Uso de Dados de Mídias Sociais para Desenvolvimento de Metodologia de Posicionamento de Táxis

Diego O. Rodrigues[1], Thiago H. Silva[2], Marilia Curado[3]
Antonio A. F. Loureiro[4] e Leandro Villas[1]

[1]Instituto de Computação, Universidade Estadual de Campinas. Campinas, Brasil
[2] Dep. de Informática, Universidade Tecnológica Federal do Paraná. Curitiba, Brasil
[3]Dep. de Engenharia Informática, Universidade de Coimbra. Coimbra, Portugal
[4]Dep. de Ciência da Computação, Univ. Federal de Minas Gerais. Belo Horizonte, Brasil

Abstract. *Smart cities emerge as a topic that applies information and communication technology in urban centers to monitor their dynamics and allow the improvement of services for their citizens. This monitoring occurs, for example, when analyzing data produced by citizens in their daily lives. A significant amount of this data has spatio-temporal annotations, which may be used to analyze the city dynamics, such as the mobility flow. Due to these characteristics and also the possibilities brought by their use and analyses, this work presents a novel approach to use social media data to enhance the positioning of taxis within the city. The results show that data from location-based social networks may be used as people's concentration virtual sensor, which can be used by the urban transportation system. The present shows how different urban data sources can be related using spatio-temporal correlation of three different sources was verified. The proposal was validated using data from the taxi system of New York City and also data from the Twitter platform. The analysis was handled using the SMAFramework, a framework to perform analysis in urban mobility data.*

Resumo. *Cidades Inteligentes surgem como um tópico que utiliza tecnologia da informação e comunicação em centros urbanos para monitorar suas dinâmicas e possibilitar que serviços prestados aos seus cidadãos possam ser melhorados. Esse monitormento se dá, por exemplo, por meio da observação de dados gerados pelos cidadãos em suas vidas cotidianas. Uma parcela significativa desses dados contém anotações espaço-temporais que podem ser utilizadas para analisar características específicas das cidades como, por exemplo, seus fluxos de mobilidade. Considerando essas características, este trabalho propõe o uso de dados de mídias sociais para melhorar o posicionamento dos táxis na cidade. Os resultados mostram indícios de que dados de mídias sociais podem ser utilizados como sensores virtuais de concentrações de pessoas em determinados locais, podendo ser usados pelo sistema de transporte urbano. Com o presente trabalho foi possível verificar como diferentes fontes de dados urbanos podem ser relacionadas por meio da correlação espaço-temporal de amostras provenientes de três fontes diferentes. A proposta foi validada usando dados reais de táxis da cidade de Nova Iorque e do Twitter. Essa validação foi feita por meio do SMAFramework, um arcabouço para análise de dados de mobilidade urbana.*

1. Introdução

A complexidade das cidades tem aumentado consideravelmente devido a uma série de fatores, por exemplo, ao elevado número de habitantes vivendo em ambientes urbanos [Nam and Pardo 2011]. Com isso, a Computação Urbana surge como um tópico em Ciência da Computação que utiliza as tecnologias da informação e comunicação em diversos aspectos de uma cidade de forma a prover soluções mais eficientes aos seus cidadãos [Zheng et al. 2014, Nam and Pardo 2011, Pellicer et al. 2013]. Por meio da melhoria na prestação de serviços e, consequente, aumento da interação deles com os cidadãos, usuários desses serviços, as novas cidades, também chamadas cidades inteligentes, visam tornar-se centros urbanos com uma dinâmica diferente da qual estamos acostumados. Tais diferenças vão desde a mobilidade urbana, como conduzir os cidadãos da cidade até seus respectivos destinos de forma eficiente e sem sobrecarregar as vias públicas; até questões governamentais, onde informações sobre o comportamento dos cidadãos podem ser utilizadas para auxiliar na construção de novas políticas públicas. Essas mudanças na dinâmica dos centros urbanos criam condições para melhorar a qualidade de vida dos seus habitantes.

Existem diferentes pontos de vista sobre as cidades inteligentes, sobre o que são e como elas deveriam surgir. Porém, uma das características amplamente aceitas é o uso de tecnologia da informação para gerir os dados gerados na cidade a fim de utilizá-los na melhoria de seus serviços [Nam and Pardo 2011, Pellicer et al. 2013, Hall et al. 2000]. Esses dados são coletados de diversos eventos que ocorrem cotidianamente na cidade, como uma pessoa ao entrar em um ônibus ou comprar comida em um restaurante. Para realizar a coleta desses dados surgem diversas abordagens de modo a reduzir os custos envolvidos com a coleta de dados enquanto melhoram a eficiência do processo. Em ambientes tão complexos como as cidades têm se tornado, manter multidões de agentes humanos realizando entrevistas a pessoas pode não ser a melhor abordagem, já que o custo para manter esses agentes trabalhando é elevado e esse não é um modelo facilmente escalável que consiga acompanhar o ritmo de crescimento das cidades. Além disso, a visão das pessoas sobre vários aspectos, obtida através de uma entrevista, nem sempre é uma representação fiel ou adequada da realidade [Veenhoven 1987, Loughnan et al. 2011].

Para alcançar os ideais de eficiência de execução dos serviços públicos nas cidades, é fundamental a obtenção de dados para entender seu funcionamento. Dados gerados em cidades podem ser originados de diferentes fontes, que podem ser vistas como uma camada de sensoriamento de um fenômeno em particular. Por exemplo, uma alternativa para o monitoramento de cidades são os sensores, desde os mais robustos, que tentam capturar uma grande quantidade de dados, até redes de sensores de baixo custo, que podem ser usados em um maior volume. A utilização desses dois tipos de sensores acontece para que se possibilite a cobertura e, consequente, o monitoramento de maiores áreas no perímetro urbano. Os sensores mais robustos e mais caros são menos numerosos e monitoram eventos à distância. Em contrapartida, os sensores de baixo custo podem ser distribuídos de forma a cobrir maiores regiões e monitorar de perto os eventos nas cidades. A fim de garantir o rápido envio de dados dos sensores às centrais de processamento, esses sensores são comumente conectados em uma rede sem fio.

Mesmo a abordagem de criação de redes de sensores de baixo custo não se mostra suficientemente escalável em determinadas situações [Silva et al. 2014]. Por exem-

plo, para o monitoramento de uma região metropolitana seria necessário uma grande quantidade de sensores. Nesse sentido, uma outra abordagem de sensoriamento é o *mobile crowdsensing*, onde os dispositivos móveis dos próprios indivíduos (e.g., *smartphones*) são utilizados para realizar o sensoriamento. Esse tipo de abordagem tem ganhado destaque com a queda nos preços dos dispositivos móveis e a popularização das redes sociais baseadas em localização, como Twitter[1], Instagram[2], e Foursquare[3] [Silva et al. 2014, Frias-Martinez et al. 2012]. Nessas abordagens, os dispositivos móveis dos usuários dos serviços urbanos agem como sensores e integram a arquitetura de camadas de sensoriamento da cidade. O papel de dados de mídias sociais, como os *tweets*, tem crescido significativamente em análise de trânsito por causa da sua alta disponibilidade, podendo ser coletada em muitos lugares no mundo em tempo real (ou quase real). Vários trabalhos utilizam dados de mídias sociais para monitorar, analisar e entender diferentes fenômenos urbanos. Um exemplo de uso de dados de mídias sociais é auxiliar no processo de compreensão do comportamento de dados extraídos de outras fontes. No presente trabalho, esses dados são utilizados para identificar a presença de aglomerados de pessoas na cidade de Nova Iorque e investigar se essa presença tem efeito sobre o sistema de táxis.

Diferentes abordagens foram utilizadas para analisar essas fontes de dados de forma independente e têm levado a resultados interessantes, como discutido neste trabalho. Todavia, o uso de combinações dessas fontes de dados a fim de compreender melhor as dinâmicas dos centros urbanos ainda é recente. Dessa forma, no presente trabalho utilizamos o arcabouço SMAFramework [Rodrigues et al. 2017] para analisar dados de mídias sociais e dados de viagens de táxi da cidade de Nova Iorque a fim de propor uma metodologia de descoberta de pontos para posicionamento de táxis e veículos de serviços semelhantes, como Uber e Cabify. Este trabalho foca em dados provenientes de redes sociais baseadas em localização, mais especificamente o Twitter, que permite o compartilhamento de mensagens curtas com anotações espaço-temporais – dados que permitem a identificação de local e momento do seu compartilhamento na rede. Os dados do Twitter são utilizados como um sensor virtual de presença humana em determinada localização, de modo que a existência de uma maior quantidade de interações sociais pode indicar a presença de um maior número de pessoas naquele local e, consequentemente, um maior número de indivíduos que possam utilizar os serviços de táxi e outros na mesma linha. Já os dados de viagens de táxi da cidade de Nova Iorque são utilizados para validar a metodologia proposta em um cenário do mundo real.

O presente trabalho tem como objetivo propor uma técnica para melhorar o posicionamento dos táxis e outros veículos de serviços semelhantes nas cidades utilizando dados de mídias sociais. Análises semelhantes já foram realizadas em determinadas cidades e utilizando os dados gerados pelo próprio serviço de táxi [Commission 2017]. Para replicar tais metodologias seria necessário uma adaptação da infraestrutura usada pelos veículos atuais. Além disso, os dados gerados pelo próprio serviço são limitados, uma vez que apenas representam usuários que conseguiram acessar o serviço. No presente trabalho propomos uma metodologia que pode ser aplicada em diferentes cidades sem adaptações de infraestrutura, uma vez que os dados de mídias sociais já estão disponíveis em volume significativo para um grande número de cidades. Além disso, esses dados não possuem a

[1]https://twitter.com/

[2]https://www.instagram.com/

[3]https://foursquare.com/

limitação de representatividade de indíviduos que já acessaram o serviço de transporte, na verdade há indícios de que as mídias sociais já tenham atingido diversos segmentos da sociedade. Dos 7,5 Bilhões de habitantes na terra, 2,7 Bilhões são usuários ativos de midias sociais em dispositivos móveis (número que fica ainda mais representativo se considerada a urbanização mundial de 54% da população) [Hootsuite 2017].

Este trabalho está organizado como descrito a seguir. A Seção 2 apresenta alguns conceitos relacionados à análise de dados de mobilidade urbana, bem como alguns estudos realizados com dados semelhantes aos usados neste trabalho. A Seção 3 descreve como foi realizado o experimento para validação da hipótese de pesquisa e apresenta as principais características do arcabouço utilizado. Para usar o arcabouço foram desenvolvidas extensões para coleta de dados dos táxis verdes, amarelos e do Twitter. Além disso, foi desenvolvido o algoritmo Fuzzy Matcher, que avalia a relação de amostras espaço-temporais de diferentes fontes de dados. A Seção 4 apresenta os resultados obtidos por meio dos dados da cidade de Nova Iorque. Com esses resultados foi possível verificar que mídias sociais podem ser usadas como fontes de dados complementares no processo de análise de mobilidade urbana. Mais específicamente, foi mostrado o valor desses dados para auxiliar no processo de posicionamento de veículos de táxis e outros semelhantes (e.g., uber). Finalmente, a Seção 5 apresenta as considerações finais.

2. Trabalhos Relacionados

Nos últimos anos, diferentes metodologias têm sido utilizadas no projeto de sistemas de transporte inteligentes, para melhorar o desempenho de transportes urbanos. Uma mudança relevante nessas metodologias surgiu com o uso de análise de dados gerados pelos cidadãos em suas vidas cotidianas, criando assim aplicações para sistemas de transporte inteligentes baseados em análise de dados. Zhang et al. [Zhang et al. 2011] estabelecem uma classificação dessas aplicações em: (i) Visão: que reune várias técnicas que exploram dados capturados de sensores visuais, como câmeras – segundo os autores, do ponto de vista da percepção, os seres humanos são mais familiares com esse tipo de percepção em detrimento de outros; (ii) Diferentes Fontes de Dados: área que contempla as análises realizadas a mais de uma fonte de dados explorando sua complementariedade – os autores mencionam que essa é uma área com muitos problemas em aberto, além disso a maior parte dos trabalhos apresentados no *survey* busca unir as informações em um modelo único sem analisar o significado dos dados de acordo com a forma que foram coletados; e (iii) aprendizado: que trata da observação dos meios de tranporte atuais a fim de aprender e melhor entender seus mecanismos ocultos – vale mencionar que não necessariamente são usadas técnicas de aprendizado de máquina. O presente trabalho pode se enquadrar nas duas últimas categorias citadas por Zhang et al., pois é criado um modelo para extração de conhecimento e entendimento dos mecanismos de transporte a partir de múltiplas fontes de dados.

Na literatura, existem trabalhos que exploram dados gerados a partir de viagens de táxi para modelar o serviço de táxi nos centros urbanos. Por exemplo, Zhang et al. [Zhang et al. 2015] exploram traces de viagens de táxi gerados na China a fim de melhor compreender as estratégias usadas pelos motoristas nos diferentes momentos. Os autores classificam esses instantes como: (i) procura por passageiros, que diz respeito aos padrões de circulação do motorista ao esperar por novos passageiros; (ii) entrega de passageiros, que consiste na escolha de rotas para levar o passageiro ao seu destino e na observação

de condições de trânsito e valor final da corrida; e (iii) escolha de área de preferência, já que em determinados momentos do dia os taxistas tendem a escolher determinadas zonas para trabalharem, como aquelas mais familiares ou com mais possibilidades de contratação. Ao criarem essa divisão, os autores mapeiam as diferentes estratégias usadas pelos taxistas com a renda por eles obtida. Esse mapeamento serve para classificar as estratégias em eficientes ou ineficientes.

Boa parte dos estudos sobre mobilidade urbana e, mais especificamente táxis, é focada na cidade de Nova Iorque, principalmente por causa da política de liberação dos dados de viagens. Um desses trabalhos, desenvolvido por Dimitriou et al. [Dimitriou et al. 2016], estuda a distribuição de inícios e fins de viagens na cidade e as relaciona com parâmetros com a duração e distância delas. Esse estudo objetiva encontrar zonas com altas probabilidades para táxistas encontrarem clientes. O estudo é limitado, pois apenas conta com os dados dos próprios táxis, de modo que apenas representa as pessoas que conseguiram contratar o serviço, mas não as que precisaram dele. Além disso, enviar todos os táxis para os mesmos locais de maior necessidade de seus serviços nem sempre é a melhor solução, já que essa estratégia pode ocasionar um aumento na oferta de serviços nessas regiões além da demanda existente.

Bialik et al. [Bialik et al. 2015] realizaram outro trabalho com dados da cidade de Nova Iorque, porém com dados do Uber[4]. O trabalho mostra que carros desse aplicativo servem mais as zonas periféricas da cidade de Nova Iorque que os táxis. Apesar de na data do estudo os táxis ainda possuírem um maior número total de viagens na cidade, esse é um estudo inicial que aponta a existência de uma demanda que não é suprida pelo atual posicionamento do sistema de táxis. Em nosso trabalho, por meio de fontes de dados complementares, buscamos modelar melhor essa demanda em regiões mais afastadas do centro. Dadas as restrições para obtenção de dados tanto do Twitter quanto do Uber, não foi possível realizar um estudo que compare essas duas fontes de dados com a abordagem de análise de correlação utilizada neste trabalho e apresentada na Seção 3.

3. Metodologia

Este trabalho propõe o uso de dados de mídias sociais para melhorar o posicionamento de veículos de táxis e outros semelhantes, como Uber. Esses dados são utilizados como um sensor virtual da presença de pessoas em determinadas regiões da cidade e, consequentemente, da existência de uma maior quantidade de candidatos a contratarem serviços de táxi. Assim, estabelece-se uma hipótese de que dados baseados em localização de interações em mídias sociais podem ser utilizados para melhorar o posicionamento de táxis e veículos de serviços semelhantes. Para validar essa hipótese foi utilizado o arcabouço de análise de dados de mobilidade urbana chamado SMAFramework [Rodrigues et al. 2017]. Esse arcabouço foi selecionado, pois permite a análise de diferentes fontes de dados de mobilidade urbana e suas complementaridades. Além disso, foram usados dados reais coletados na cidade de Nova Iorque.

A cidade de Nova Iorque possuí dois sistemas de táxis: os amarelos e os verdes. Os táxis amarelos podem circular livremente por toda a cidade para realizar suas viagens e devem ser contratados diretamente na rua, isto é, os usuários devem ver o táxi, sinalizar para o motorista e contratá-los. Na cidade de Nova Iorque, foi possível perceber que, com

[4]https://www.uber.com/

o tempo, os táxis amarelos estavam servindo a zonas limitadas da cidade: no centro da ilha de Manhattan e próximas aos aeroportos. Essa constatação foi feita a partir de uma análise de dados de posição de início e fim das viagens de táxi. Para solucionar esse problema a prefeitura da cidade criou os táxis verdes, que seriam proibidos de iniciar viagens nessas zonas, servindo assim zonas mais periféricas, como o Brooklin, Queens e Bronx. Essa divisão em zonas dos dois sistemas de táxi foi utilizada para validar a hipótese de pesquisa desse trabalho, por meio do SMAFramework e o desenvolvimento de uma ferramenta de análise, o FuzzyMatcher – que permite a avaliação da relação existente entre duas amostras de diferentes fontes de dados com base em sua distância espaço-temporal. Com essa ferramenta foi possível criar uma visualização da correlação das viagens de táxi e as interações em mídias sociais.

A análise foi feita com dados de táxi da cidade de Nova Iorque de janeiro de 2016, coletados do portal de dados públicos da cidade[5] e os dados do Twitter foram coletados por meio da API pública[6]. No total foram coletados 398,887 tweets com anotações espaço-temporais e, respectivamente 10,415,045 e 1,141,933 viagens válidas nos táxis amarelos e verdes naquele mês. O algoritmo FuzzyMatcher permitiu realizar a comparação dos dados do Twitter vs. táxis amarelos e Twitter vs. táxis verdes, os resultados obtidos são apresentados na Seção 4. As Seções 3.1 e 3.2 apresentam com mais detalhes o arcabouço utilizado e o algoritmo que permitiu a realização da análise efetuada, respectivamente.

3.1. Arcabouço para Integraçao de Dados de Mobilidade Urbana

O SMAFramework [Rodrigues et al. 2017], utilizado para reconhecimento de padrões nos dados de mobilidade urbana, auxilia na coleta de dados disponíveis em diferentes fontes nas cidades e os padroniza para facilitar o seu gerenciamento e análise. O arcabouço provê uma base comum para execução de tarefas triviais de análise de dados, como limpar dados inválidos, remover dados duplicados e filtrar os dados. Finalmente, o arcabouço proporciona formas de lidar com desafios de análise de dados de mobilidade urbana. Mais especificamente, o mapeamento de dados provenientes de diferentes camadas de sensoriamento e, também, a análise da correlação entre essas camadas a partir de diferentes perspectivas.

A Figura 1 mostra a arquitetura do arcabouço. No topo, é exibida a cidade gerando os dados por meio de diferentes fontes de dados. Cada ícone dentro da nuvem representa uma fonte de dados brutos diferente, que deve ser coletada para análise usando os Coletores de Dados. Esses componentes coletam dados brutos da sua fonte e os salvam em um formato inicial básico chamado de Amostras. Por exemplo, um Coletor de Dados pode usar uma API para realizar a coleta em uma plataforma de rede social e, então, converter esses dados para o formato básico de Amostras. Esses dados também podem ser coletados de formas variadas e não apenas através de APIs. O objetivo dos Coletores de Dados é observar as peculiaridades de cada fonte de informação e realizar o passo inicial de extração. No exemplo mostrado na Figura 1, os dados do Twitter são extraídos por dois Coletores de Dados, que podem ser, por exemplo, um coletor de *stream*, que coleta os dados em tempo real, e um outro coletor para dados obtidos previamente, que estejam arquivados.

[5]https://opendata.cityofnewyork.us/

[6]https://developer.twitter.com/

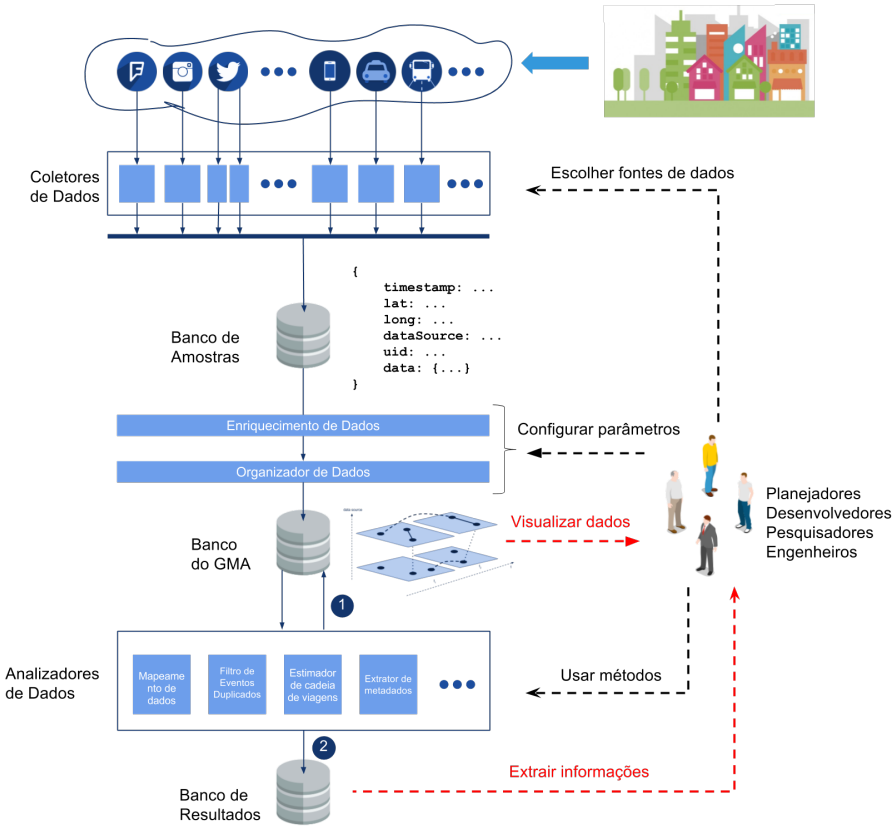

Figura 1. Arquitetura do Arcabouço SMAFramework

Cada amostra contém dados anotados sobre espaço e tempo. Essas amostras consistem em geo-localização (i.e., latitude e longitude), tempo, fonte de dados e UIDs, i.e., identificadores das entidades que geraram os dados. Além disso, essas amostras podem conter dados adicionais que auxiliam eventuais casos de análise. Essas amostras são inicialmente armazenadas em uma base de dados, como mostrado na Figura 1. Uma vez na base de dados, existe um componente para ler as amostras e as organizar na estrutura de um Grafo Multi-Aspecto (GMA) [Kivelä et al. 2014, Wehmuth et al. 2014]. O GMA é a estrutura de dados principal do SMAFramework e é nessa estrutura que os dados são colocados para realização das tarefas de análise ou visualização dos dados. Além da organização dos dados nessa estrutura, também são executadas tarefas de enriquecimento de dados conforme as configurações do usuário. Esse enriquecimento pode inferir novas amostras, que não estavam no conjunto de dados inicial, de acordo com os dados existentes. Por exemplo, Mahrsi et al. [Mahrsi et al. 2016] apresentam duas formas de enriquecer dados extraídos de posições onde passageiros pegam ônibus na cidade. Nesse cenário, é conhecido quando e onde os usuários entram em um ônibus, porém, não se têm dados de quando eles deixam o ônibus. Assim, para inferir mais dados, os autores criaram duas hipóteses: (i) ao trocar de ônibus, é assumido que o usuário desceu na parada mais próxima na linha em que estava da parada na qual ele pegou o próximo ônibus; (ii) na última viagem do dia, os passageiros vão descer do ônibus na parada mais próxima da primeira parada utilizada no início do dia.

Depois do enriquecimento dos dados, eles podem ser estruturados segundo o mo-

delo de GMA. Essa estrutura é armazenada pelo arcabouço na base de dados para o GMA. Nesse ponto, os dados podem ser verificados pelos usuários do arcabouço, por exemplo, usando ferramentas de visualização. Uma vez organizados na estrutura de GMA, é possível executar os Analisadores de Dados para extrair informações de mobilidade. De forma semelhante aos Coletores de Dados, os Analisadores de Dados podem ser convenientemente adicionados para estender o arcabouço e adequá-lo às necessidades do usuário. Os resultados dos procedimentos de análise são salvos na estrutura GMA, ou ainda armazenados na base de resultados, onde os usuários têm acesso a informações na forma de totalizadores, resumos, índices, metadados, mapas de calor, dentre outros.

O arcabouço é organizado em um fluxo dos dados passando por três fases para reduzir o trabalho necessário na adição de novos módulos, i.e., primeiro as Amostras, depois o GMA e, finalmente, os Resultados. Assim, se usuários precisarem usar Analizadores existentes para analisar dados de uma fonte de dados nova, eles apenas terão que desenvolver o Extrator de Dados para isso. O mesmo ocorre caso o usuário use uma fonte de dados já suportada pelo arcabouço, porém com um novo Analizador. Essa abordagem na arquitetura facilita que se criem extensões para o arcabouço, que podem ser usadas para adição de novos módulos ao SMAFramework de modo a mantê-lo atualizado, que é uma importante característica para lidar com análise de dados [Luckow and Kennedy 2017, Biem et al. 2010], um campo onde novas metodologias surgem com uma frequência significativa. Essa divisão em fases também organiza o arcabouço de forma a facilitar a inserção de novas fases no fluxo dos dados entre as fases já existentes para melhoria de fatores como desempenho ou escalabilidade; por exemplo a adição de uma fase de indexação do conteúdo entre a fase do GMA e a análise poderia ser realizada de modo aumentar o desempenho de algum método de análise.

3.2. Fuzzy Matcher

As tarefas de análise disponíveis no arcabouço objetivam auxiliar no processo de entendimento das dinâmicas das cidades pelas pessoas envolvidas no seu planejamento. Por exemplo, cidadãos para melhor entender como utilizar os recursos e serviços disponíveis; ou administradores públicos, ao criar políticas para gerência da cidade. O Analisador de Dados Fuzzy Matcher investiga a correlação espaço-temporal dos dados em diferentes camadas de sensoriamento, analisando a influência/correlação local de uma camada de sensoriamento em outra. Por influência/correlação local é entendido que nós em uma camada devem se localizar perto no espaço e no tempo de nós em outra camada. O propósito dessa análise é verificar se dados de uma camada de sensoriamento podem ser usados para estudar outra. Dessa forma, dados do Twitter foram utilizados para capturar o comportamento dos usuários assumindo que a presença de uma grande quantidade de usuários do Twitter em um local poderia indicar também a existência de pessoas em busca de viagens de táxi. Dessa forma, permitindo o uso dados da camada de sensoriamento do Twitter para estimar a quantidade de pessoas em determinado local e analisar se a sua presença está relacionada a contratação de táxis.

Fuzzy Matcher é um algoritmo que identifica pares espaço-temporais entre nós no GMA de diferentes camadas de sensoriamento de uma cidade. Depois da identificação desses pares, o algoritmo também avalia uma pontuação. Essa pontuação é calculada com base na distância espaço-temporal entre os nós pareados e uma função de dispersão. Para usar esse componente do arcabouço, os usuários devem especificar parâmetros, como

uma precisão em termos de distância no espaço e no tempo, e também uma função de depreciação no espaço e no tempo. Essa função pode ser alterada para casos de uso específicos, onde a dispersão das entidades analisadas pode ser modelada com uma função. Por exemplo, a movimentação de uma multidão de pessoas pode variar de acordo com o cenário no qual o deslocamento ocorre. Um desfile, por exemplo, tende a percorrer uma grande distância na cidade, por outro lado, uma multidão assistindo a um show fica concentrada em um único local. A forma como multidões, e outros fluxos de mobilidade, se comportam na cidade pode ser descrita de diversas formas com essas funções de dispersão.

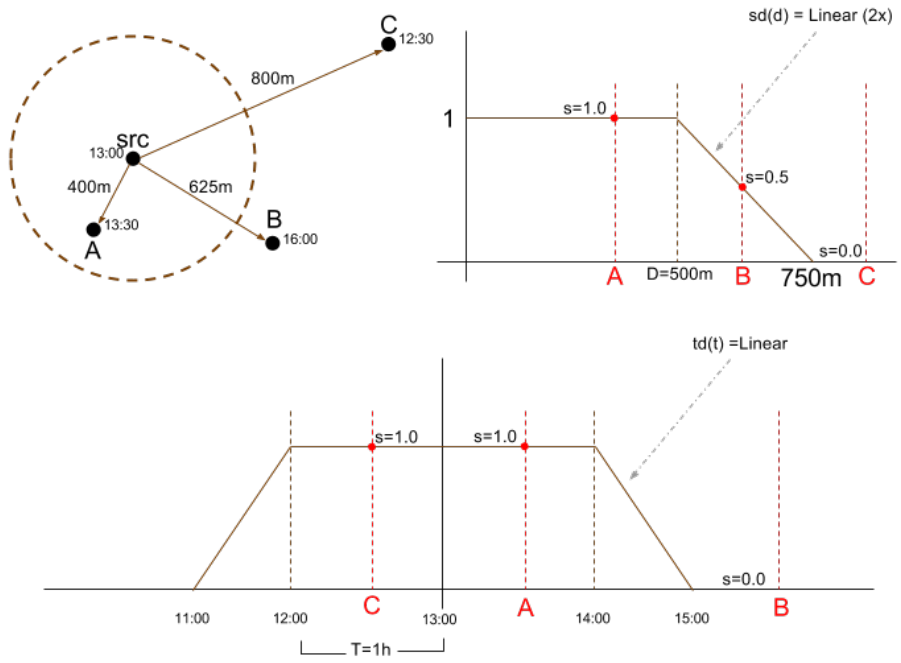

Figura 2. Funcionamento do algoritmo do Fuzzy Matcher

A Figura 2 mostra como o algoritmo Fuzzy Matcher funciona e também como os seus parâmetros são usados. Na figura temos: a precisão de distância D, precisão de tempo T, e as funções de depreciação temporal e espacial $td(t)$ e $sd(d)$, respectivamente, onde d e t são as distâncias temporal e espacial usadas para criar a curva para avaliar a pontuação de cada par. A pontuação espacial é dada por $SS(d) = f(d)/D$ onde $f(d) = \{D$, se $d < D$; $sd(d-D)$, caso contrário$\}$, e de forma similar para a pontuação temporal: $TS(t) = f(t)/T$ onde $f(t) = \{T$, se $t < T$; $st(t-T)$, caso contrário$\}$. Por exemplo, dado o cenário na Figura 2, o par $(\textit{SRC}, B)$ tem uma pontuação espacial $SS(625) = 0.5$, para $D = 500$, e a função de depreciação espacial linear (i.e., $sd(d) = D - d$). No mesmo cenário, como o par $(\textit{SRC}, C)$ tem $SS(800) = 0$, esse par não é considerado pelo Fuzzy Matcher. O mesmo é válido em caso de $TS(t) = 0$.

O Fuzzy Matcher adiciona algumas possibilidades para auxiliar na análise de dados urbanos. Por exemplo, bancos de dados geográficos oferecem uma variedade de ferramentas para trabalhar com dados espaciais, como consultas geo-localizadas, que permitem buscas em regiões específicas; ou até mesmo funções observadoras, que tornam possível observar mudanças na base de dados que ocorrem em determinada área. Toda-

via, muitas dessas abordagens não contabilizam a dimensão temporal. Além disso, além de apenas parear amostras, o Fuzzy Matcher proporciona pontuações que representam o quão forte/fraco é um dado par. Ao usar essas pontuações, limites podem ser definidos de forma a classificar os pares, por exemplo. Finalmente, o Fuzzy Matcher introduz uma forma de usar diferentes funções de depreciação para analisar conjuntos de dados que possuem diferentes comportamentos de dispersão.

4. Resultados

O objetivo desse experimento é identificar padrões na quantidade de viagens de táxi contratadas e a quantidade de *tweets* em uma área; analisar a correlação no espaço e tempo entre esses conjuntos de dados. Usuários do Twitter, identificados pelos seus *tweets*, agem como um sensor de aglomerações de pessoas – muitas pessoas compartilhando *tweets* indicam a existência de ainda mais pessoas em uma área. Essas aglomerações de pessoas podem conter candidatos a usar os serviços de táxi. Assim, é realizado um estudo para identificar regiões onde dados do Twitter sugerem que viagens de táxi poderiam ser estimuladas.

4.1. Condições do Experimento

Para analisar a correlação desses conjuntos de dados, foi definida uma área de interesse que cobre Manhattan e algumas regiões vizinhas, e foram coletados dados conforme descrito na Seção 3. O experimento consiste em usar as pontuações do algoritmo Fuzzy Matcher para verificar a correlação de interações dos usuários do Twitter e a contratação de viagens de táxis em determinadas zonas da cidade. Nesse experimento, é analisado se dados do Twitter podem ser utilizados para melhorar o posicionamento de táxis na cidade, de modo que eles fiquem mais acessíveis aos cidadãos, além de aumentar o número de solicitações de viagens. A hipótese de que uma região com um número relevante de *tweets* indica uma quantidade significativa de possíveis passageiros de táxi foi testada. Se comprovada, dados do Twitter, coletados em tempo quase real, podem ser usados para indicar regiões onde táxis podem se posicionar para servir melhor aos cidadãos. Esses dados ainda podem ajudar a reduzir o tempo necessário para identificar uma mudança nos fluxos de passageiros da cidade causada por eventos não esperados, como manifestações ou congestionamentos.

4.2. Resultados e Análise

O experimento foi dividido em duas fases. Em um experimento inicial, foi usado o algoritmo Fuzzy Matcher para analisar a correlação dos dados coletados do Twitter e dos táxis amarelos. Depois desse experimento inicial, foi executado um segundo similar ao primeiro, porém comparando os dados do Twitter e dos táxis verdes. O algoritmo identificou pares espaço-temporais com uma precisão de distância de 100 metros e uma precisão temporal de 2 horas. Os dados do Twitter foram usados para criar o mapa de calor mostrado na Figura 3 I. Além disso, as pontuações dos pares entre os conjuntos de dados foram calculadas e usadas para construir os mapas de calor com os táxis amarelos (Figura 3 II) e com os táxis verdes (Figura 3 III). É importante destacar que os mapas de calor nas partes II e III da Figura 3 refletem a pontuação encontrada pelo Fuzzy Matcher, e não apenas as quantidades como na parte I. Ademais, a correlação avaliada pelo algoritmo não exclui a possibilidade de existir um único dado do Twitter, por exemplo, sendo pareado

com diferentes amostras do conjunto de dados dos táxis. Esses dois fatores podem levar a maiores valores nos mapas de calor do Fuzzy Matcher do que no mapa do Twitter, o que é esperado. Finalmente, mas não menos importante, o mapa de calor criado com o algoritmo Fuzzy Matcher considera a variação temporal em suas análises, i.e., as pontuações sobre as regiões no mapa de calor apenas aumentam se nessa região existirem amostras próximas no tempo e no espaço, em detrimento de uma avaliação que não concebe o fator temporal.

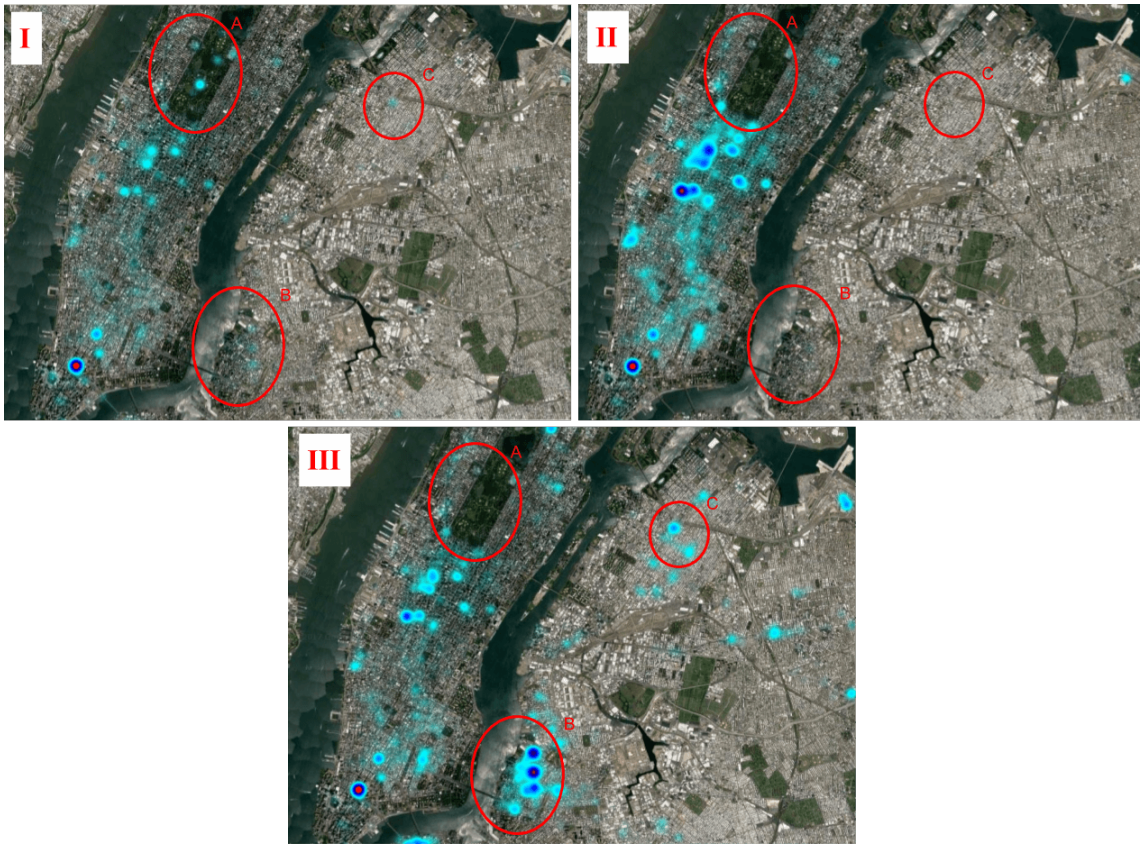

Figura 3. Mapas de calor da distribuição de *twwets* (I), e pares entre dados do Twitter e os táxis amarelos (II) e os táxis verdes (III).

Observando os mapas de calor da Figura 3 I e II, é possível identificar que regiões com grandes volumes de *tweets* resultaram em regiões com altos volumes de pares no centro da Ilha de Manhattan. Essa informação por si só não é suficiente para provar a hipótese inicial de que a presença de usuários do Twitter poderia indicar a necessidade de viagens de táxi. Após observar todo o mapa de calor, existem algumas regiões que devem ser destacadas, identificadas pelos círculos A, B e C na Figura 3. Nessas regiões, existe uma quantidade significativa de *tweets* na parte I, porém, as pontuações dos pares encontrados não são relevantes na parte II. A primeira região a ser observada é a delimitada pelo circulo A. Esse circulo está sobre o Central Park, assim táxis não podem acessar essa área, justificando a ausência de pares. Nos outros dois círculos, é possível notar uma pequena quantidade de *tweets* (em I) que não são refletidos em pares (em II). Uma informação importante a ser conhecida nesse ponto é que os táxis amarelos possuem uma boa cobertura na região central de Manhattan, todavia de acordo com os dados da Figura 3 I e II a cobertura geral da cidade poderia ser melhorada ao enviar táxis a outras regiões,

como os círculos B e C; o que poderia aumentar o número de viagens e a qualidade do sistema de transporte da cidade.

Para investigar mais a possibilidade de melhorar o sistema de transporte enviando táxis a outras regiões da cidade baseado nos dados coletados do Twitter, um segundo experimento foi executado com os dados dos táxis verdes. Historicamente em Nova Iorque, a administração da cidade percebeu que os táxis amarelos não estavam servindo igualmente as zonas da cidade fora de Manhattan, assim um segundo serviço de táxi foi criado, os táxis verdes. Esses táxis não são permitidos circular nas regiões centrais de Manhattan. A expetativa no segundo experimento era de que poderia-se provar os resultados do experimento inicial combinando os dados com esse terceiro conjunto de dados. Assim, se um grande volume de pares fossem identificados nas regiões dos círculos B e C, e outras áreas similares, isso significa que de fato enviar táxis para essas zonas resulta no aumento de contratações de viagens. Ao observar os círculos B e C na Figura 3 III, é possível notar o aumento de pares, o que é um forte indício de comprovação da hipótese inicial. Nesse caso, as conclusões extraídas dos conjuntos de dados já haviam sido percebidas pela administração da cidade, o que foi importante para validar o trabalho proposto. Todavia, existem várias outras cidades onde esse entendimento do sistema de transporte ainda não foi obtido por outros meios e os dados do Twitter poderiam ser utilizados para tal fim. Além disso, mesmo na cidade de Nova Iorque, dados da camada de sensoriamento do Twitter, obtidos em tempo quase-real, poderiam ser utilizados para monitorar mudanças rápidas nos fluxos de mobilidade da cidade que possam resultar no surgimento de novos pontos com alta probabilidade de contratação de viagens de táxi.

5. Considerações Finais

Soluções de Computação Urbana têm tornado possível melhorar os serviços prestados aos cidadãos. Por meio da análise de dados, por exemplo, é possível obter uma melhor compreensão das dinâmicas urbanas a fim de modelá-las e estudar a fundo quais as melhores soluções a serem aplicadas a fim de mitigar os problemas da cidade como, por exemplo, a mobilidade, contexto que se insere o presente trabalho. Com o estudo de dados de mídias sociais e dados disponibilizados por órgãos públicos, foi possível verificar uma hipótese que pode conduzir a uma melhora no posicionamento dos táxis e veículos de serviços semelhantes, melhorando o sistema de trasnporte da cidade e provendo serviços mais eficientes aos cidadãos. Para validar a hipótese apresentada foram desenvolvidas algumas ferramentas para estender o SMAFramework, como os extratores de dados do Twitter e dos táxis amarelos e verdes, bem como a ferramenta de análise de correlação espaço-temporal Fuzzy Matcher.

Além das contribuições com o desenvolvimento de ferramentas de análise de mobilidade urbana para o arcabouço, a principal contribuição do trabalho é a constatação de que existe valor em usar dados de interações de usuários em mídias sociais baseadas em localização como sensores virtuais de aglomerados de pessoas na cidade. Esse tipo de sensor poderia ser usado também para auxiliar a alocação de outros recursos que servissem a grandes números de cidadãos, e não apenas na alocação de veículos. No futuro, o uso de dados de mídias sociais pode ser levado para outras cidades a fim de melhorar o posicionamento dos táxis e outros serviços. Além disso, essa mesma metodologia, porém com uma maior quantidade de dados (provenientes até mesmo de outras fontes, não apenas o Twitter) poderia ser utilizada para melhorar o posicionamento de táxis mesmo em

Nova Iorque. A análise realizada na cidade para criação do serviço dos táxis verdes, que aumentaria a oferta de táxis nas regiões fora de Manhattan, foi feita com dados dos próprios táxis amarelos. Ou seja, o dados utilizados na análise representa as informações de pessoas que conseguiram contratar seus serviços. Ao usar dados de mídias sociais a análise ainda é limitada a representar pessoas que têm acesso a esse tipo de recurso, entretanto existe uma grande tendência desses meios de penetrar em diversos segmentos da sociedade.

Agradecimentos

Este trabalho foi parcialmente apoiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo 403260/2016-7 e 401802/2016-7; e pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), processo 2015/07538-1 e 2015/24494-8.

Referências

Bialik, C., Flowers, A., Fischer-Baum, R., and Mehta, D. (2015). Uber Is Serving New York's Outer Boroughs More Than Taxis Are.

Biem, A., Bouillet, E., Feng, H., Ranganathan, A., Riabov, A., Verscheure, O., Koutsopoulos, H., and Moran, C. (2010). IBM Infosphere Streams for Scalable, Real-time, Intelligent Transportation Services. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 1093–1104, New York, NY, USA. ACM.

Commission, N. T. &. L. (2017). Background on the Boro Taxi program.

Dimitriou, L., Kourti, E., Christodoulou, C., and Gkania, V. (2016). Dynamic Estimation of Optimal Dispatching Locations for Taxi Services in Mega-Cities based on Detailed GPS Information. *IFAC-PapersOnLine*, 49(3):197–202.

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

Hall, R. E., Bowerman, B., Braverman, J., Taylor, J., and Todosow, H. (2000). The vision of a smart city. *2nd International Life . . .* , page 7.

Hootsuite (2017). Global Statshot Digital in Q3 2017. Technical Report Q3-2017.

Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of Complex Networks*, 2(3):203–271.

Loughnan, S., Kuppens, P., Allik, J., Balazs, K., de Lemus, S., Dumont, K., Gargurevich, R., Hidegkuti, I., Leidner, B., Matos, L., Park, J., Realo, A., Shi, J., Sojo, V. E., yue Tong, Y., Vaes, J., Verduyn, P., Yeung, V., and Haslam, N. (2011). Economic inequality is linked to biased self-perception. *Psychological Science*, 22(10):1254–1258. PMID: 21948855.

Luckow, A. and Kennedy, K. (2017). Chapter 5 – Data Infrastructure for Intelligent Transportation Systems. In *Data Analytics for Intelligent Transportation Systems*, pages 113–129.

Mahrsi, M. K. E., Côme, E., Oukhellou, L., and Verleysen, M. (2016). Clustering Smart Card Data for Urban Mobility Analysis. pages 1–17.

Nam, T. and Pardo, T. A. (2011). Conceptualizing smart city with dimensions of technology, people, and institutions. In *Proceedings of the 12th Annual International Digital Government Research Conference on Digital Government Innovation in Challenging Times - dg.o '11*, page 282, New York, New York, USA. ACM Press.

Pellicer, S., Santa, G., Bleda, A. L., Maestre, R., Jara, A. J., and Skarmeta, A. G. (2013). A global perspective of smart cities: A survey. *Proceedings - 7th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, IMIS 2013*, pages 439–444.

Rodrigues, D. O., Boukerche, A., Silva, T. H., Loureiro, A. A. F., and Villas, L. A. (2017). SMAFramework: Urban Data Integration Framework for Mobility Analysis in Smart Cities. In *Proceedings of the 20th ACM International Conference on Modelling, Analysis and Simulation of Wireless and Mobile Systems*, MSWiM '17, pages 227–236, New York, NY, USA. ACM.

Silva, T., Vaz De Melo, P., Almeida, J., and Loureiro, A. (2014). Large-scale study of city dynamics and urban social behavior using participatory sensing. *Wireless Communications, IEEE*, 21(1):42–51.

Veenhoven, R. (1987). Cultural bias in ratings of perceived life quality: A comment on ostroot and snijder. *Social Indicators Research*, 19(3):329–334.

Wehmuth, K., Ziviani, A., and Fleury, E. (2014). A Unifying Model for Representing Time-Varying Graphs. *Computing Research Repository arXiv.org*, I(January):1–28.

Zhang, D., Sun, L., Li, B., Chen, C., Pan, G., Li, S., and Wu, Z. (2015). Understanding taxi service strategies from taxi GPS traces. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):123–135.

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., and Chen, C. (2011). Data-Driven Intelligent Transportation Systems: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 12(4):1624–1639.

Zheng, Y., Capra, L., Wolfson, O., and Yang, H. (2014). Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38.