos *tweets*, disponibilizada a partir de 2010, tornou possível saber quais as coordenadas geográficas de onde um usuário compartilhou um texto, o que possibilitou pesquisas utilizando também este tipo de dado. Por exemplo, na pesquisa de Mislove et al. (2010), este tipo de dado é utilizado para identificação do nível de humor em diferentes regiões dos Estados Unidos. Para tal, são utilizadas palavras-chave para classificação do *tweet* entre os diferentes níveis de humor, e o dado georeferenciado para identificação da região do qual ele foi postado. Como resultado deste método obteve-se a caracterização do estado emocional das pessoas em diferentes regiões dos Estados Unidos.

Num primeiro momento, a extração de informações de trânsito oriundas de *tweets* geoferenciados caracterizou-se como uma opção com potencial interessante para o desenvolvimento deste trabalho, uma vez que, em teoria, seria possível buscar *tweets* de uma região de interesse utilizando coordenadas geográficas e então utilizar um conjunto de palavras-chave para extrair informações referentes à trânsito deles. Entretanto, experimentos realizados mostraram que dificilmente os usuários desta rede social postam este tipo de informação. Um outro problema relacionado à esta abordagem refere-se ao grau de confiabilidade das informações extraídas, isto porque a opinião de indivíduos acerca do que representa uma situação de congestionamento pode variar. Além disso, existe o problema da baixa qualidade dos textos publicados, pois no geral eles são repletos de gírias, abreviações e erros de grafia, o que dificulta o processo de extração de informações.

Em contrapartida, o uso de informações publicadas em contas do Twitter dedicadas exclusivamente à publicação de informações de trânsito mostrou-se mais adequada ao contexto deste trabalho, pois es-

tas informações são publicadas por entidades confiáveis (ex: entidades policiais), garantindo assim a veracidade das informações obtidas, e também possuem um menor grau de erros de grafia, pois são publicados com fins jornalísticos. Neste caso, o próprio texto do *tweet* contém a identificação da *região* sobre a qual a condição de trânsito está sendo informada, eliminando a necessidade de utilização de dados geoferenciados.

2.3.1 Reconhecimento de Entidades Nomeadas em dados do Twitter

O termo Entidades Nomeadas, no contexto de extração de informação em textos (IE), refere-se à unidades de informação que identificam nomes de pessoas, organizações, locais, informações temporais e quantitativas (NADEAU; SEKINE, 2007). A identificação de menções à Entidades Nomeadas, também conhecida como reconhecimento de entidades nomeadas (NER), é a classificação de palavras do texto como sendo referente à uma classe de entidade (ex: localidade) e caracteriza-se como uma etapa importante em alguns métodos de extração de informação em textos.

Utilizando-se NER, por exemplo, é possível fazer a identificação formal de conceitos para anotações semânticas em textos, permitindo distinguir o contexto de uso de uma palavra (ex: São Paulo referindo-se à cidade ou ao time de futebol). Outro exemplo são métodos que fazem a mineração de opnião em textos publicados na Web, onde utilza-se NER para classificar textos que abordam opiniões referentes à um mesmo contexto (ex: textos expressando opiniões políticas) (MARRERO et. al., 2013).

A utilização de técnicas de NER em textos publicados no Twitter enfrenta muitos desafios, visto que a maioria das técnicas existentes foram desenvolvidas para a identificação de entidades nomeadas em textos longos (ex: textos jornalísticos, obras literárias), diferentemente dos textos publicados no Twitter, que podem ter no máximo 140 caracteres. Em um estudo publicado por Derczynski et al. (2014) foi identificado que técnicas de NER apresentam entre 85 e 90% de precisão quando aplicadas à textos longos, dimunuindo significativamente, entre 30 e 50%, quando utilizadas em textos de *tweets*. Um dos problemas relacionados à esta redução é o tamanho limitado de textos de *tweets*, que dispõem de pouca informação relacionada ao contexto das informações. Por exemplo, no *tweet*: "São Paulo minha paixão", não

há informação suficiente para aferir com segurança se a palavra "São Paulo"refere-se ao time, ao estado, ao santo ou à cidade.

Em Sorato et al. (2016) foi realizada uma avaliação de ferramentas existentes para o reconhecimento de palavras relevantes (RWR) em textos de microblogs, que caracteriza-se por ser o processo de identificação e classificação de entidades nomeadas que tenham relevante papel sintático e/ou semântico em um texto (DOWNEY et al., 2007). Na pesquisa de Sorato et al. (2016) foram utilizadas para avaliação ferramentas de RWR em *tweets* escritos na lingua portuguesa e foram identificados uma série de problemas, como a não disponibilização de documentação que permita estudar o processo de RWR utilizado, a disponibilização apenas de versões online (demo) para um número limitado de *tweets*, a escassez de recursos apropriados para processar textos em língua portuguesa, entre outros.

O uso de ferramentas de NER foi desconsiderado para o processo de extração de informação proposto neste trabalho devido aos problemas citados e também porque elas apresentaram dificuldade em identificar nos *tweets* de trânsito entidades úteis no contexto deste trabalho como, por exemplo, entidades que referem-se à localidade. A tabela 2 apresenta exemplos de reconhecimento de entidades nomeadas em *tweets* contendo informações de trânsito realizado pelas ferramentas *LX-NER*[1] e a *OpenNLP*[2]. A escolha dessas ferramentas para testes com os *tweets* de trânsito foi devido ao fato de elas suportarem NER para a lingua portuguesa e estarem disponíveis para uso em pesquisa (SORATO et al., 2016). Na tabela, a primeira coluna mostra o *tweet* avaliado e a segunda e a terceira coluna mostram as saídas das ferramentas. Para cada saída são apresentadas as palavras que fazem menções à entidades e, entre colchetes, a classe da entidade identificada. Nota-se que no primeiro *tweet* ambas as ferramentas reconheceram "Norte da Ilha"como sendo uma menção pertencente à classe localidade, no entando, a localidade de interesse no contexto deste trabalho seria "SC-401". No segundo *tweet* as palavras "Av", "Centro"e "Norte"foram também identificadas como menções pertencentes à classe localidade, sendo que a localidade de interesse seria "Beira-Mar". A ferramenta *OpenNLP* foi também utilizada para comparação com um dos algoritmos desenvolvidos neste trabalho que é apresentada no capítulo de experimentos.

[1] http://lxcenter.di.fc.ul.pt/services/en/LXServicesNer.html
[2] https://opennlp.apache.org/

Tabela 2: Exemplo de saídas de ferramentas de NER para *tweets* contendo informações de trânsito

Texto do tweet	OpenNLP	LX-NER
SC-401: movimento é intenso sentido Norte da Ilha, com trechos lentos. #t24horas	Norte Ilha [place]	Norte da Ilha [location]
Beira-Mar Norte: trânsito congestionado sentido Centro desde a saída da Av. da Saudade. # t24horas	Av [place] Norte [place]	Norte [location] Centro [location] Av [location] Saudade [works]

2.4 TÉCNICAS DE EXTRAÇÃO DE INFORMAÇÃO EM TEXTOS

O processo de Recuperação de Informação em textos (RI), consiste em uma série de técnicas para fazer uma redução dimensional do texto a ser analisado, bem como, identificar similaridades de palavras ou expressões, utilizando-se um conjunto de palavras ou expressões de interesse previamente definidos (EBECKEN et al., 2003). Dentre as diferentes técnicas existentes, são utilizadas neste trabalho as técnicas de tokenização, busca de termos relevantes por similaridade de palavras, e padronização de *strings*. Cada uma delas é devidamente detalhada nas próximas subseções.

2.4.1 Tokenização

A técnica de tokenização é a primeira etapa na extração de informações em textos, pois é ela que irá permitir a análise das palavras do texto de forma individual. Segundo Carvalho (2012), a tokenização consiste na separação de um texto em unidades básicas denominadas tokens. Um token pode ser uma palavra, ou algum outro elemento como um número ou sinal de pontuação. Para realizar esta separação em tokens, o principal critério utilizado é o uso de um espaço em branco, mas podem também ser utilizados critérios baseados em tabulação ou no início de uma nova linha.

Para exemplificar o processo de tokenização, suponha a seguinte expressão: “Beira-mar norte: trânsito lento, houve um acidente com moto.”. Considere ainda que o processo de tokenização utiliza como critério de separação de palavras os espaços em branco entre elas. Após este processo a frase ficaria representada da seguinte forma: [Beira-mar]

[norte:] [trânsito] [lento,] [houve] [um] [acidente] [com] [moto].

Um problema comum durante o processo de tokenização está relacionado aos sinais de pontuação. Retomando o exemplo anterior, a palavra "lento"juntamente com um sinal de ",", foram consideradas como um token único, o que poderia não ser interessante dependendo do tipo de análise pretendida. Uma das soluções para este problema, poderia ser a remoção do sinal de pontuação. No entanto, isto também seria dependente do tipo de análise, visto que a remoção de um sinal de "."poderia comprometer a identificação de uma abreviação por exemplo.

2.4.2 Similaridade de Palavras

Nos Sistemas de Recuperação de Informações Textutais (SRI Textual), é bastante comum a definição de um conjunto de palavras-chave referentes a um determinado domínio que são utilizadas para identificação dos textos mais relevantes. Esta identificação é realizada através da quantidade de palavras semelhantes que cada texto possui com um conjunto de palavras-chave (WIVES, 2002). A forma utilizada para verificar esta semelhança é através do uso de uma Função de Similaridade.

A Função de Similaridade é responsável por encontrar relações entre termos chave e um texto. A forma mais simples de implementação é realizando uma comparação direta entre os termos chave e as palavras do texto, utilizando o critério de igualdade, onde duas palavras são consideradas semelhantes, se e somente se elas possuem a mesma sequência de caracteres. No entanto, devido a problemas de sinonímia, polissemia e ainda outros relacionados a linguagem, são necessários métodos mais sofisticados para realizar esta comparação (MORAIS, 2007). Neste trabalho foi realizada uma análise de duas técnicas dentre as existentes na literatura: o algoritmo de Levenshtein (LEVENSHTEIN, 2007) e o algoritmo de Jaro-Winkler (PORTER e WINKLER, 1997). No trabalho de Pires (2011), elas foram listadas como as mais citadas em trabalhos científicos, o que serviu de motivação para o uso delas.

- Função de Levenshtein: A função de Levenshtein tem como objetivo calcular a similaridade através da distância entre duas *strings*. Para isso ela calcula o mínimo de operações necessárias para transformar uma *string* em outra. As transformações podem ser entendidas como sendo: substituição, remoção e inserção, cada uma com custo 1. Por exemplo, considerando as palavras

"Forte"e "Fonte", a distância calculada seria 1, visto que para a conversão da primeira na segunda seria necessária a substituição do caracter "r"pelo caracter "n".

- Função de Jaro-Winkler: A função de JaroWinkler (WINKLER, 1990) é uma variação da função de Jaro (JARO, 1989, 1995), e tem por objetivo medir a similaridade entre duas strings, levando-se em consideração o tamanho delas. Como resultado, ela retorna um valor entre 0 e 1, onde quanto mais próximo de 1 for este valor, mais similares as duas strings são. Para que haja um melhor entendimento desta medida, será primeiramente explicado o algoritmo de Jaro, e posteriormente o algoritmo Jaro-Winker.

  Matematicamente, o algoritmo de Jaro é dado pela fórmula:

$$d_j = \frac{1}{3}\left(\frac{c}{d} + \frac{c}{r} + \frac{c-\tau}{c}\right) \tag{2.1}$$

  Onde:

d_j : função de Jaro para comparação entre strings;

c: número de caracteres em comum entre as duas strings;

d: número de caracteres da primeira string;

r: número de caracteres da segunda string;

τ: número de transposições de caracteres. Estas transposições são caracteres que estão localizados em posições diferentes nas duas strings em comparação.

O cálculo de Winkler (1994), utilizou a função de Jaro, no entanto adicionou um cálculo para levar em consideração se os quatro primeiros caracteres das duas strings são iguais. Sendo assim, ele é definido pela fórmula:

$$d_w = d_j + ik\,(1 - d_j) \tag{2.2}$$

Onde:

d_w : função de Winkler;

i: quantidade de caracteres iguais nas duas strings até a quarta posição;

k: constante utilizada por Winkler para definir o quanto a similaridade será ajustada para cima, considerando a quantidade de caracteres em comum até a quarta posição. No trabalho de Winkler foi utilizada a constante k = 0.1. Sendo que, ele também afirmou que ela não deve ser maior que 0.25, pois poderá resultar em um cálculo de similaridade maior que 1.

d_j: função de Jaro.

Em Pires (2011), foi realizado um experimento para comparar as funções de Levenshtein e Jaro-Winkler. As funções foram utilizadas para comparar strings referentes à nomes de pessoas, variando-se o posicionamento de alguns caracteres, e até mesmo adicionando alguns. A figura 7 apresenta uma tabela com estes resultados. Note que as primeiras colunas (mais à esquerda) mostram as duas strings que foram comparadas, e as duas últimas colunas (mais à direita) mostram o percentual de semelhança calculado respectivamente pela função de Levenshtein e de Jaro-Winker.

Figura 7: Comparação de strings através dos algoritmos de Levenshtein e Jaro-Winkler. Fonte: Pires, Fábio A. 2011, p.8

		% Semelhança	
String1	**String2**	**Levenshtein**	**Jaro-Winkler**
LUIZA SOUSA	LUISA SOUZA	82	88
DEISE BARBOSA	DEIZE BARBOZA	85	92
CASSEMIRO BITENCOURT	CASEMIRO BITENCORT	90	97
THEREZINHA CRIVELLI	TEREZINHA CRIVELI	90	91
VICTOR MAGAWA	VITOR MAGAVA	85	93
JACYRA LOCHIDIO	JACIRA LOXIDIO	15	53
JOSÉ JOAQUIM XAVIER	CHAVIER, JOZÉ JOAQUIM	20	65

Analisando-se os resultados dessa tabela, é possivel verificar que a técnica de Jaro-Winker é menos sensível a diferenças que a técnica de Levenshtein. Por exemplo, as duas strings da última linha da tabela são bastante similares, sendo que a diferença entre elas são os caracteres iniciais e o posicionamento das palavras XAVIER e CHAVIER. No entanto, a função de Levenshtein calculou uma porcentagem de apenas 20% de semelhança, ao passo que a de Jaro-Winkler calculou 65% de semelhança. Devido a essa diferença a função de Jaro-Winker foi considerada mais adequada para o desenvolvimento deste trabalho. Na

seção 2.2.3 será apresentada a técnica de padronização de strings, cujo objetivo é melhorar o cálculo de similaridade entre palavras.

2.4.3 Padronização de strings

Um dos principais problemas em processos de comparação de palavras são as possíveis formas de grafia, e isto foi um dos motivos que levou pesquisadores a desenvolver as funções de cálculo de similaridade entre palavras. No entanto, essas funções podem apresentar resultados insatisfatórios, ou seja, apresentar um cálculo de similaridade baixo para strings que aos olhos humanos são bastante similares.

No intuito de melhorar o cálculo realizado pela função de similaridade, alguns trabalhos utilizam técnicas para fazer uma padronização entre as strings sendo comparadas. Por exemplo, no trabalho de Pires (2011) foi utilizado um método de fonetização de palavras, que tem como objetivo substituir a forma escrita pela forma de fonemas (forma como a palavra é pronunciada) e com isto minimizar erros de grafia. Na pesquisa foi feita uma comparação do resultado da aplicação das funções de Jaro-Winker e Levenshtein antes e depois desse processo, sendo que, foi verificado que o cálculo de similaridade para as strings que passaram pelo método de fonetização foi mais satisfatório. A figura 8 apresenta os resultados obtidos com esta análise.

Figura 8: Comparação de strings através dos algoritmos de Levenshtein e Jaro-Winkler incluindo registros fonetizados Fonte: Pires, Fábio A. 2011, p.86

			% Semelhança	
Tipo	**String1**	**String2**	**Levenshtein**	**Jaro-Winkler**
Normal	LUIZA SOUSA	LUISA SOUZA	82	88
Fonética	LUIZA SUZA	LUIZA SUZA	100	100
Normal	DEISE BARBOSA	DEIZE BARBOZA	85	92
Fonética	DIZI BARBUZA	DIZI BARBUZA	100	100
Normal	CASSEMIRO BITENCOURT	CASEMIRO BITENCORT	90	97
Fonética	KASIMIRU BITINKURTI	KAZIMIRU BITINKURTI	95	97
Normal	THEREZINHA CRIVELLI	TEREZINHA CRIVELI	90	91
Fonética	TIRIZINIA KRIVILI	TIRIZINIA KRIVILI	100	100
Normal	VICTOR MAGAWA	VITOR MAGAVA	85	93
Fonética	UITUR MAGAVA	UITUR MAGAVA	100	100
Normal	JACYRA LOCHIDIO	JACIRA LOXIDIO	15	53
Fonética	GIASIRA LUXIDIU	GIASIRA LUXIDIU	100	100
Normal	JOSÉ JOAQUIM XAVIER	CHAVIER, JOZÉ JOAQUIM	20	65
Fonética	GIUZI GIUAKIN XAVIR	XAVIR GIUZI GIUAKIN	37	76

No presente trabalho, após análises exploratórias das palavras presentes nos textos de *tweets* relacionadas à trânsito, decidiu-se utilizar um processo de *padronização de strings* diferente do utilizado no trabalho de Pires, que é detalhado no capítulo 3.

3 PROPOSTA PARA EXTRAÇÃO DE INFORMAÇÕES DE TRÂNSITO EM TEXTOS PUBLICADOS NO TWITTER

Este capítulo apresenta o método proposto para recuperação e armazenamento de informações de trânsito em textos de *tweets* a fim de permitir seu uso como informação de contexto em interpretações de trajetórias de veículos. A elaboração do método foi realizada em três etapas e na seguinte ordem:

- Identificação das informações presentes nos *tweets* necessárias para o uso com trajetórias de veículos (informações de interesse);
- Criação de um método que faça a extração das informações de interesse identificadas;
- Elaboração de uma estrutura que permita o armazenamento e recuperação das informações de interesse.

A seção seção 3.1 apresenta as informações de interesse, a seção 3.2 apresenta o método desenvolvido para recuperação dessas informações, e a seção 3.3 apresenta a estrutura escolhida para o armazenamento das mesmas.

3.1 IDENTIFICAÇÃO DAS INFORMAÇÕES DE INTERESSE NOS TEXTOS DOS *TWEETS*

A primeira etapa no processo de desenvolvimento do método proposto neste trabalho foi a realização de um levantamento das informações presentes nos *tweets* de trânsito que precisam ser extraídas para a utilização com trajetórias de veículos. São elas:

- *Data*: Data de publicação do *tweet*. Ela serve para identificar a data da ocorrência de uma situação de trânsito, visto que esses *tweets* são utilizados para a publicação de condições de trânsito em tempo real. Esta informação permite recuperar informações de trânsito publicadas na mesma data de coleta das trajetórias;
- *Lugar*: Lugar do qual a informação de trânsito está sendo informada. Esta informação serve para identificar as condições de trânsito referentes ao lugar no qual trajetórias foram coletadas (ex: Av. Beira-mar, rua Gustavo Richard, Ponte Colombo Salles, etc);

- *Situação de Trânsito*: Informação da presença ou não de situações de congestionamento (ex: fila, retenção, trânsito livre, etc);

- *Evento de Trânsito*: Informação da presença ou não de eventos que possam estar influenciando no trânsito (ex: alagamentos, manifestações, colisões de veículos, etc).

3.2 MÉTODO PARA RECUPERAÇÃO DE INFORMAÇÕES DE TRÂNSITO EM TEXTOS DE TWEETS

Para implementação do método foi primeiramente realizada uma análise exploratória de um conjunto de textos de *tweets* de trânsito no intuito de identificar a maneira como as informações de interesse são publicadas para então definir uma metodologia que permita sua extração. A tabela 3 apresenta exemplos dessas informações. Nela, a primeira coluna apresenta o texto do *tweet*, e as três colunas subsequentes apresentam, respectivamente, o lugar, a situação de trânsito e o evento identificados.

Tabela 3: Exemplo de tweets e as informações de interesse

Texto do tweet	Lugar	Situação	Evento
Br 101 sentido sul km 225 ao 227 **obras** deixam o trânsito em meia pista no momento **sem fila**.	BR-101	Sem Fila	Obras
SC-401: devido ao **acidente** em ambos os sentidos trânsito apresenta bastante **lentidão**.	SC-401	Lentidão	Acidente
Manifestação segue pela Avenida **Paulo Fontes** causando **retenção**.	Paulo Fontes	Retenção	Manifestação

Após a análise manual dos *tweets* pôde-se identificar alguns padrões como:

- A primeira letra de cada palavra que compõe o nome de um lugar está em maiúsculo na maioria dos *tweets*. (ex. “João Pio Duarte Silva”, “BR-101”, “Av. Gustavo Richard”);

- As situações de trânsito e os eventos podem ser identificados por uma palavra (ex: “congestionamento”, “colisão”), mas também podem ser identificados por expressões compostas de duas palavras (ex: “trânsito lento”, “fluxo intenso”, “carro quebrado”, “rodovia interditada”);

- A maioria dos *tweets*, quando apresentam a negação de uma situação de trânsito, contém esta informação antes da situação. (ex: "sem registro de fila", "não há retenção"), sendo as palavras mais utilizadas "não"e "sem".

A partir das análises realizadas, foi realizada a implementação de um algoritmo para a extração de cada uma das informações de interesse: um para extração do lugar, outro para extração da situação de trânsito e outro para extração de eventos que estejam influenciando no trânsito. Não foi necessário desenvolver algoritmo para extrair a data de publicação do *tweet*, visto que ela já é disponibilizada pelo próprio Twitter via API. O restante dessa seção está organizada da seguinte forma: A subseção 3.2.1 apresenta as técnicas de recuperação de informação utilizadas nos algoritmos propostos, a subseção 3.2.2 apresenta o algoritmo para extração de lugares, e a subseção 3.2.3 apresenta os algoritmos para extração da situação e eventos no trânsito.

3.2.1 Técnicas para recuperação de informação utilizadas nos algoritmos propostos

Os algoritmos propostos neste trabalho analisam as palavras que compõem os textos dos *tweets* de maneira individualizada e/ou combinadas em conjuntos de tamanhos distintos. Para isto, é realizado um processo de tokenização utilizando como critério de separação dos tokens os espaços em branco entre as palavras. Além disso, são descartados os tokens considerados inúteis para o processo de recuperação de informação realizado neste trabalho, que são: palavras seguidas dos sinais # e @, links, e o token RT. A tabela 4 mostra um exemplo do texto de um *tweet* antes e depois desse processo.

Tabela 4: Exemplo de tweet antes e depois de tokenização

Texto original do tweet	Texto após tokenização
"RT @PRF191SC SC-405: Trânsito segue lento no sentido bairros, entre o Elevado da Seta e o Trevo do Rio Tavares. #t24horas"	[SC-405:] [Trânsito] [segue] [lento] [no] [sentido] [bairros,] [entre] [o] [Elevado] [da] [Seta] [e] [o] [Trevo] [do] [Rio] [Tavares.]

Para realizar o reconhecimento das informações de interesse nos textos dos *tweets* é feita a comparação dos tokens e/ou conjunto de tokens com palavras-chave conhecidas, utilizando o cálculo de similaridade entre palavras de JaroWinkler. No intuito de minimizar erros na aplicação da função de similaridade, foi realizado um processo de

padronização de *strings* contendo as seguintes características:

i Antes do armazenamento das palavras-chave referentes a lugar, situação de trânsito e evento é realizado um mapeamento de caracteres para letras minúsculas, bem como a remoção de hífen, underline, sinais de pontuação, e de espaços em branco presentes em nomes compostos por mais de uma palavra. Este processo gera as *strings* que são armazenadas no atributo *matching* das tabelas *Lugar*, *Situacao* e *Evento*. A tabela 5 apresenta exemplos dessas palavras antes e depois deste processo de padronização.

ii Os algoritmos para recuperação de informação realizam um processo de padronização de *strings* nas palavras do texto do *tweet* a ser analisado que é similar ao apresentado anteriormente. No entanto, como as palavras do texto do *tweet* são primeiramente separadas em tokens, não é necessária a remoção de espaços em branco entre elas. A tabela 6 apresenta exemplos dessas palavras antes e depois do processo de padronização de *strings*.

Tabela 5: Exemplo de palavras-chave antes e depois do processo de padronização

Tabela	Antes da Padronização	Depois da Padronização
Lugar	João Pio Duarte Silva	joaopioduartesilva
	Gama Déça	gamadeca
	Av. Gustavo Richard	avgustavorichard
Situação	trânsito lento	transitolento
	Congestionamento	congestionamento
	retenção	retencao
	colisão	colisao
Evento	manifestação	manifestacao
	alagamento	alagamento

Tabela 6: Exemplo de palavras dos tweets antes e depois do processo de padronização

Palavra antes da padronização	Palavra depois da padronização
SC-405	sc405
Trânsito	transito
Capoeiras,	capoeiras
BR-101	br101
Déça	deca
Silva:	silva

3.2.2 Algoritmo para identificação de lugares em textos de tweets

O algoritmo de identificação de lugar é responsável por analisar o texto do *tweet*, buscando identificar o lugar ao qual ele faz referência. Como foi identificado na etapa de análise manual dos textos dos *tweets* que grande parte deles apresentam o nome do lugar em letra maiúscula, isto serviu como um dos critérios para a implementação deste algoritmo. Dessa maneira, foi estabelecido que somente são analisadas as palavras com letra maiúscula, ou que são números, pois estas últimas são comumente utilizadas para identificar rodovias (ex: "BR 101"). O fluxograma da figura 9 apresenta os passos envolvidos na execução do algoritmo, e o número de cada operação (retângulo) será usado para descrever o seu funcionamento. Ele recebe como entrada 3 parâmetros: o texto do *tweet* que será analisado, um conjunto de palavras-chave relacionadas a lugar utilizadas para o cálculo de similaridade com palavras do *tweet* e o valor mínimo do cálculo de similariade de Jaro-Winkler para considerar duas *strings* como sendo similares.

Primeiramente é realizado um processo de tokenização das palavras do texto do *tweet*. Essas palavras são então armazenadas em uma lista chamada de lista de palavras (1). Além disso, é criada uma outra lista, inicialmente vazia, onde serão colocadas palavras que podem identificar lugares, chamadas de candidatos a lugar (2). Uma vez que as duas listas são criadas, o algoritmo faz a leitura de cada uma das palavras da lista de palavras, identificando aquelas que satisfazem o critério de candidato a lugar (3-6). Caso hajam candidatos identificados, o algoritmo irá realizar o processo de padronização dessas palavras para que sejam posteriormente utilizadas no cálculo de similaridade

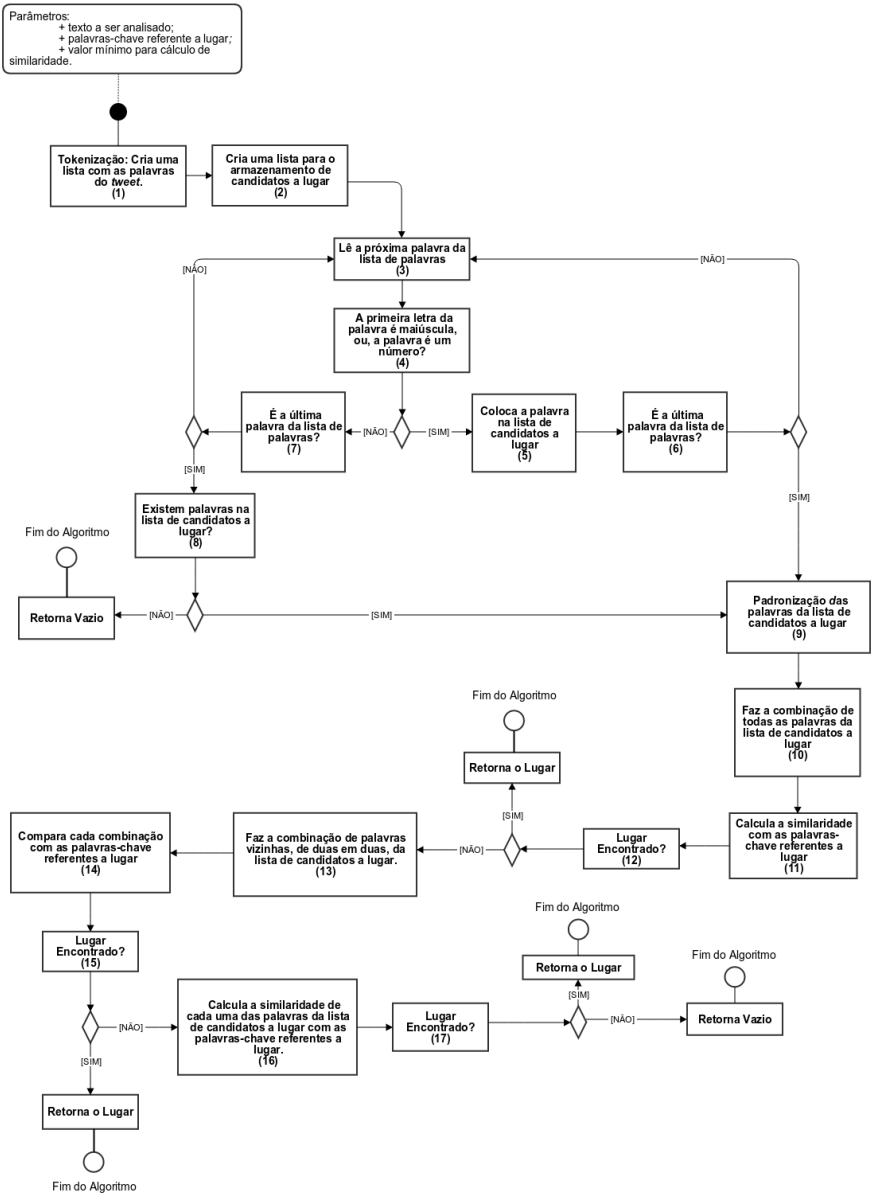
com as palavras-chave referentes a lugar recebidas como parâmetro. Caso não hajam candidatos, o algoritmo termina (7-9).

Uma vez selecionados e padronizados, os candidados a lugar são combinados em diferentes conjuntos para realização do cálculo de similaridade. A primeira combinação é formada pela concatenação de todos os candidatos a lugar. Ex: [‘av.’, ‘gustavo’, ‘richard’] → [‘avgustavorichard’]. Caso o cálculo de similaridade desta com alguma palavra do conjunto de palavras-chave relacionadas a lugar resulte em um valor maior ou igual àquele definido no parâmetro de valor mínimo do cálculo de similariade, então o algoritmo considera que o lugar foi encontrado e termina (10-12).

Caso a combinação anterior não identifique um lugar, então os candidatos a lugar são combinados de dois em dois, e somente entre palavras vizinhas. Ex: [‘br’, ‘101’, ‘centro’, ‘capital’] → [‘br101’, ‘101centro’, ‘centrocapital’]. Após o processo de combinação, as palavras resultantes são utilizadas para o cálculo de similaridade com cada uma das palavras-chave referentes a lugar. Caso haja uma comparação que resulte em um valor maior ou igual àquele definido no parâmetro de valor mínimo do cálculo de similariade, então o algoritmo considera que o lugar foi encontrado e termina (13-15).

Por fim, caso nenhuma das combinações anteriores identifique um lugar, os candidatos a lugar são utilizados um a um para o cálculo de similaridade com as palavras-chave referentes a Lugar. Caso alguma comparação resulte em um valor maior ou igual àquele definido no parâmetro de valor mínimo do cálculo de similariade, então o algoritmo considera que o lugar foi encontrado. Em caso negativo, o algoritmo não retornará nenhum lugar (16-17).

Figura 9: Fluxograma do algoritmo para identificação de lugar



3.2.3 Algoritmos para Indentificação de Situação de Trânsito e Eventos

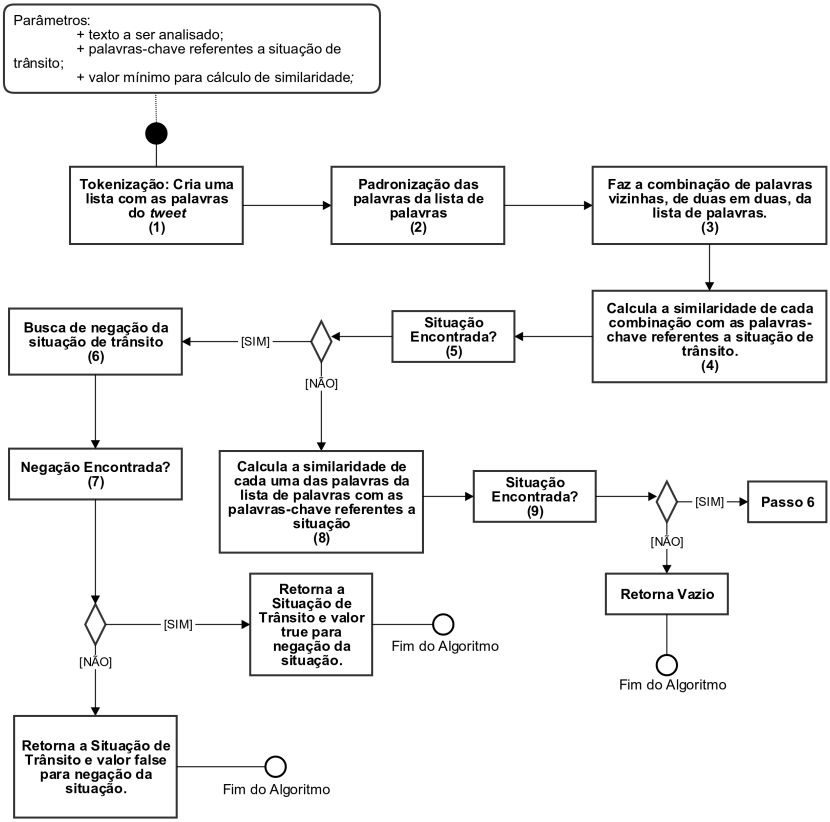
O algoritmo de identificação de situação do trânsito busca palavras que identificam o congestionamento ou não de vias de uma cidade. O fluxograma da figura 10 apresenta uma representação de alto nível dos passos do algoritmo, e o número de cada operação (retângulo) será usado para descrever o seu funcionamento. Ele recebe como entrada 3 parâmetros: o texto do *tweet* que será analisado, um conjunto de palavras-chave relacionadas a situações de trânsito utilizadas para o cálculo de similariadade com palavras do *tweet* e o valor mínimo do cálculo de similariade de Jaro-Winkler para considerar duas *strings* como sendo similares.

Assim como no algoritmo de identificação de lugares, é primeiramente realizado um processo de tokenização das palavras do texto do *tweet* que são então armazenados em uma lista (1). Após este processo, os tokens passam por uma padronização para que possam ser posteriormente utilizados no cálculo de similaridade com as palavras-chave referentes a situação de trânsito (2). Essas palavras são então combinadas de duas em duas e somente entre palavras vizinhas. Ex: [‘br’, ‘101’, ‘transito’, ‘sem’, ‘fila’, ‘no’, ‘momento’] → ‘br101’,‘101transito’, ‘transitosem’, ‘semfila’, ‘filano’, ‘nomomento’ (3). Optou-se por esta combinação pelo fato da maioria dos *tweets* apresentarem duas palavras para identificar uma situação de trânsito como em “trânsito lento”, “fluxo intenso”, “fluxo acentuado”. As palavras combinadas são então utilizadas para o cálculo de similaridade com cada uma das palavras-chave relacionadas à situação de trânsito (4).

Caso a comparação com alguma palavra-chave resulte em um valor maior ou igual àquele definido no parâmetro de valor mínimo para o cálculo de similaridade, o algoritmo considera que a situação de trânsito foi encontrada (5). Após isto, ele verifica se houve uma negação da situação de trânsito, procurando a ocorrência das palavras ‘sem’ e ‘não’ antes da situação encontrada (6-7). Ao final deste processo, o algoritmo retorna a situação encontrada e um valor booleano para a negação, sendo *true* em caso afirmativo e *false* caso contrário.

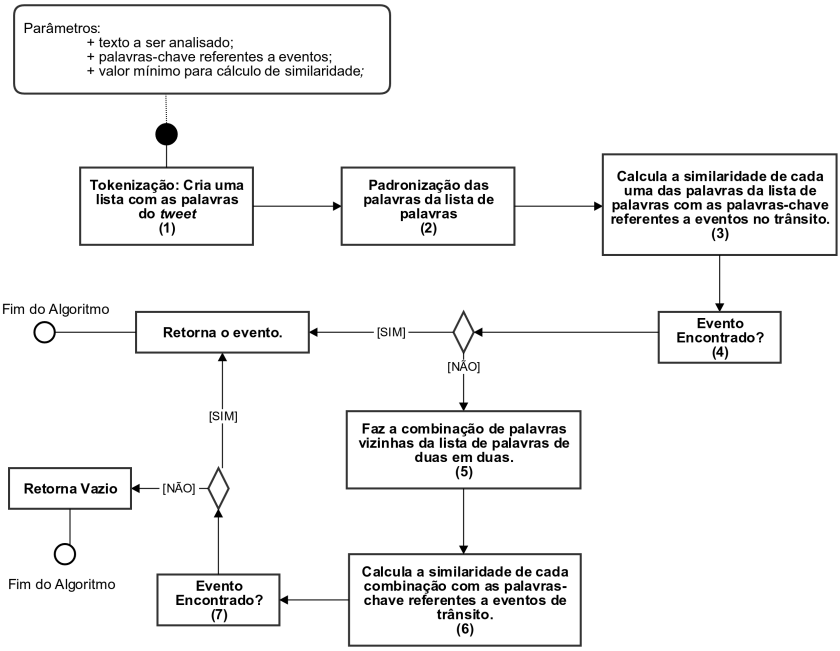
Se a combinação de palavras realizada anteriormente não resultar em uma situação identificada, as palavras são então utilizadas uma a uma para o cálculo de similaridade com as palavras-chave. Ex: [‘br’, ‘101’, ‘transito’, ‘sem’, ‘fila’] → ‘br’, ‘101’, ‘transito’, ‘sem’, ‘fila’. Caso uma situação seja encontrada, então o algoritmo irá realizar a busca da negação da situação (6-7), caso contrário, o algoritmo termina (8-9).

Figura 10: Fluxograma do algoritmo para identificação de situação de trânsito



O algoritmo de identificação de eventos é responsável pela identificação de eventos de trânsito em vias de uma cidade e é bastante similar ao algoritmo de identificação de situação de trânsito. Como exemplo de eventos de trânsito pode-se citar: 'tombamento', 'alagamento', 'colisão', 'acidente', etc. Este algoritmo recebe como parâmetros de entrada o *tweet* a ser analisado, uma lista de palavras-chave referentes a eventos de trânsito e o valor mínimo do cálculo de similariade de Jaro-Winkler para considerar duas *strings* como sendo similares. A diferença entre este algoritmo e o de identificação de situação esta no fato de que o cálculo de similaridade é primeiramente realizado utilizando cada um dos tokens separadamente, isto porque, no caso de eventos, é mais comun o uso de uma única palavra para identificá-lo. Caso nenhum seja encontrado no passo anterior, os tokens são então combinados dois a dois para uma nova comparação. Um outro aspecto que diferencia os dois algoritmos é o fato de que este não contém um passo para verificação de negação, já que esta, quando presente no *tweet*, refere-se à negação de uma situação de trânsito e não de um evento. O fluxograma da figura 11 apresenta uma representação de alto nível dos passos envolvidos na execução deste algoritmo.

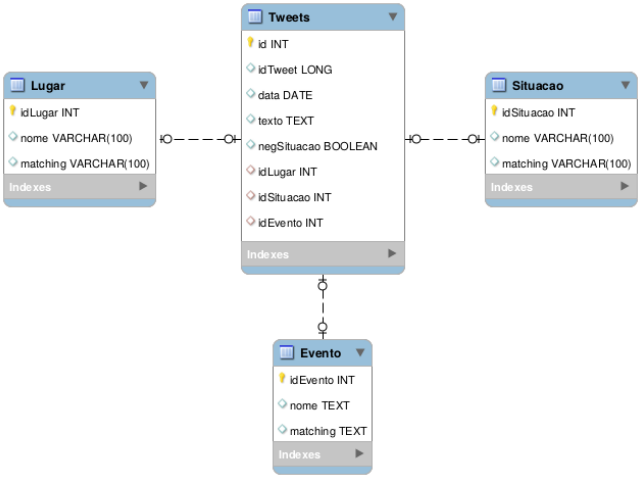
Figura 11: Fluxograma do algoritmo para identificação de eventos no trânsito



3.3 PROPOSTA PARA ARMAZENAMENTO DE INFORMAÇÕES RECUPERADAS DE TEXTOS DE TWEETS

Para o armazenamento das informações de interesse recuperadas dos textos dos *tweets* foi criado um banco de dados relacional. Esta decisão veio do fato deste tipo de estrutura oferecer um mecanismo formal para organização dos dados extraídos, e também permitir a consulta das informações armazenadas utilizando-se linguagem SQL (*Structured Query Language*), o que facilita o uso posterior das mesmas. A figura 12 apresenta o modelo lógico do banco de dados proposto.

Figura 12: Modelo lógico do banco de dados relacional proposto



No modelo proposto, a tabela *Tweets* funciona como elemento central para o relacionamento entre os dados recuperados de cada *tweet*. Sendo assim, ela é responsavel pelo relacionamento de um *tweet* com o lugar, e/ou a situação de trânsito e/ou o evento recuperados dele. Além disso, ela armazena a data de publicação do *tweet* permitindo recuperar, por exemplo, informações da condição de trânsito de um determinado lugar num dado momento. Nesta tabela, o atributo *negSituacao* é utilizado para marcar se houve uma negação da situação de trânsito identificada. Por exemplo, no texto: "não há indícios de congestionamento", a situação de trânsito identificada seria "congestionamento", entretanto, houve uma negação pelo uso da expressão "não há", o que caracteriza o fato de que o trânsito não esta congestionado. Neste caso,

o atributo *negSituacao* deverá ser salvo com o valor booleano *true*.

A tabela *Lugar* armazena um conjunto de palavras-chave referentes a nomes de vias (ruas, avenidas, rodovias, etc) situadas em uma região. A tabela *Situacao* armazena um conjunto de palavras-chave referentes à condições de trânsito como: "fila", "congestionado", "trânsito lento", etc. Por fim, a tabela *Evento* armazena um conjunto de palavras-chave relacionadas à eventos que podem influenciar no trânsito como: "acidente", "alagamento", "tombamento", etc. Essas palavras-chave são utilizadas como as palavras-chave de entrada nos algoritmos de extração de informação propostos. Nas três tabelas, o atributo *matching* contém uma versão mais simplificada das palavras-chave, conforme exemplificado na Tabela 5. Esse atributo é utilizado para operações de *string matching*.

4 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta os experimentos realizados para a avaliação do método de recuperação de informações em textos de *tweets* proposto neste trabalho, e também os testes realizados para analisar a viabilidade de uso das informações extraídas na interpretação de trajetórias. O restante deste capítulo esta dividido da seguinte forma: A seção 4.1 apresenta os experimentos realizados para avaliação dos algoritmos de extração de informações de trânsito em textos de *tweets* e a seção 4.2 apresenta a análise da viabilidade de uso dessas infromações para a interpretação de dados brutos de trajetórias.

4.1 EXPERIMENTOS COM DADOS DO TWITTER

O objetivo destes experimentos foi avaliar a eficácia dos algoritmos propostos em extrair corretamente as informações de interesse dos textos dos *tweets*: *Lugar*, *Situação de trânsito* e *Eventos* que influenciem no trânsito. Para realização dos experimentos foram coletados *tweets* da conta Trânsito 24 horas utilizando a *REST API* do Twitter. Esta API permite que se recupere *tweets* de uma determinada conta publicados em um intervalo de tempo. Dessa maneira, foi então implementado um *script* em linguagem *Ruby*[1] juntamente com a biblioteca *twitter*[2], escrita nesta mesma linguagem, que possibilita o uso da *REST API* do Twitter.

A coleta de *tweets* foi realizada utilizando-se a função da API do Twitter *user_timeline*, que retorna a coleção de *tweets* mais recentes postados em uma conta do Twitter. Após coletados, os *tweets* foram armazenados na tabela *Tweets* do banco de dados proposto neste trabalho (seção 3.3), que foi implementado utilizando o SGDB PostGreSQL[3].

Os experimentos utilizaram como valor mínimo do cálculo de similaridade uma taxa de 95%. Ela foi definida após análises exploratórias onde variou-se esta taxa para identificar o valor que resultasse em um maior número de informações extraídas corretamente. Foram utilizadas como palavras-chave de entrada um conjunto de palavras previamente armazenadas nas tabelas *Lugar*, *Situacao* e *Evento* do banco de dados proposto neste trabalho.

[1] https://www.ruby-lang.org/pt/
[2] https://github.com/sferik/twitter
[3] https://www.postgresql.org/

Para elencar palavras-chave relacionadas a situação e evento de trânsito foi realizada uma análise manual de palavras/expressões pertencentes à este contexto nos textos de um conjunto de *tweets* referentes à trânsito. As palavras/expressões resultantes desta análise foram então inseridas nas tabelas *Situacao* e *Evento*. Além disso, uma versão simplificada de cada palavra/expressão foi armazenada no atributo *matching* de cada uma dessas tabelas (ver seção 2.4.3).

Para a obtenção de palavras-chave relacionadas à lugares foram utilizados dados no fomato OSM publicados pelo *Open Street Maps*[1], que é um mapa colaborativo que contém diferentes tipos de informações sobre regiões de todo o mundo. Foram extraídas do arquivo OSM mais de 1600 nomes de vias da grande Florianópolis, que foram então inseridas na tabela *Lugar* do banco de dados proposto neste documento. Da mesma forma como nas palavras/expressões referentes a situação e evento de trânsito, uma versão simplificada de cada um dos nomes de vias extraídos foi armazenada no atributo *matching* da tabela *Lugar* (ver seção 2.4.3).

A avaliação dos algoritmos propostos foi realizada pelo cálculo das medidas de precisão e *recall* para cada um deles. Utilizou-se para o cálculo de precisão e *recall* as definições de KENT et. al. (1955). As fórmulas abaixo apresentam a descrição de como essas métricas foram calculadas nos experimentos realizados:

$$\text{precisão} = \frac{\text{total de informações extraídas corretamente}}{\text{total de informações extraídas}} \tag{4.1}$$

$$\text{recall} = \frac{\text{total de informações extraídas corretamente}}{\text{total de } \textit{tweets} \text{ utilizados na avaliação}} \tag{4.2}$$

Uma informação foi considerada como corretamente extraída se e somente se o algoritmo identificou uma palavra-chave referente a esta informação e esta palavra-chave estava presente no texto do *tweet*. Os *tweets* utilizados para a avaliação de cada um dos algoritmos foram selecionados manualmente para garantir que eles possuíam a informação de interesse para ser extraída. A tabela 7 apresenta exemplos dessa avaliação. A primeira coluna mostra o texto do *tweet* analisado, a segunda coluna mostra a palavra-chave identificada, e a terceira coluna

[1] https://www.openstreetmap.org

apresenta a avaliação se a informação foi corretamente extraída.

Tabela 7: Exemplo de avaliação da extração de informação dos *tweets*

Texto do tweet	Palavra-chave	Avaliação
Bom dia! Trânsito intenso na região da Grande Florianópolis, com lentidão na **BR 282**, Via Expressa, acesso a Ilha de Florianópolis. #t24horas	Rua Florianópolis	Incorreto
BR-101 São José: acidente no Km 207 ocasiona 3 Km de congestionamento sentido Palhoça. #T24horas	Rua São José	Incorreto
Trânsito bem intenso na **Av. Marinheiro Max Schramm** entre o Cambirela Hotel e a Globo Nissan. Foto: Morales #t24horas https://t.co/WwQL0PgFhX	Avenida Marinheiro Max Schramm	Correto

Foi utilizada uma ferramenta de NER (Reconhecimento de Entidades Nomeadas) para comparação com os resultados do algoritmo de identificação de lugares, visto que esta ferramenta tem potencial interessante para extrair este tipo de informação. Para o treino da fe-rramenta de NER foi utilizado o *corpus* anotado Amazônia[1] que contém mais de 4 milhões de palavras anotadas no português. Ao todo foram coletados 1500 *tweets* publicados entre julho e setembro de 2016 para realização dos experimentos, sendo que um subconjunto de 1220 foi utilizado para avaliação dos algoritmos de extração de *Lugar* e *Situação* de trânsito, e um subconjunto de 450 foi utilizado para avaliação do algoritmo de extração de *Eventos* no trânsito.

4.1.1 Experimento 1 - Extração de Lugar

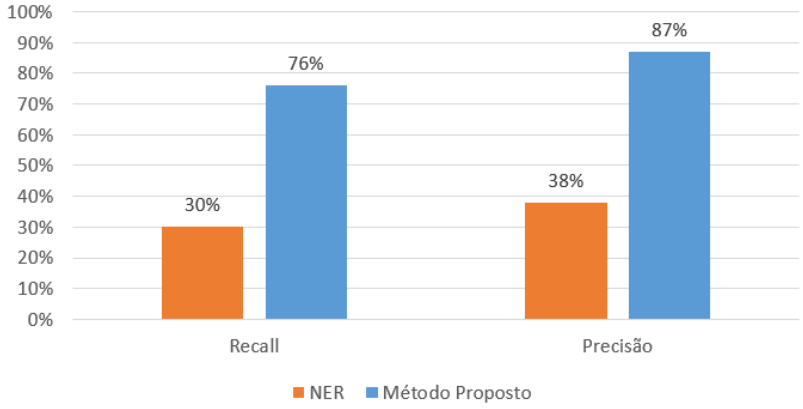
O primeiro experimento avaliou o *recall* e precisão do algoritmo de identificação de lugares. Além disso, foi realizada a comparação entre os resultados do algoritmo e a ferramenta Open NLP (BALDRIDGE 2005), utilizada para o reconhecimento de entidades nomeadas (NER). Ela é uma biblioteca Java baseada em aprendizado de máquina que suporta o reconhecimento de entidades nomeadas em textos para diferentes categorias, sendo as categorias relevantes para comparação neste

[1] http://www.linguateca.pt/floresta/ficheiros/gz/

experimento: lugar e pessoa. Esta última porque nomes de vias geralmente referem-se a nomes de indivíduos.

O gráfico da figura 13 mostra o resultado deste experimento. Nele, é possível notar que o *recall* e precisão do algoritmo proposto para identificação de lugares ficaram acima de 70%. É possível notar ainda que os valores medidos para o algoritmo proposto foram bastante significativos em relação a ferramenta de NER, a qual obteve menos de 40% de precisão e *recall*.

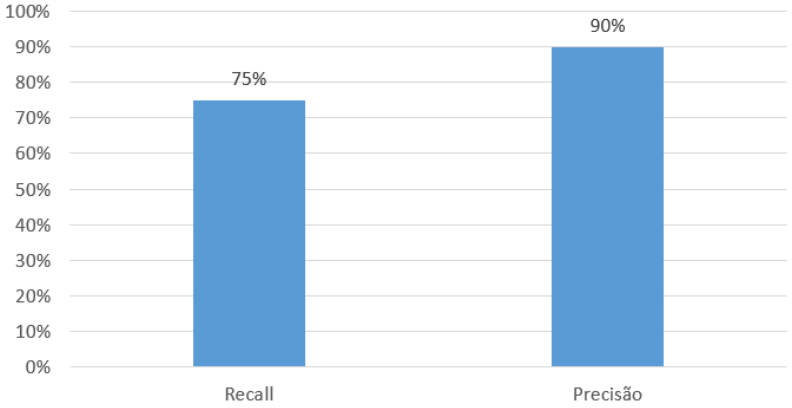
Figura 13: Resultado da Avaliação do Algoritmo de Identificação de Lugar



4.1.2 Experimento 2 - Extração de Situação de Trânsito

No segundo experimento foi realizada a avaliação do algoritmo de identificação de situações de trânsito. O gráfico da figura 14 apresenta o resultado obtido. Nota-se que a acurácia deste algoritmo foi quase idêntica aquela do algoritmo de identificação de lugares. No entanto, a precisão foi um pouco maior. Um dos principais problemas observados neste algoritmo foi a identificação de situações de trânsito onde as palavras que identificam a situação encontram-se muito separadas como em: "O trânsito na Beira-mar norte flui com bastante dificuldade". Neste caso, a situação de trânsito é identificada pelas palavras "trânsito", "flui", e "dificuldade", que estão distanciadas no texto.

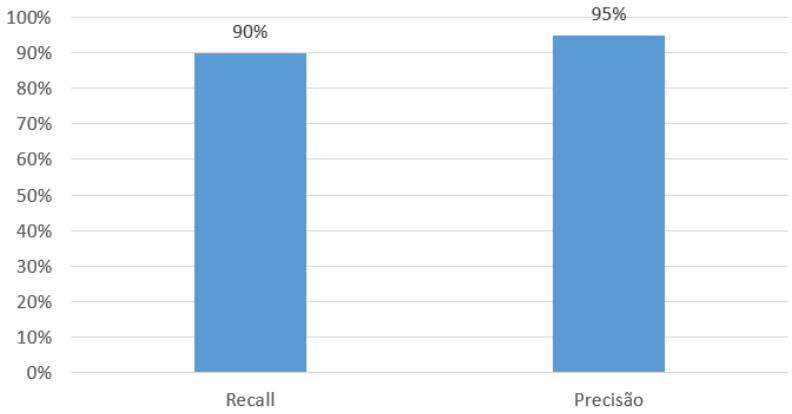
Figura 14: Resultado da Avaliação do Algoritmo de Identificação de Situação de Trânsito



4.1.3 Experimento 3 - Extração de Eventos que Influenciam no Trânsito

O terceiro experimento avaliou o algoritmo de identificação de eventos em trânsito. O gráfico da figura 15 apresenta os valores de acurácia e precisão. Nota-se que os valores foram significamente maiores que os medidos para os algoritmos de identificação de lugar e situação de trânsito. Uma das principais razões para este aumento está no fato da maioria dos eventos serem identificados por uma única palavra, o que melhora o processo de identificação.

Figura 15: Resultado da Avaliação do Algoritmo de Identificação de Eventos



4.2 USO DE INFORMAÇÕES DE TRÂNSITO NA INTERPRETAÇÃO DE TRAJETÓRIAS

Os experimentos realizados nesta seção objetivam avaliar a viabilidade de uso das informações de trânsito extraídas dos *tweets* com trajetórias de veículos que trafegam pelas vias de uma cidade. Todas as trajetórias utilizadas foram coletadas na cidade de Florianópolis-SC. A coleta dessas trajetórias foi realizada utilizando-se dispositivos GPS e também o applicativo *GPSLogger*[1] disponível para o sistema Android[2].

Ao todo foram realizados 3 experimentos: O primeiro mostrando o uso das informações extraídas dos *tweets* para identificação da situação de trânsito em uma via no mesmo momento de uma trajetória coletada na mesma via. O segundo utilizando as informações extraídas dos *tweets* para identificar eventos de trânsito ocorrendo no momento de uma trajetória. O terceiro utilizando as informações de trânsito extraídas de *tweets* para validar um dos algoritmos propostos por Aquino (2014), que identifica desvios em trajetórias e infere se o desvio tem relação com uma possível situação de congestionamento.

[1] https://play.google.com/store/apps/details?id=com.mendhak.gpslogger&hl=en
[2] https://www.android.com/

4.2.1 Experimento 1 - Uso de Informações de Congestionamentos

O objetivo deste experimento foi anotar trajetórias coletadas em uma movimentada avenida da cidade de Florianópolis-SC (avenida Beira-Mar) utilizando-se informações acerca da condição de trânsito nesta avenida publicadas pelo Trânsito 24 horas. Para este experimento foram utilizadas duas trajetórias, uma coletada no dia 01/10/2013 em um momento de congestionamento da via, e outra no dia 02/10/2013 em um momento sem congestionamento.

Para cada uma das trajetórias coletadas foi utilizada a faixa de horário de coleta delas para a busca das informações extraídas dos *tweets* no banco de dados. A trajetória do dia 01/10/2013 foi coletada entre os horários de duas e três da tarde, sendo que foi encontrada uma situação de trânsito oriunda de um *tweet* publicado às 14:15 do mesmo dia, indicando a ocorrência de congestionamento da via. A trajetória do dia 02/10/2013, foi coletada entre uma e duas da tarde, sendo que foi encontrada a informação de que não havia uma situação de congestionamento da via, identificada pelo uso da expressão "trânsito moderado", publicada às 13:07. Em ambas as análises, verificou-se que não houve a negação da situação de trânsito.

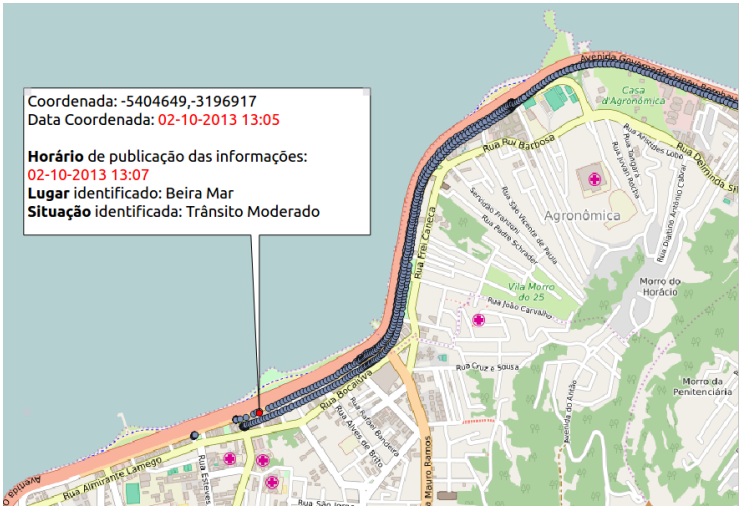
A figura 16 apresenta parte do resultado da consulta realizada para obtenção da situação de trânsito na Beira-mar entre os horários de coleta das trajetórias. Nela, encontram-se em destaque as situações de trânsito encontradas. As figuras 17 e 18 mostram as trajetórias coletadas plotadas em mapa. Os balões de cada figura apontam para pontos da trajetória que foram coletados em horário próximo ao horário de publicação das informações de trânsito recuperadas.

Figura 16: Resultado da consulta para recuperação de situação de trânsito

	data timestamp with time zone	lugar character varying(100)	situacao character varying(100)	negacao boolean
1	2013-10-01 14:15:27-03	Avenida Beira Mar	movimento intenso	f
2	2013-10-01 14:35:02-03	Ponte Colombo Salles	trânsito fluindo	f
3	2013-10-02 12:19:00-03	Rodovia SC-405	fila	f
4	2013-10-02 12:24:27-03	Rua Antônio Luz	fluxo intenso	f
5	2013-10-02 13:07:06-03	Avenida Beira Mar	transito moderado	f
6	2013-10-02 13:47:12-03	Rodovia SC-405	trânsito lento	f
7	2013-10-02 14:00:54-03	Rodovia SC-405	trânsito lento	f

Figura 17: Uso de situação de trânsito extraído de tweets em situação de congestionamnto

Figura 18: Uso de situação de trânsito extraído de tweets em situação de não congestionamento



4.2.2 Experimento 2 - Uso de Informações de Eventos

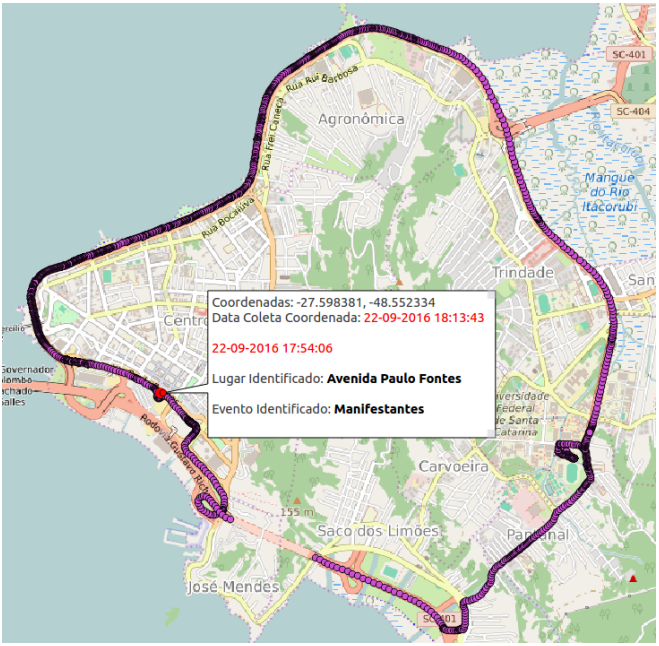
Neste experimento foi realizada a coleta de uma trajetória no mesmo horário em que ocorria uma manifestação na cidade de Flo-

rianópolis-SC. A coleta da trajetória ocorreu entre as 17 e 21 horas e abrangeu a maioria das regiões percorridas pelos manifestantes. Foi então utilizado o algoritmo de identificação de *eventos* para extração de informações relacionadas a esta manifestação. Neste experimento também foi utilizado um ponto da trajetória como referência para a busca da informação do evento. A figura 19 apresenta o resultado da consulta utilizada para recuperação das informações sobre a ocorrência do evento na região de coleta da trajetória. Nela, encontra-se em destaque a informação de que haviam manifestantes na via onde a trajetória foi coletada em um horário próximo ao horário de coleta do ponto da trajetória utilizado como referência para a análise. A figura 20 apresenta a trajetória coletada plotada em mapa, bem como a identificação do ponto de referência utilizado.

Figura 19: Resultado da consulta para busca de eventos

	data timestamp with time zone	lugar character varying(100)	evento character varying(100)
1	2016-09-22 08:44:45-03	Rua Biguaçu	acidente
2	2016-09-22 09:27:49-03	Avenida Beira Mar	acidente
3	2016-09-22 11:20:25-03	Rua Pedro Ivo	acidente
4	2016-09-22 13:20:35-03	Rua Florianópolis	paralisação
5	2016-09-22 16:19:16-03	Rua Florianópolis	paralisação
6	2016-09-22 17:06:20-03	Rua João Pinto	manifestantes
7	2016-09-22 17:54:06-03	Avenida Paulo Fontes	manifestantes
8	2016-09-22 18:28:19-03	Avenida Paulo Fontes	manifestantes
9	2016-09-22 20:01:27-03	Avenida Beira Mar	passeata
10	2016-09-22 20:42:40-03	Rua Joinville	acidente
11	2016-09-22 20:43:56-03	Rodovia Gustavo Richard	protesto
12	2016-09-22 20:44:33-03	Avenida Beira Mar	manifestação

Figura 20: Exemplo do uso de evento extraído de tweets na análise de trajetórias de objetos móveis



4.2.3 Experimento 3 - Validação do Método de Aquino (2014)

Este último experimento foi realizado no intuito de utilizar as informações de trânsito extraídas de *tweets* para validar o algoritmo de detecção de desvios em trajetórias relacionados à congestionamentos (*Traffic Avoiding Outliers*) proposto no trabalho de Aquino (2014). O método dele utiliza cálculos matemáticos para a identificação de desvios relacionados à congestionamentos. Dessa meneira, utilizou-se as informações de trânsito oriundas de *tweets* para averiguar se realmente havia um congestionamento na via no momento em que um desvio foi identificado.

Para o experimento foram utilizadas trajetórias seguindo um caminho padrão e outras fazendo um desvio em relação à este caminho entre os horários de meio-dia e três da tarde. Foi então utilizado o algoritmo de Aquino (2014) para identificar desvios realizados com o objetivo de evitar uma possível situação de congestionamento no caminho padrão, neste caso a avenida Beira-mar norte. A figura 21 apresenta o resultado deste experimento. A trajetória em preto mostra um desvio identificado na avenida Beira-mar e que foi classificado pelo método de Aquino (2014) como um desvio devido a um congestionamento.

Para validar a veracidade da classificação realizada pelo método de Aquino (2014), foram coletados *tweets* publicados na mesma data de coleta da trajetória (Outubro de 2013) e então foram executados os algoritmos propostos neste trabalho para extrair informações de trânsito desses *tweets*. As informações extraídas foram armazenadas em um banco de dados que foi então consultado para verificar se haveria alguma informação relacionada à condição do trânsito na avenida Beria-mar no momento de ocorrência do desvio identificado.

A figura 22 apresenta em destaque as informações de trânsito encontradas. É possível notar que no momento de ocorrência do desvio identificado havia uma situação de fila na avenida Beira-mar, mostrando que o algoritmo de Aquino (2014) conseguiu identificar um desvio devido a uma situação de congestionamento. Não foram encontradas informações de eventos relacionadas ao congestionamento identificado.

Figura 21: Identificação de *traffic avoiding outliers*, adaptado de Aquino (2014)

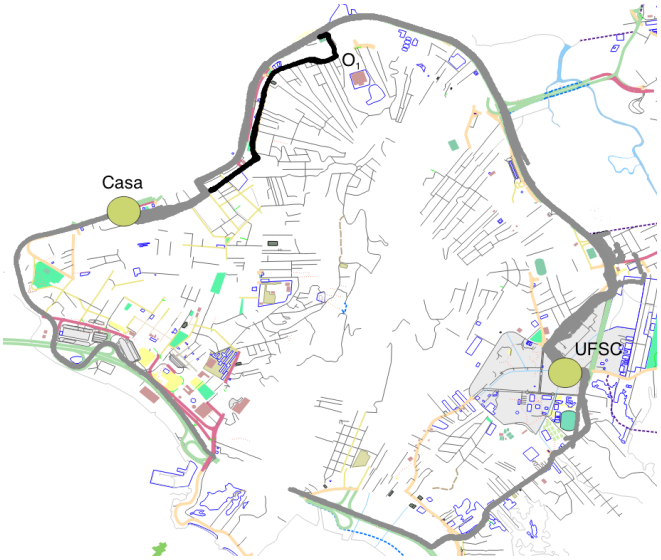


Figura 22: Resultado de Consulta para Busca de Situações de Congestionamento

	data timestamp with time zone	lugar character varying(100)	situacao character varying(100)	negacao boolean
1	2013-10-03 12:55:00-03	Ponte Colombo Salles	movimento intenso	f
2	2013-10-03 12:55:00-03	Ponte Colombo Salles	movimento intenso	f
3	2013-10-03 13:02:00-03	Avenida Blumenau	lentidão	f
4	2013-10-03 13:24:00-03	Rodovia SC-405	trânsito lento	f
5	2013-10-03 13:49:00-03	Rua Tijucas	trânsito lento	f
6	2013-10-03 14:03:00-03	Avenida Beira Mar	fila	f
7	2013-10-03 14:33:00-03	Rodovia SC-401	trânsito flui	f
8	2013-10-03 14:41:00-03	Rua São José	trânsito lento	f
9	2013-10-03 14:46:00-03	Rua Flórida	fila	f
10	2013-10-03 15:44:00-03	Rua São José	trânsito lento	f
11	2013-10-03 15:47:00-03	Rodovia Gustavo Richar	flui bem	f
12	2013-10-03 15:48:00-03	Avenida Beira Mar	fila	f

5 CONCLUSÃO E TRABALHOS FUTUROS

A proliferação de disposivos capazes de coletar trajetórias de indivíduos gera uma massa de dados que podem ser estudados no in-tuito de extrair diferentes padrões desses dados. Muitos trabalhos foram publicados com foco na mineração dos dados de trajetórias na sua forma bruta e estudos mais recentes buscam agregar informações do contexto no qual uma trajetória foi coletada, como a identificação dos lugares de parada durante uma trajetória (ALVAREZ et al., 2007) e a identificação de situações de congestionamentos durante um desvio em relação a um caminho padrão (AQUINO et. al, 2014). No entanto, pouco tem sido explorado sobre o uso de dados de redes sociais neste processo, o que foi um fator motivador para o desenvolvimento deste trabalho.

O Twitter, por ter como principal conceito a estrutura de microblog, possui contas de usuário focadas na publicação de informações específicas a um determinado contexto e, dentre elas, as contas voltadas para publicação de informações de trânsito. Neste trabalho, foi proposto um método para extrair e organizar este tipo de informação do Twitter de forma que sejam utilizadas para a interpretação de trajetórias de veículos trafegando pelas vias de uma cidade.

Os resultados obtidos com o método proposto se mostraram bastante satisfatórios já que obtiveram medidas de *recall* acima de 70% e taxas de precisão em torno de 90% em todas as análises. Estes valores são considerados significativos porque mostram uma cobertura expressiva das informações que devem ser extraídas e um baixo número de falsos positivos (informações extraídas erroneamente). Além disso, conforme apresentado nos experimentos, o método proposto neste trabalho permite utilizar dados de *eventos* relacionados à trânsito com trajetórias de veículos, o que não havia sido explorado até o momento. Por fim, outra contribuição foi propor uma metodologia que permite relacionar dados de trajetórias de veículos com dados de trânsito publicados no Twitter, uma vez que se tem o lugar e a data de uma situação/evento de trânsito permitindo fazer a relação com o lugar/data de coleta de trajetórias de veículos.

Como trabalhos futuros cita-se:

- Adicionar um passo de identificação do sentido de uma situação de trânsito. Por exemplo, no texto: “Via Expressa: Lentidão no sentido Ilha, flui bem em direção à BR-101.”, este passo poderia identificar que a lentidão ocorre no sentido ilha da via expressa;

- Adicionar um passo de quebra do texto do *tweet* em frases, pois alguns *tweets* citam mais de um lugar ao mesmo tempo, como por exemplo: “Fila na BR-101 norte. Movimento moderado na Av. Gustavo Richard. #t24horas”. Nesta caso, seria melhor quebrar o texto em duas frases, “Fila na BR-101 norte”e “Movimento moderado na Av. Gustavo Richard.”;

- Testar o uso de ferramentas de NER treinadas com um *corpus* gerado dos próprios *tweets* para realizar uma nova avaliação de desempenho frente ao método proposto;

- Avaliação das métricas de precisão e *recall* na extração de informações de trânsito de *tweets* publicadas em contas do Twitter de outras cidades;

- Utilização das informações de trânsito extraídas de *tweets* para melhorar as interpretações realizadas por algoritmos que identificam padrões em dados de trajetórias brutas, por exemplo, o algoritmo de identificação de comportamentos anômalos em trajetórias de motoristas (CARBONI, 2014);

- Estender o método proposto para que seja possível relacionar trajetórias brutas de veículos com as situações de trânsito de forma automática. Isto pode ser realizado através do uso de dados espaço-temporais associados aos nomes de lugares.

Além dos items citados acima, as informações extraídas dos *tweets* podem também ser usadas para analizar o trânsito na cidade de Florianópolis-SC, uma vez que se sabe as ruas problemáticas, as ruas com maior ocorrência de congestionamento e os diferentes tipos de eventos de trânsito que acontecem na cidade.

REFERÊNCIAS

ALVARES, L. O.; BORGONY, V.; KUIJPERS, B.; MACEDO, J. A. F.; MOELANS, A.; VAISMAN, A. **A model for enriching trajectories with semantic geographical information**: Proceedings of the 15th acm international symposium on advances in geographic information systems (acm gis 2007). USA, 2007. 162-169 p.

AQUINO, ARTUR R.; ALVARES, L. O.; RENSO, C.; BOGORNY, V. **Towards Semantic Trajectory Outlier Detection**: Xiv brazilian symposium on geoinformatics. Brasil, 2014.

BEBER, MARCO A.; FERRERO, CARLOS A.; FILETO, RENATO.; BOGORNY, VANIA. **Towards Activity Recognition in Moving Object Trajectories from Twitter Data**: Xvii brazilian symposium on geoinformatics. Brasil, Novembro 2016.

BHAT, F. et al. **A Software System for Data Mining with Twitter**: 10th ieee international conference on date of conference. Londres, Setembro 2011.

BOWLING, A. **Measuring Health: a Review of Quality of Life Measurements Scales**. 2nd. ed. Baltimore: Open University Press, 1997. 91 - 109 p.

BRAZ, FERNANDO J.; BOGORNY, VANIA. **Introdução a Trajetórias de Objetos Móveis**. Joinville - SC, 2012. Editora da Univille.

CARBONI, M. E; BOGORNY, V. **Inferring Drivers Behavior through Trajectory Analysis**: Advances in intelligent systems and computing. [S.l.], 2014. 837 - 848 p.

CARVALHO, WESLEY S. **Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina**. São Paulo, fevereiro de 2012.

CULOTTA, ARON. **Detecting influenza outbreaks by analyzing Twitter messages**. Department of Computer Science, Southeastern Louisiana University, Julho de 2010.

DERCZYNSKI, L. et al. **Analysis of named entity recognition and linking for tweets**. [S.l.], 2014. 32 - 39 p.

DESCONHECIDO. **Social Networks and Blogs now 4th most popular online activity**: http://www.nielsen.com/us/en/press-room/2009/social-networks–.html. acessado em: Junho de 2013. [S.l.], 2009.

DOWNEY, D; BROADHEAD, M; ETZIONI, O. **Locating Complex Named Entities in Web Text**. [S.l.], 2007. 2733 - 2739 p.

EBECKEN, N; LOPES, M; COSTA, M. **Mineração de Textos**. Manole, 2003. 337 - 370 p.

FONTES, V. C.; BORGONY, V. **Discovering Semantic Spatial and Spatio-Temporal Outliers from Moving Object Trajectories**. [S.l.], Março 2013.

FOSCA GIANNOTTI; MIRCO NANNI; FABIO PINELLI; DINO PEDRESCHI. **Trajectory pattern mining**. [S.l.], 2007.

J BALDRIDGE. **The opennlp project**. [S.l.], 2005. Disponível em: <http://opennlp.apache.org/index.html>.

KENT, A; BERRY, M; LEUHRS, F. U; PERRY, J. W. **Operational criteria for designing information retrieval systems**: Machine literature searching viii. [S.l.], 1955. 93 - 101 p.

LAUBE, P; IMFELD, S; WEIBEL, R. **Discovering relative motion patterns in groups of moving point objects**. [S.l.], Abril 2005.

LEE, J.; HAN, J.; LI, X. . **Trajectory Outlier Detection: A Partition-and-Detect Framework**. [S.l.], Abril 2008.

MARRERO, M. et al. **Named Entity Recognition: Fallacies, challenges and opportunities**: Computer standards and interfaces. [S.l.], 2013. 482 - 489 p.

MISLOVE, A; VISWANATH, B; GUMMADI, K. P; DRUSCHEL, P. **You are who you know: Inferring user profiles in Online Social Networks**: Proceedings of acm international conference of web search and data mining. Nova York (EUA), 2010.

MORAIS, EDISON A. M; AMBRÓSIO. ANA. P. L. **Mineração de Textos**: Relatório técnico. Universidade Federal de Goiás, 2007.

NADEAU, D; SEKINE, S. **A survey of named entity recognition and classification**: Lingvisticae investigationes. John Benjamins publishing company, 2007. 3 - 26 p.

PIRES, FÁBIO ANTERO. **Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde**. São Paulo, 2012.

SORATO, D; GOULARTE, B. F; NASSAR, M. S; FILETO, R. **Análise de Métodos e Ferramentas para Reconhecimento de Palavras Relevantes em Microblogs**: Xii brazilian symposium on information systems. Universidade Federal de Santa Catarina, 2016.

WIVES, L. **Tecnologias de descoberta de conhecimento em textos aaplicadas à inteligência competitiva**: Exame de qualificação eq-069. PPGC-UFRGS, 2002.

APÊNDICE A – Artigo

ANÁLISE DE DADOS DO TWITTER PARA INTERPRETAÇÃO DE TRAJETÓRIAS DE OBJETOS MÓVEIS

Thiago Thalison Firmino de Lima[1], Vania Bogorny[2]

[1]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Florianópolis - SC - Brazil

[2]Departamento de Informática e Estatística – Universidade Federal de Santa Catarina (UFSC)
Caixa Postal P.476 - Florianópolis - SC - Brazil

thalisonfirmino@gmail.com, vania.bogorny@ufsc.br

Abstract. *The Moving Objects Trajectory Analysis aim attention at the creation of methods that identi es behaviour of individuals during their movement. One of the major challenges related with this kind of analysis is the gathering of data that describes the environment in which these trajectories belong. In this article we present a method that extract, organize and store tra c information published on Twitter to be used as a source of information in the analysis of trajectories of vehicles.*

Resumo. *A análise de trajetórias de objetos móveis busca a criação de métodos que permitam identi car o comportamento de indivíduos através da análise de trajetórias realizadas por eles. Um dos principais desa os relacionados à este tipo de análise está em como obter dados que descrevam o ambiente onde estas trajetórias ocorrem, melhorando assim as análises realizadas. Neste artigo é apresentado um método para extração, organização e armazenamento de informações de trânsito publicadas no Twitter para que sejam usadas como dados de contexto na análise de trajetórias de veículos.*

1. Introdução

A área de trajetórias de objetos móveis é o campo de pesquisa da Ciência da Computação que busca propor métodos que permitam fazer a análise de trajetos de indivíduos através da extração de diferentes informações como o caminho mais utilizado

entre duas regiões [Fosca et al., 2007], a identi cação de comportamentos anômalos em trajetórias de motoristas [Carboni et al., 2014], a identi cação de desvios realizados em relação à um caminho padrão [Lee et al., 2008], entre outros.

A grande maioria dos trabalhos publicados visam a análise de trajetórias utilizando-se somente os dados brutos da trajetória, que são um conjunto de pontos (*x, y, t*) onde *x* e *y* representam coordenadas geográ cas e *t* o instante de tempo no qual essas coordenadas foram coletadas. Esses dados limitam as interpretações que podem ser realizadas pelo fato de não possuírem informações do contexto no qual as trajetórias ocorrem, impossibilitando responder perguntas interessantes como: "O que teria levado um indivíduo a fazer um desvio de trajeto?", "Por que um motorista fez um movimento anômalo durante um percurso?", "Qual seria o motivo para um grupo de indivíduos seguissem um trajeto comum?", entre outras. Neste cenário, acredita-se que dados publicados em redes sociais (OSNs - Online social Networks) possuem um potencial interessante para ser utilizado como este tipo de fonte, isto porque, apesar de serem relativamente recentes (criadas em meados dos anos 2000), são utilizadas por mais de dois terços da população que acessa a internet [N.O Report., 2013], o que gera um grande volume de informações diariamente.

O Twitter possui contas contas de usuário utilizadas para publicação de mensagens acerca das condições de trânsito em determinadas regiões de uma cidade como a conta "O dia 24 horas" (Rio de Janeiro), a "Trânsito SP" (São Paulo), a "Trânsito Zero Hora" (Porto Alegre), e a "Trânsito 24 horas" (Florianópolis). Essas informações de trânsito, se extraı das e bem estruturadas, podem ser úteis para se entender o ambiente onde trajetórias de veículos são analisadas. Neste artigo é proposto um método para extrair, organizar e armazenar informações de trânsito publicadas no Twitter de forma que possam ser utilizadas como fonte de informação na interpretação de trajetórias de veículos. Também são apresentados experimentos mostrando a viabilidade de uso dessas informações no contexto de trajetórias de veículos.

2. Conceitos Básicos e Trabalhos Relacionados

Nesta seção são os apresentados os principais conceitos que embasam o desenvolvimento do método descrito neste artigo.

2.1. Trajetórias de Objetos Móveis

Uma trajetória é a sequência dos pontos que caracteriza a localização espacial de um objeto em um intervalo de tempo [Braz e Bogorny, 2012]. A representação mais simples de uma trajetória, denominada de trajetória bruta, consiste em um conjunto de pontos representados pelas seguintes informações: *tid*, *x*, *y*, *t*. O atributo *tid* é o identi cador da trajetória; os atributos *x* e *y* são, respectivamente, as coordenadas

geográ cas referentes a latitude e a longitude do objeto; e o atributo *t* refere-se ao instante de tempo no qual o ponto foi gerado.

A análise de trajetórias brutas com o objetivo de propor técnicas computacionais capazes de extrair padrões de comportamento delas tem sido um tema explorado em diversos estudos cientı cos. No método proposto em Carboni (2014), por exemplo, os dados brutos de trajetórias são utilizados para a identi cação de comportamentos anômalos em trajetórias de motoristas. Tais comportamentos são identi cados pelo uso de cálculos matemáticos que detectam acelerações, frenagens e mudanças bruscas de direção no decorrer da trajetória. Além disso, o método utiliza as informações detectadas para fazer a classi cação dos motoristas em 4 categorias distintas: *Careful Driver*, motoristas que não realizaram nenhum comportamento anômalo durante a trajetória, *Distracted Driver*, motoristas que zeram mudanças bruscas de movimento quando da ocorrência de algum evento na trajetória, *Dangerous Driver*, motoristas que zeram mudanças bruscas de movimento sem motivo aparente e, por m, *Very Dangerous Driver*, motoristas que zeram mudanças bruscas de velocidade como acelerações acima da velocidade média da via ou frenagens repentinas, podendo também envolver mudanças bruscas de movimento. A gura 1 mostra um exemplo de aplicação do método de Carboni (2014). Nela, a imagem (a) mostra a trajetória analisada e a imagem (b) apresenta a marcação dos pontos desta trajetória onde foram identi cados comportamentos anômalos.

Figura 1. Análise de comportamento anômalo em trajetórias de motoristas, adaptado de Carboni (2014)

No trabalho de Fontes e Bogorny (2013), é proposto um algoritmo para detectar trajetórias *outliers*. Um outlier em trajetórias é um objeto que se move de forma diferente em relação à maioria dos objetos que realizam trajetória semelhante

a ele. No método dele são de nidas regiões de interesse entre as quais as trajetórias são analisadas, o que permite a detecção de trajetórias outliers em relação à um caminho padrão entre essas regiões. A gura 2 apresenta um exemplo. Nela são apresentadas 6 trajetórias denominadas *T1*, *T2*, *T3*, *T4*, *T5* e *T6*. Supondo que as regiões *R1* e *R2* sejam um hotel e um restaurante respectivamente. Seria possı vel utilizar o algoritmo de Fontes e Bogorny (2013) para identi car que nas trajetórias *T2*, *T3* e *T4* foi seguida uma rota padrão entre os dois locais, ao passo que nas trajetórias *T1*, *T5* e *T6* foi seguida uma rota alternativa.

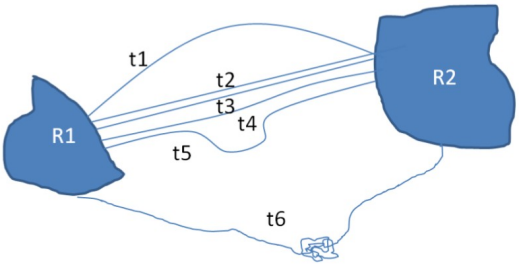


Figura 2. Exemplo de *outliers*, adaptado de Fontes et al. (2013)

Uma das grandes di culdades encontradas na análise de dados brutos de trajetórias está no fato de que este tipo de dado não dispõe de informações referentes ao contexto no qual a trajetória foi gerada. Sendo assim, é possível identi car comportamentos nessas trajetórias, mas é di cultoso elencar possíveis motivos para o mesmo. Por exemplo, no trabalho de Carboni (2014), não se tem a informação do tipo de evento que teria in uenciado em um comportamento anômalo por parte do motorista, que pode ser um congestionamento, uma obra na pista, um alagamento, entre outros.

Um dos primeiros esforços para solucionar o problema de identi cação de motivos que in uenciam comportamentos em trajetórias foi realizado em Aquino (2014), que estende o trabalho de Fontes (2013), propondo uma solução capaz de identi car motivos de desvios em trajetórias. Apesar do método de Aquino (2014) fazer a identi cação de desvios relacionados à eventos, não se tem a informação do tipo de evento identi cado. Além disso, são necessários um número mínimo de trajetórias para identi car se o desvio foi realizado no intuito de evitar congestionamentos, visto que este método realiza cálculos matemáticos sobre dados de trajetórias brutas para identi car esta informação.

2.2. Rede Social Twitter e Contas de Usuário Focadas na Divulgação de Informações Sobre Condições de Trânsito

Dentre as principais Redes Sociais utilizadas na atualidade, a rede social Twitter destaca-se pelo número signi cativo de usuários. Ela foi criada em 2006 e está organizada na forma de um microblog que permite aos seus usuários a publicação de textos breves (máximo de 140 caracteres), que poderão ser visualizados publicamente, ou apenas por um grupo restrito de seguidores, conforme o usuário achar mais conveniente. Com a popularização do Twitter, também se popularizaram contas de usuário utilizadas para publicação de *tweets* relacionados à um tema especí co. Neste contexto, estão as contas utilizadas para publicação de informações sobre as condições de trânsito em uma cidade, informando situações de congestionamento nas diferentes vias e possíveis eventos que estejam in uenciando esses congestionamentos.

2.3. Extração de Informação em Dados do Twitter

Uma das formas de analisar os dados do Twitter é através da extração de informações presentes nos textos dos *tweets*. Um exemplo é o trabalho de Aron Culotta (2010), no qual foi realizado um estudo de dados do Twitter para detectar surtos de gripe nos Estados Unidos pela análise de cerca de 570 milhões de tweets coletados durante os meses de setembro de 2009 a maio de 2010. No método proposto, foram utilizadas técnicas simples de string matching entre o tweet e palavras-chave relacionadas à gripe. Após esta etapa, foi realizada a contagem de tweets para os quais houve matching, e então foi calculada a proporção deles em relação ao total de tweets coletados. Como resultado da pesquisa foi possı vel detectar uma forte correlação entre a proporção de tweets para os quais houve matching, e as estatı sticas semanais do U.S. Centers for Disease Control and Prevention (CDC) acerca dos surtos de gripe nos Estados Unidos. Por exemplo, utilizando apenas a palavra-chave *u* (gripe em Inglês), foi possı vel detectar uma correlação de 81% entre a proporção de tweets e as estatı sticas do CDC.

Além do texto do tweet, a capacidade de agregar dados geográ cos nos *tweets*, disponibilizada a partir de 2010, tornou possı vel saber quais as coordenadas geográ cas de onde um usuário compartilhou um texto, o que possibilitou pesquisas utilizando também este tipo de dado. Por exemplo, na pesquisa de Mislove et al. (2010), este tipo de dado é utilizado para identi cação do nı vel de humor em diferentes regiões dos Estados Unidos. Para tal, são utilizadas palavras-chave para classi cação do tweet entre os diferentes nı veis de humor, e o dado georreferenciado para identi cação da região do qual ele foi postado. Como resultado deste método obteve-se a caracterização do estado emocional das pessoas em diferentes regiões dos Estados Unidos.

Os *tweets* de trânsito utilizados não possuem dados georreferenciados para identi car a via da qual uma condição de trânsito está sendo informada, no entanto, o próprio texto do *tweet* contém a identi cação da região sobre a qual a condição de trânsito está sendo informada. Dessa forma foi possível desenvolver um método que utiliza palavras-chave para extração tanto da condição de trânsito como do referido lugar.

3. Proposta para Extração de Informações de Trânsito em Textos Publicados no Twitter

Para o desenvolvimento do método de extração de informações referentes à trânsito foi realizada a identi cação das informações presentes nos *tweets* de trânsito que precisam ser extraı das para a utilização com trajetórias de veı culos. São elas: a *Data* de publicação do *tweet,* que serve para identi car a data de ocorrência de uma situação de trânsito, o *Lugar* do qual a informação de trânsito está sendo informada, que serve para identi car as condições de trânsito referentes ao lugar de ocorrência de trajetórias, a *Situação de Trânsito*, que informa a presença ou não de situações de congestionamento e, por m, os *Eventos de Trânsito,* que informam a presença ou não de eventos que possam estar in uenciando no trânsito (ex: alagamentos, manifestações, colisões de veı culos, etc).

A gura 3 apresenta as entradas do método proposto. Ele recebe como entrada um conjunto de palavras-chave referentes às informações de interesse, *lugares*, *situações de trânsito* e *eventos de trânsito*, que são utilizadas para a busca dessas informações no texto do *tweet.* As palavras-chave referentes à lugar foram obtidas através de dados do *Open Street Maps,* as palavras-chave referentes à situação e eventos de trânsito foram obtidas através de uma análise manual de um conjunto de textos de *tweets.* Outro parâmetro recebido pelo método é um valor de *threshold* para o cálculo de similaridade entre as diferentes palavras que compõem o texto do *tweet* e as palavras-chave. Dessa maneira, uma vez que um cálculo resulte em um valor menor ou igual ao *threshold,* a informação é considerada encontrada. Por m, o método também recebe o texto do *tweet* que passará pelo processo de extração de informação.

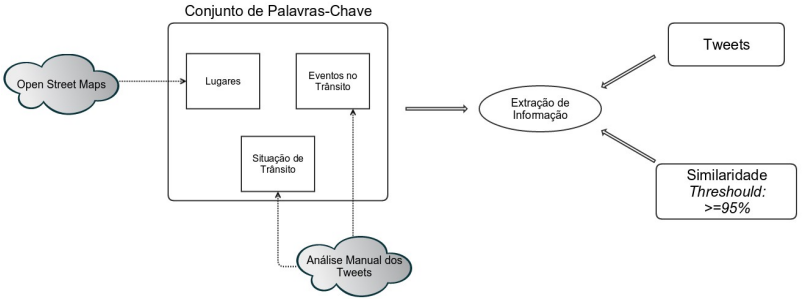


Figura 3. Método para Extração de Informações de Trânsito em Textos de *tweets*

O processo de extração das informações de interesse segue um conjunto de passos que são *tokenização, padronização de strings,* e o *cálculo de similaridade* entre as palavras do texto do *tweet* com as palavras-chave da informação de interesse que será extraída. A gura 4 mostra a sequência dos passos citados que seguem a ordem apresentada da esquerda para a direita:

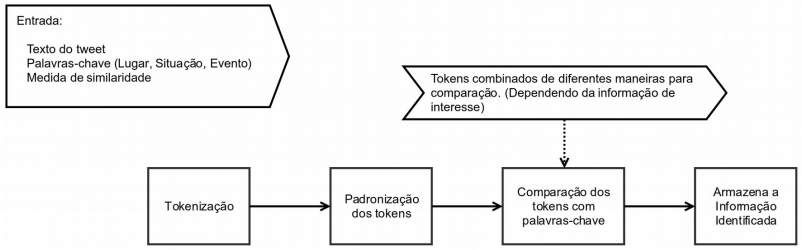


Figura 4. Sequência dos passos no processo de extração de informação

Passo 1 - Tokenização: A tokenização é a quebra do texto em palavras. No processo de tokenização utilizado o critério utilizado foram os espaços em branco entre as palavras. Além disso, são descartados os tokens considerados inúteis para o processo de recuperação de informação realizado neste trabalho, que são: palavras seguidas dos sinais # e @, links, e o token RT. A tabela 1 apresenta um exemplo desse processo.

Texto original do *tweet*	Texto após tokenização
“RT @PRF191SC SC-405: Trânsito segue lento no sentido bairros, entre o Elevado da Seta e o Trevo do Rio Tavares. #t24horas”	[SC-405:] [Trânsito] [segue] [lento] [no] [sentido] [bairros,] [entre] [o] [Elevado] [da] [Seta] [e] [o] [Trevo] [do] [Rio] [Tavares.]

Tabela 1. Exemplo de texto de *tweet* antes e depois do processo de tokenização

Passo 2 - Padronização dos Tokens: Antes do armazenamento das palavras-chave referentes a lugar, situação de trânsito e evento é realizado um mapeamento de caracteres para letras minúsculas, bem como a remoção de hı fen, underline, sinais de pontuação, e de espaços em branco presentes em nomes compostos por mais de uma palavra. Os algoritmos para recuperação de informação realizam um processo de padronização de strings nas palavras do texto do tweet a ser analisado que é similar ao apresentado anteriormente. No entanto, como as palavras do texto do tweet são primeiramente separadas em tokens, não é necessária a remoção de espaços em branco entre elas. A tabela 2 apresenta exemplos desse processo.

Palavra-Chave	Antes da Padronização	Depois da Padronização
Lugar	João Pio Duarte Silva	joaopioduartesilva
Situação De Trânsito	Congestionamento	congestionamento
Evento	colisão	colisao

Tabela 2. Exemplo de Palavras Antes e Depois do Processo de Padronização

Passo 3 - Comparação dos tokens com palavras-chave: Os tokens padronizados são combinados em diferentes grupos para que possam ser comparados com as diferentes palavras-chave através de um cálculo de similaridade. A Função de Similaridade é responsável por encontrar relações entre termos chave e um texto. Para este cálculo é utilizada a função Jaro-Winkler. A função de JaroWinkler (Winkler, 1990) é uma variação da função de Jaro (Jaro, 1995), e tem por objetivo medir a similaridade entre duas strings, levando-se em consideração o tamanho delas. Como resultado, ela retorna um valor entre 0 e 1, onde quanto mais próximo de 1 for este valor, mais similares as duas strings são. A combinação de palavras ocorre entre palavras vizinhas que são

concatenadas de duas em duas e depois de três em três. Caso não haja matching com este processo de combinação as palavras são utilizadas uma a uma para comparação. Exemplo de combinação de tokens dois à dois: [‘br’, ‘101’, ‘transito’, ‘sem’, ‘ la’, ‘no’, ‘momento’] ‘br101’, ‘101transito’, ‘transitosem’, ‘sem la’, ‘ lano’, ‘nomomento’.

As informações extraídas são armazenadas em um banco de dados para posterior utilização com trajetórias. A gura 5 apresenta o modelo conceitual do banco de dados utilizado. A tabela *Tweets* funciona como elemento central para o relacionamento entre os dados recuperados de cada *tweet*. Sendo assim, ela é responsável pelo relacionamento de um *tweet* com o lugar, e/ou a situação de trânsito e/ou o evento recuperados dele. Além disso, ela armazena a data de publicação do *tweet* permitindo recuperar, por exemplo, informações da condição de trânsito de um determinado lugar num dado momento. A tabela *Lugar* armazena um conjunto de palavras-chave referentes à nomes de vias (ruas, avenidas, rodovias, etc) situadas em uma região. A tabela *Situacao* armazena um conjunto de palavras-chave referentes à condições de trânsito como: “ la”, “congestionado”, “trânsito lento”, etc. Por m, a tabela Evento armazena um conjunto de palavras-chave relacionadas à eventos que podem in uenciar no trânsito como: “acidente”, “alagamento”, “tombamento”, etc. Essas palavras-chave são utilizadas como as palavras-chave de entrada nos algoritmos de extração de informação propostos.

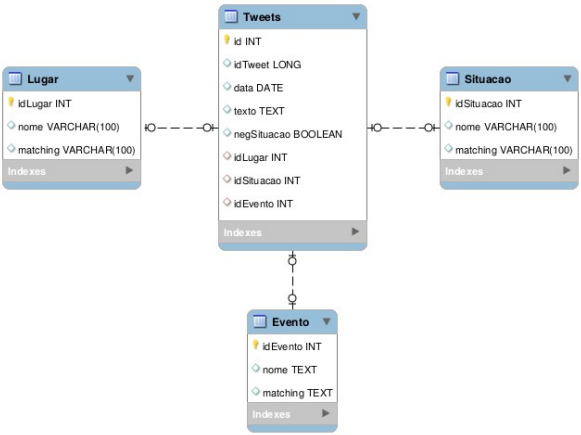


Figura 5. Banco de Dados para Armazenamento de Informações Extraídas

4. Experimentos

Nesta seção são apresentados os experimentos que foram realizados no intuito de avaliar o método proposto, bem como avaliar a viabilidade de uso de informação de trânsito publicadas em *tweets* com trajetórias de veículos.

4.1. Experimento 1 - Extração de *Lugar*

O primeiro experimento avaliou a acurácia e precisão do algoritmo de identi cação de lugares. Além disso, foi realizada a comparação entre os resultados do algoritmo e a ferramenta Open NLP (Baldridge, 2005), utilizada para o reconhecimento de entidades nomeadas (NER). O grá co da gura 6 mostra o resultado deste experimento. Nele, é possı vel notar que a acurácia e precisão do algoritmo proposto para identi cação de lugares caram acima de 70%. É possı vel notar ainda que os valores medidos para o algoritmo proposto foram bastante signi cativos em relação a ferramenta de NER.

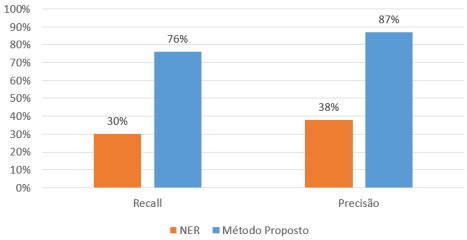


Figura 6. Resultado da Avaliação do Algoritmo de Identificação de Lugares

4.2. Experimento 2 - Extração de *Situação de Trânsito*

No segundo experimento foi realizada a avaliação do algoritmo de identi cação de situações de trânsito. O grá co da gura 7 apresenta o resultado obtido. Nota-se que a acurácia deste algoritmo foi quase idêntica àquela medida para o algoritmo de identi cação de lugares, no entanto a precisão foi um pouco maior. Um dos principais problemas observados neste algoritmo foi a identi cação de situações de trânsito onde as palavras que identi cam a situação encontram-se muito separadas como em: “O trânsito na Beira-mar norte ui com bastante di culdade”. Neste caso, a situação de trânsito é identi cada pelas palavras “trânsito”, “ ui”, e “di culdade”, que estão distanciadas no texto.

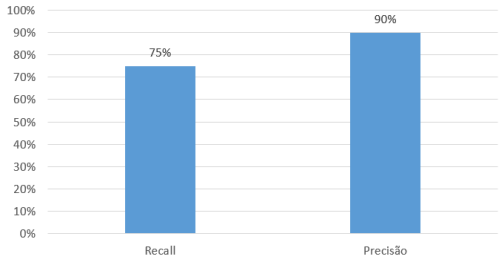


Figura 7. Resultado da Avaliação do Algoritmo de Identificação de Situação de Trânsito

4.3. Experimento 3 - Extração de *Eventos no Trânsito*

O terceiro experimento avaliou o algoritmo de identi cação de eventos em trânsito. O grá co da gura 8 apresenta os valores de acurácia e precisão. Nota-se que os valores foram signi cativamente maiores que os medidos para os algoritmos de identi cação de lugar e situação de trânsito. Uma das principais razões para este aumento está no fato da maioria dos eventos serem identi cados por uma única palavra, o que melhora o processo de identi cação.

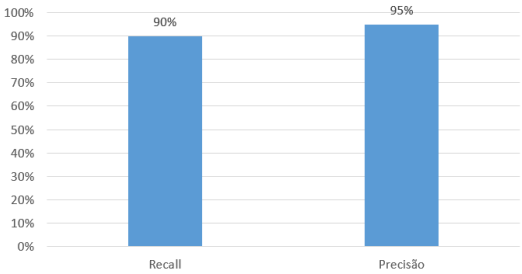


Figura 8. Resultado da Avaliação do Algoritmo de Identificação de Eventos

4.4. Experimento 4 - Avaliação de viabilidade de uso das informações extraídas com trajetórias

Os experimentos apresentados nesta seção objetivam avaliar a viabilidade de uso das informações de trânsito extraı das dos *tweets* com trajetórias de veı culos que trafegam pelas vias de uma cidade. Todas as trajetórias utilizadas foram coletadas na cidade de Florianópolis-SC e os *tweets* foram coletados da conta Trânsito 24 horas. O primeiro experimento foi realizado coletando-se trajetórias em uma movimentada avenida da cidade de Florianópolis-SC (avenida Beira-Mar). Para este experimento foi utilizada uma trajetória coletada no dia 01/10/2013 em um momento de congestionamento da via. A gura 9 mostra a trajetória coletada plotada em mapa. Os balões de cada gura apontam para pontos da trajetória que foram coletados em horário próximo ao horário de publicação de informações de trânsito recuperadas do Twitter. Nota-se que que foram encontradas informações con rmando a situação da via no momento de coleta da trajetória.

Figura 9. Uso de situação de trânsito extraído de *tweets* em situação de congestionamento

Neste experimento foi realizada a coleta de uma trajetória no mesmo horário em que ocorria uma manifestação na cidade de Florianópolis-SC. A coleta da trajetória ocorreu entre as 17 e 21 horas e abrangeu a maioria das regiões percorridas pelos manifestantes. A gura 10 apresenta a trajetória coletada plotada em mapa, bem como a informação de que havia manifestantes na via onde a trajetória foi coletada em um horário próximo ao horário de coleta da trajetória.

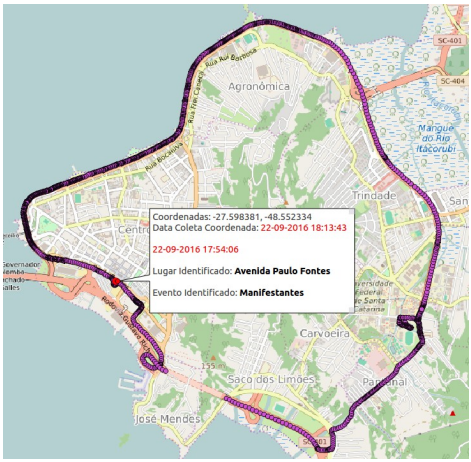


Figura 10. Exemplo do uso de eventos na análise de trajetórias de objetos móveis.

5. Conclusão e Trabalhos Futuros

Os resultados obtidos com o método proposto se mostraram bastante satisfatórios já que obtiveram medidas de recall acima de 70% e taxas de precisão em torno de 90% em todas as análises. Estes valores são considerados signi cativos porque mostram uma cobertura expressiva das informações que devem ser extraı das e um baixo número de falsos positivos (informações extraı das erroneamente). Além disso, conforme apresentado nos experimentos, o método proposto neste trabalho permite utilizar dados de eventos relacionados à trânsito com trajetórias de veı culos, o que não havia sido explorado até o momento. Por m, outra contribuição foi propor uma metodologia que permite relacionar dados de trajetórias de veı culos com dados de trânsito publicados no Twitter, uma vez que se tem o lugar e a data de uma situação/evento de trânsito permitindo fazer a relação com o lugar/data de coleta de trajetórias de veı culos.

Como trabalhos futuros cita-se:

- Avaliação das métricas de precisão e recall na extração de informações de trânsito de tweets publicadas em contas do Twitter de outras cidades;
- Utilização das informações de trânsito extraı das de tweets para melhorar as interpretações realizadas por algoritmos que identi cam padrões em dados de

trajetórias brutas, por exemplo, o algoritmo de identi cação de comportamentos anômalos em trajetórias de motoristas (CARBONI, 2014);

- Estender o método proposto para que seja possı vel relacionar trajetórias brutas de veı culos com as situações de trânsito de forma automática. Isto pode ser realizado através do uso de dados espaço-temporais associados aos nomes de lugares.

Além dos itens citados acima, as informações extraı das dos *tweets* podem também ser usadas para analisar o trânsito na cidade de Florianópolis-SC, uma vez que se sabe as ruas problemáticas, as ruas com maior ocorrência de congestionamento e os diferentes tipos de eventos de trânsito que acontecem na cidade.

Referências

ALVARES, L. O.; BORGONY, V.; KUIJPERS, B.; MACEDO, J. A. F.; MOELANS, A.; VAISMAN, A. **A model for enriching trajectories with semantic geographical information**: Proceedings of the 15th acm international symposium on advances in geographic information systems (acm gis 2007). USA, 2007. 162-169 p.

AQUINO, ARTUR RIBEIRO DE. **Um método para adicionar semântica em outliers de trajetórias de objetos móveis**. Brasil, 2014.

BHAT, F. et al. **A Software System for Data Mining with Twitter: 10th ieee international conference on date of conference**. Londres, Setembro 2011.

BRAZ, F. J.; BOGORNY, V. **Introdução a Trajetórias de Objetos Móveis**. Joinville - SC, 2012. Editora da Univille.

BOWLING, A. **Measuring Health: a Review of Quality of Life Measurements Scales**. 2nd. ed. Baltimore: Open University Press, 1997. 91 - 109 p.

CARBONI, M. E; BOGORNY, V. **Inferring Drivers Behavior through Trajectory Analysis**: Advances in intelligent systems and computing. [S.l.], 2014. 837 - 848 p.

CARVALHO, WESLEY S. **Reconhecimento de entidades mencionadas em português utilizando aprendizado de máquina**. São Paulo, fevereiro de 2012.

CULOTTA, ARON. **Detecting in uenza outbreaks by analyzing Twitter messages**. Department of Computer Science, Southeastern Louisiana University, Julho de 2010.

DERCZYNSKI, L. et al. **Analysis of named entity recognition and linking for tweets**. [S.l.], 2014. 32 - 39 p.

DESCONHECIDO. **Social Networks and Blogs now 4th most popular online activity**: http://www.nielsen.com/us/en/pressroom/2009/social-networks–.html. acessado em: Junho de 2013. [S.l.], 2009.

DOWNEY, D; BROADHEAD, M; ETZIONI, O. **Locating Complex Named Entities in Web Text**. [S.l.], 2007. 2733 - 2739 p.

EBECKEN, N; LOPES, M; COSTA, M. **Mineração de Textos**. Manole, 2003. 337 - 370p.

FONTES, V. C.; BORGONY, V. **Discovering Semantic Spatial and Spatio-Temporal Outliers from Moving Object Trajectories**. [S.l.], Março 2013.

FOSCA GIANNOTTI; MIRCO NANNI; FABIO PINELLI; DINO PEDRESCHI. **Trajectory pattern mining**. [S.l.], 2007.

J BALDRIDGE. **The opennlp project**. [S.l.], 2005. Disponı vel em: <http://opennlp.apache.org/index.html>.

KENT, A; BERRY, M; LEUHRS, F. U; PERRY, J. W. **Operational criteria for designing information retrieval systems**: Machine literature searching viii. [S.l.], 1955. 93 - 101 p.

LAUBE, P; IMFELD, S; WEIBEL, R. **Discovering relative motion patterns in groups of moving point objects**. [S.l.], Abril 2005.

LEE, J.; HAN, J.; LI, X. . **Trajectory Outlier Detection: A Partition-and-Detect Framework**. [S.l.], Abril 2008.

MARRERO, M. et al. **Named Entity Recognition: Fallacies, challenges and opportunities**: Computer standards and interfaces. [S.l.], 2013. 482 - 489 p.

MISLOVE, A; VISWANATH, B; GUMMADI, K. P; DRUSCHEL, P. **You are who you know: Inferring user pro les in Online Social Networks**: Proceedings of acm international conference of web search and data mining. Nova York (EUA), 2010.

MORAIS, EDISON A. M; AMBRÓSIO. ANA. P. L. **Mineração de Textos: Relatório técnico**. Universidade Federal de Goiás, 2007.

NADEAU, D; SEKINE, S. **A survey of named entity recognition and classi cation**: Lingvisticae investigationes. John Benjamins publishing company, 2007. 3 - 26 p.

PIRES, FÁBIO ANTERO. **Ambiente para extração de informação epidemiológica a partir da mineração de dez anos de dados do Sistema Público de Saúde**. São Paulo, 2012.

SORATO, D; GOULARTE, B. F; NASSAR, M. S; FILETO, R. **Análise de Métodos e Ferramentas para Reconhecimento de Palavras Relevantes em Microblogs**: Xii brazilian symposium on information systems. Universidade Federal de Santa Catarina, 2016.

WIVES, L. **Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva**: Exame de quali cação eq-069. PPGC-UFRGS, 2002.