MODELOS DE INFERÊNCIA DE LOCALIZAÇÃO DE RESIDÊNCIA PARA USUÁRIOS DO FOURSQUARE

MICHELLE HANNE SOARES ANDRADE

MODELOS DE INFERÊNCIA DE LOCALIZAÇÃO DE RESIDÊNCIA PARA USUÁRIOS DO FOURSQUARE

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: JUSSARA MARQUES DE ALMEIDA GONÇALVES

Belo Horizonte

Setembro de 2016

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Andrade, Michelle Hanne Soares de.

A553m Modelos de inferência de localização de residência para usuários do foursquare . / Michelle Hanne Soares de Andrade. – Belo Horizonte, 2016.
xxiv, 83 f.: il.; 29 cm.

Dissertações (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientadora: Jussara Marques de Almeida Gonçalves.

1. Computação - Teses. 2. Redes sociais on-line. 3. Probabilidades. 4. Foursquare. 5. Direito a privacidade. I. Orientadora. II. Título.

CDU 519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Modelos de inferência de localização de residência para usuários do foursquare

MICHELLE HANNE SOARES DE ANDRADE

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROFA. JUSSARA MARQUES DE ALMEIDA GONÇALVES - Orientadora
Departamento de Ciência da Computação - UFMG

PROF. CLODOVEU AUGUSTO DAVIS JÚNIOR
Departamento de Ciência da Computação - UFMG

PROF. MARCOS ANDRÉ GONÇALVES
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 26 de setembro de 2016.

Ao meu novo amor, Vinícius Andrade Teixeira.

Agradecimentos

Agradeço a todos aqueles que me auxiliaram nesta trajetória do mestrado. Em especial a Professora Orientadora Jussara Almeida, por conduzir este trabalho de forma justa e imparcial, porém, com leveza me orientou com dedicação e empenho. Ao Professor Marcos Gonçalves, responsável por alicerçar métodos e consolidar ideias.

É com enorme satisfação que agradeço a todos os colegas dos laboratórios, VOD e CAMPS, os quais me auxiliaram em tarefas cotidianas, entre conversas, troca de ideias e boas risadas.

Agradeço aos meus familiares, pelo apoio e compreensão neste últimos anos. Especialmente aos meus pais, ao meu esposo e minhas irmãs. Fugindo do habitual, este percurso teve que ser interrompido em alguns meses, para eu dar à luz um filho, um lindo menino de nome Vinícius.

E essa nova vida trouxe entusiasmo, para a finalização de mais este trabalho. Desafio vencido, que venham os outros.

"Há momentos, e você chega a esses momentos, em que de repente o tempo pára e acontece a eternidade."

(Fiódor Dostoiévski - Escritor Russo)

Resumo

Vivemos uma época em que tudo o que acontece está nas redes sociais *online*, sejam eventos cotidianos ou de grandes proporções. Além do compartilhar e curtir, a tendência agora é estar conectado e geolocalizado. Por exemplo, o Foursquare, rede social *online* de compartilhamento de localização, possui atualmente mais de 55 milhões de usuários e mais de 65 milhões de *venues*, totalizando mais de 8 bilhões de *check-ins*[1]. A crescente exposição de dados geolocalizados pode trazer benefícios, como suporte a serviços de buscas personalizadas e sistemas de recomendação. Esses dados também podem ser utilizados negativamente, por exemplo para *marketing* viral e sistemas que buscam inferir a residência do usuário visando, utilizar essa informação para ações legais e ilegais. Este trabalho propõe diferentes modelos de inferência de localização de residência de usuários a partir de dados públicos compartilhados no Foursquare. Foram propostos diferentes modelos de inferência, que exploram as coordenadas geográficas obtidas de dados públicos dos usuários. Foram analisados mais de 7 milhões de usuários e foi concluído que a união de dados públicos pode melhorar a cobertura e acurácia de inferência, especialmente quanto à granularidade da cidade de residência do usuário. A avaliação experimental realizada mostrou que, em comparação com os modelos de referência, os propostos atingem melhores resultados, com aumento de 21% da acurácia, comparados com o mesmo número de cobertura de usuários. Em particular, um dos modelos propostos, que explora a combinação de vários classificadores (modelo Híbrido), inferiu corretamente a cidade de residência de mais de 5 milhões de usuários e obteve uma acurácia de 70,04%. Em granularidades mais finas como bairro, foi possível obter uma acurácia de até 69,73%, considerando a inferência dentro do bairro do usuário. Na granularidade de coordenadas geográficas, obtivemos uma acurácia de 67,12%, considerando a localização exata de residência do usuário em até 5 *km* de distância.

Palavras-chave: inferência, privacidade, foursquare, redes sociais online, residência.

[1] http://pt.foursquare.com/about acesso em Julho de 2016

Abstract

We live in a time when everything that happens on online social networks, are daily or major events. Besides the share and like, the trend is now to be connected and geolocated. For example, Foursquare, online social network location sharing currently has over 55 million users and over 65 million venues, more than 8 billion check-ins[2]. The growing exposure of geolocated data can bring benefits like support services custom searches and recommendation systems. These data also can be used negatively, e.g., viral marketing and systems that seek to infer the home location for illegal actions. This work proposes different models of inference home location users from public data shared Foursquare. We propose different models of inference that exploit the geographical coordinates obtained from public data users. We analysed more than 7 million users and concluded that public data binding can improve the coverage and accuracy of inference, especially the city of granularity. Experimental evaluation showed that compared with baseline models and found that they attain better results, with an increase of 21% accuracy, with the same user coverage. Particularly, one of models proposed exploiting the combination of multiple classifiers (Hybrid model) correctly inferred the home location of over 5 million users and obtained 70.04% of accuracy. The granularity of neighborhood, it was possible to accuracy of up to 69.73%, considering the inference within the user's neighborhood. The granularity of geographic coordinates, obtained an accuracy of 67.12%, considering the exact home location within a 5 km away.

Keywords: inference, privacy, foursquare, social networks, home location.

[2]http://pt.foursquare.com/about acesso em Julho de 2016

Lista de Figuras

Lista de Tabelas

Lista de Algoritmos

Sumário

Capítulo 1

Introdução

Os serviços colaborativos envolvendo os conceitos de *wikis* e tecnologia da informação são responsáveis por grande parte do conteúdo atualmente gerado na Internet. Nesse cenário, destacam-se as chamadas redes sociais *online*, cujo principal objetivo é a propagação de informações. O recente aumento da popularidade dos dispositivos móveis, como *smartphones* e *tablets*, contribuiu sobremaneira para a disseminação das redes sociais *online*, e incentivou o surgimento e a popularização dos serviços baseados em geolocalização.

O Facebook [2004] possui atualmente mais de 1.13 bilhão de usuários ativos por dia em média[1]. O microblog Twitter [2006] possui 313 milhões de usuários ativos mensais[2]. Já o YouTube [2006] possui mais de um bilhão de usuários e milhões de horas de vídeo são assistidas diariamente, gerando bilhões de visualizações [3]. A Wikipedia [2006] em inglês possui 5.212.999 artigos e uma média de 800 novos por dia, além de 28.818.166 usuários[4].

O Foursquare [Foursquare 2009] é uma das maiores redes sociais *online* geolocalizadas (*Location-based social networks* ou LBSNs). Atualmente, o Foursquare tornou-se apenas um sistema de recomendação e transferiu para o aplicativo móvel denominado *Swarm* o serviço de marcação de lugares (*check-in*). O acesso ao Forsquare ocorre via *site* ou aparelho móvel, enquanto o *Swarm* está acessível somente através de dispositivos como *smartphones*, *tablets* ou similares. Em ambos, o usuário habilita a checagem e identificação do local onde está via GPS *(Global Positioning System)*, de acordo com sua localização física. No *Swarm*, ao marcar e confirmar o local, o usuário realiza um *check-in* e pode compartilhar sua localização com outras pessoas, via Facebook

[1] https://newsroom.fb.com/company-info/ acesso em Julho de 2016
[2] https://about.twitter.com/company acesso em Julho de 2016
[3] http://www.youtube.com/yt/press/statistics.html acesso em Julho de 2016
[4] http://en.wikipedia.org/wiki/Wikipedia:Statistics acesso em Julho de 2016

ou Twitter. Os *check-ins* são realizados em locais especiais, chamados *venues*, que representam lugares físicos, tais como universidades, monumentos, empresas e marcas comerciais. Esta dissertação é baseada em dados coletados antes da criação do *Swarm*, e por isso considera que todos os atributos aqui abordados são originados da rede social *online* Foursquare.

As políticas de privacidade desse tipo de serviço nem sempre são explícitas, possibilitando que o usuário publique de forma automática os locais por onde está passando ou receba informações sem o seu consentimento [Doty & Wilde 2010]. Um dos propósitos de utilizar marcações de lugares visitados é a gamificação, a qual favorece o engajamento dos usuários e motiva comportamentos particulares por meio do uso de elementos de jogos [Fitz-Walter et al. 2011].

Na atualidade, as principais redes sociais *online* utilizam a marcação de lugares físicos por vontade própria dos usuários, o que permite e/ou facilita a inferência de localização. Paralelamente à exposição do usuário a situações de risco à privacidade, a inferência de localização do usuário pode ser bastante útil para sistemas de recomendação de lugares e eventos, *marketing* digital, sistemas de monitoramento, como alertas de trânsito e epidemias, entre outras funcionalidades. Neste contexto, torna-se relevante uma investigação sobre até que ponto é possível inferir a localização do usuário (por exemplo, a cidade onde o usuário reside), utilizando apenas dados públicos, obtidos do uso de uma rede social *online* georreferenciada.

1.1 Motivação

Atualmente, o grande volume de dados públicos expostos diariamente na Internet são utilizados, por exemplo, para alimentar buscas cada vez mais personalizadas e sistemas de recomendação dos mais variados tipos, o que pode ocasionar questionamentos sobre a privacidade dos usuários.

A privacidade está relacionada ao controle sobre a exposição de informações, as quais são exploradas pelos que desejam ter acesso a um conteúdo teoricamente protegido. Neste contexto, destacam-se as redes sociais *online* georreferenciadas, por apresentarem ao usuário a possibilidade de publicação da marcação física do local onde se encontra. Permitindo, entanto, que se conheça seus hábitos.

Com a exposição geográfica dos locais frequentados diariamente, as redes sociais ganham status de vitrine, onde o usuário posta a sua localização ao fazer um *check-in* e pode receber milhares de *likes* e comentários a respeito do local. Por isso, inferir características do usuário é bastante útil para as empresas, que buscam personalizar

serviços, mantendo e fidelizando o cliente. Atualmente, o Foursquare possui o registro de mais de 65 milhões de *venues* ao redor do mundo[5].

Uma pesquisa com usuários do Foursquare, realizada em 2011, mostrou que 58% dos usuários possuíam "amigos"que não conheciam pessoalmente e 66% conheceram novas pessoas através do Foursquare, revelando que a maioria dos usuários não tem objeções quanto a compartilhar informações com desconhecidos [Lindqvist et al. 2011]. Quando o usuário realiza um *check-in*, as informações do local (*venue*), incluindo coordenadas e localização no mapa, são exibidas para os amigos e também podem ser exibidas para terceiros, pois é possível compartilhar essa informação em outras redes sociais, como o Facebook. Nesse caso, o usuário está sujeito a ter sua privacidade violada, já que informações de localização podem revelar gostos, comportamento e até mesmo a rotina do usuário.

Nesse contexto, a principal motivação para este trabalho é investigar até que ponto dados públicos, provenientes de redes sociais *online*, notadamente o Foursquare, podem ser usados para inferir o local onde o usuário reside. a inferência do local de residência de um usuário pode ser utilizada em sistemas que exploram a recomendação personalizada, por exemplo, a busca de um restaurante para jantar. Além de apontar para uma possível vulnerabilidade da privacidade. Desse modo, torna-se importante uma investigação sobre métodos de inferência da localização de residência de usuários. Tais métodos poderão contribuir para a criação de ferramentas, tanto para sistemas de recomendação direcionados para o perfil do usuário, seja no resultado de uma busca ou na indicação sugestiva, quanto para a proteção da informação. Uma aplicação atual que utilizou a inferência de localização de residência do usuário foi o monitoramento das eleições municipais do Brasil em 2012, realizado pelo Observatório da Web[6] [Filho et al. 2015].

1.2 Objetivos

O objetivo principal deste trabalho é a investigação de modelos de inferência da localização de residência de usuários, utilizando dados públicos coletados da rede social *online* Foursquare. O propósito é desenvolver modelos que sejam precisos, mas também que garantam uma boa cobertura, em termos da fração de usuários elegíveis para o modelo de inferência. Essa fração é função dos tipos de dados explorados pelo modelo de inferência, uma vez que nem todos os usuários compartilham todos os tipos de dados no Foursquare. Pretende-se explorar dados de *venues* (locais existentes fisi-

[5]https://pt.foursquare.com/about
[6]http://www.observatorio.inweb.org.br

camente, como restaurantes, universidades, shoppings entre outros), *tips* (comentários relacionados aos *venues*), *likes* (confirmações de 'eu gosto', direcionadas aos *venues*), *mayorships* (premiações dadas a usuários que realizam muitos *check-ins* em um dado *venue*) e lista de amigos.

1.3 Contribuições

As contribuições principais desta dissertação são:

1. Modelos de inferência da localização de residência dos usuários.

   São propostas diversas estratégias visando aumentar o número de acertos de inferência da cidade de residência dos usuários.

   (i) Primeiramente, estudamos os métodos de referência, especificamente as soluções propostas por Pontes [2013] e propomos estratégias que melhoram os resultados da acurácia de inferência das cidades onde os usuários residem.

   (ii) Propomos também um modelo que considera o mapeamento do histórico de *venues* dos *mayorships, tips e likes* para cada usuário da rede social *online* Foursquare, incluindo a sua lista de amigos. Para isso, foram mapeadas todas as coordenadas geográficas de movimentação do usuário e de seus amigos, gerando um área de mobilidade.

   (iii) Propomos um modelo híbrido que combina diferentes técnicas, uma alternativa para explorar os melhores métodos de inferência, tendo em vista melhorar tanto a cobertura dos dados quanto a acurácia das inferências.

2. Avaliamos modelos para inferência da residência do usuário, na granularidade de cidade e bairro, além da sua localização exata, baseada em coordenadas geográficas.

1.4 Organização do Trabalho

Além deste capítulo introdutório, esta dissertação está dividida em 5 partes. O Capítulo 2 apresenta a revisão da literatura acerca do tema geolocalização, inferência em redes sociais *online* e privacidade. O Capítulo 3 aborda a contextualização do Foursquare e apresenta a coleção de dados usada. O Capítulo 4 define o problema de inferência tratado nesta dissertação e discute os métodos de referência e os novos modelos propostos para a inferência de localização de residência de usuário. Já o Capítulo 5 apresenta

a avaliação metodológica e mostra os resultados obtidos na análise experimental dos modelos propostos. O Capítulo 6 encerra esta dissertação com as conclusões e possíveis direções para trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Para alguns grupos de pessoas e empresas, as redes sociais *online* são o lugar certo para a exposição desenfreada, cujo propósito é alcançar o maior número possível de *check-ins*, *likes*, amigos, comentários, compartilhamentos, ou seja, a busca da visibilidade a todo custo. Cada vez mais comum, a divulgação da geolocalização é algo que reforça o conteúdo da mensagem, ganhando mais visibilidade com as LBSNs. A Seção 2.1 discute trabalhos anteriores que abordam aspectos relacionados a serviços de geolocalização. A crescente massa de dados públicos gerada nas redes sociais *online* pode ser utilizada de forma positiva em sistemas de recomendação ou de modo negativo por exploradores virtuais, *marketing* viral e *hackers*. Isto levanta questões relevantes relacionadas à privacidade dos usuários, já que o permite explorar inúmeras possibilidades, incluindo a identificação de características que levam o usuário a um maior grau de exposição, conforme será discutido na Seção 2.2. Por fim, a inferência de características do usuário também pode ser aplicada em sistemas de controle da privacidade, recomendação e monitoramento de eventos. A Seção 2.3 aborda trabalhos anteriores que tratam desse tópico.

2.1 Serviços de Geolocalização

Serviços baseados em geolocalização tornaram-se muito populares nos últimos anos devido à difusão de redes sociais *online* e aplicativos móveis que utilizam a localização geográfica para permitir vários serviços de recomendação [Liu et al. 2015]. Destacam-se nos serviços de localização geográfica os aplicativos GoogleMaps[1], Waze[2], Life360

[1]https://maps.google.com
[2]https://www.waze.com

Family Locator[3] e as redes sociais *online* FourSquare e o Facebook Places, criado em agosto de 2010.

Apesar da rápida popularização de *smartphones*, pesquisadores observaram uma lentidão inicial por parte dos usuários em adotar tecnologias de compartilhamento de geolocalização [Page et al. 2012]. Esse fato pode ser atribuído a uma preocupação do usuário em partilhar a localização com terceiros, o que pode possibilitar uma violação da sua privacidade. Entretanto, atualmente, a utilização dos dispositivos móveis para acesso às redes sociais *online* se tornou extremamente comum. De fato, dados do Facebook indicam que mais de 1 bilhão de usuários acessam a rede social *online* por dispositivos móveis[4]. Dados do Twitter também mostram que 82% dos seus usuários usavam a plataforma móvel[5].

Redes sociais *online* e aplicativos móveis com serviços de geolocalização fomentam o surgimento de redes georreferenciadas (ou simplesmente redes geossociais)[6]. O compartilhamento crescente de informações georreferenciadas, impulsionado pelo uso de dispositivos móveis com interfaces mais amigáveis, contribuiu para a criação de um volume cada vez maior de dados geossociais. Tais dados capturam e refletem as atividades humanas e são bastante úteis para fomentar aplicações do mundo real, para diversas áreas como mídia, economia, turismo, dentre outras [Li & Hsieh 2015].

De fato, a crescente popularização do uso de dispositivos móveis equipados com sensores GPS criou um vasto campo para aplicativos e serviços de geolocalização, além de ser uma fonte valiosa de dados contendo informaçoes temporais e espaciais em diferentes granularidades. Em Bauer et al. [2012], os autores analisaram o conteúdo textual de milhões de *tips* (comentários) de usuários do Foursquare, criando um modelo para recuperar os tópicos mais discutidos de uma cidade. Os autores observaram que os tópicos mais populares são semelhantes em todas as cidades e correspondem a atividades cotidianas, como trabalho, comida, vida noturna e outros.

Neste contexto, formam-se as Redes de Sensoriamento Participativos (RSPs), ou seja, uma rede de dados constituída a partir do compartilhamento voluntário de informação contextual do usuário, normalmente contendo dados georreferenciados [Silva et al. 2014a], [Burke et al. 2006]. Segundo Silva et al. [2014a], as RSPs oferecem oportunidades sem precedentes de acesso a dados de sensoriamento em escala planetária, facilitando a obtenção de informações que não estão disponíveis prontamente, podendo ser aplicadas no processo de tomada de decisão de diferentes entidades, como

[3]https://www.life360.com

[4]https://newsroom.fb.com/company-info/ acesso em Julho de 2016

[5]https://about.twitter.com/company acesso em Julho de 2016

[6]É um tipo de rede social *online* que oferece serviços de geolocalização para a marcação de lugares, texto de publicações, *tags* e imagens, bem como a busca por locais, eventos e pessoas.

por exemplo, pessoas, grupos, serviços e aplicações móveis.

Estudos de diferentes RSPs que utilizam o compartilhamento de localização exploraram o comportamento e hábitos dos usuários, demonstrando que há padrões diferentes para dias de semana e final de semana [Silva et al. 2013a], [Silva et al. 2013b], [Silva et al. 2013c]. Por exemplo, nas redes sociais *online* Foursquare, Gowalla[7] e Brightkite [8] existe um aumento claro de utilização por volta da hora do café da manhã, almoço e jantar [Cheng et al. 2011], [Silva et al. 2013a]. Já no Instagram existem apenas dois horários de grande utilização, que ocorrem por volta da hora do almoço e jantar [Silva et al. 2013b]. Já no aplicativo móvel de alertas de trânsito Waze[9], também existem dois horários de maior evidência de uso, um por volta de 7 e 8 da manhã e outro por volta de 6 da tarde, coincidindo com horários típicos de maior intensidade no trânsito [Silva et al. 2013c]. Essas observações sugerem que é possível mapear a rotina de cidades com a utilização de dados georreferenciados.

Para oferecer ao usuário o serviço de geolocalização, as redes sociais *online* utilizam, em sua maioria, bases de dados criadas pelos próprios usuários. Nestes sistemas, usuários podem cadastrar novos locais, fazer *check-ins* em locais previamente criados, bem como denunciar dados incorretos. O posicionamento de um determinado local ocorre pelas coordenadas geográficas, principalmente via GPS *(Global Positioning System)*. Os autores criaram um sistema denominado Locus, capaz de localizar e visualizar pontos de interesse (POI - *Point of Interest*). Para isso, foi criado um *gazeteer*[10], cujo armazenamento se deu através da criação de instâncias de ontologia de locais. Já Twaroch et al. [2008] apresentaram um método de detecção de nomes de lugares na Web usando expressões relacionadas ao contexto da localização geográfica, cujo objetivo é melhorar a qualidade dos dicionários para sistemas de recuperação de informação geográfica. Por exemplo, os autores buscaram possíveis nomes de lugares próximos aos arredores de Cardiff, cidade do Reino Unido. Foi selecionada uma lista inicial com possíveis nomes de lugares candidatos. Após as etapas de filtragens e pesquisa na Web, 15 possíveis nomes candidatos restaram, como por exemplo, *Cardiff Gate Business Park, Cardiff International Arena e Cardiff Gate.* Apesar de alguns nomes não corresponderem a locais físicos, a grande maioria mostrou-se representativa.

A literatura também tem exemplos do uso de dados georreferenciados gerados pelas redes sociais *online*, nas áreas de educação, saúde, transporte e eventos de grande impacto [Rodrigues et al. 2015], [Davis Jr. et al. 2011], [Ribeiro et al. 2012], [Sakaki

[7]Rede social criada em 2009 e descontinuada em março de 2012.
[8]http://brightkite.com - Rede Social descontinuada
[9]https://www.waze.com
[10]Catálogo de nomes de lugares acompanhados de sua localização geográfica

et al. 2010] e [Gomide et al. 2011] .

Ribeiro et al. [2012] apresentaram um método para identificar os eventos e as condições de tráfego a partir de dados coletados do Twitter, geolocalizá-los e exibí-los em tempo real. O método proposto tem quatro etapas principais, a saber: (i) pré-processamento do conteúdo dos *tweets*; (ii) identificação e detecção dos eventos e condições relacionadas ao trânsito; (iii) detecção de locais usando cadeias de caracteres de correspondência exata; e (iv) detecção de locais por aproximação. Os resultados obtidos mostraram que o método é capaz de detectar bairros e vias públicas com uma acurácia que varia de 50 a 90%, dependendo do número de lugares mencionados nos *tweets*.

Detectar eventos é outra aplicação que é comumente explorada por dados gerados pelo Twitter. Sakaki et al. [2010] propuseram um método para monitorar mensagens do Twitter visando detectar a ocorrência de eventos. Os autores aplicaram o método na detecção de terremotos no Japão, com base no monitoramento de *tweets*, conseguindo detectar 96% dos mesmos. Gomide et al. [2011] analisaram como a epidemia da dengue é observada no Twitter e como as informações geradas podem ser utilizadas no controle da doença. Os autores utilizaram análise de sentimentos para analisar como os usuários se referem à dengue no Twitter, utilizando esse resultado para focar apenas nos *tweets* que expressavam conteúdo relevante sobre a dengue. Além disso, eles também construíram um modelo para a previsão do número de casos de dengue, aplicado no projeto do Observatório da Dengue[11]

A rede social *online* foco deste trabalho é o Foursquare. Por sua natureza georreferenciada, o Foursquare fornece uma fonte rica de dados sobre a mobilidade do usuário e de seus amigos, bem como seus padrões de comportamento que possibilitam a recomendação de locais de interesse, amizade e até oportunidades de negócios [Long et al. 2012], [Noulas et al. 2011a], [Noulas et al. 2013], [Brown et al. 2012]. Além da possibilidade de obtenção das coordenadas geográficas de localização do usuário, o Foursquare permite o acesso fácil a um conjunto adicional de características do local, tais como o tipo (por exemplo, restaurante) e popularidade (por exemplo, número de *check-ins*) [Rossi et al. 2015]. Uma aplicação de uso de dados georreferenciados do Foursquare foi abordada por Wang et al. [2015] para a previsão de fracasso de negócios empresariais. Os autores utilizaram os dados de *check-ins* do Foursquare de clientes de diversos locais para prever o insucesso de restaurantes em Nova York, ou melhor, antecipar possíveis falhas detectadas a partir do uso de redes sociais *online*. Utilizando várias técnicas de modelagem preditiva, como Redes Neurais[12] [Kriesel 2007] e o algoritmo

[11] http://www.observatorio.inweb.org.br/dengue

[12] As Redes Neurais Artificiais (RNAs) são modelos computacionais inspirados no cérebro humano,

de classificação K-Nearest Neighbors(K-NN) [Altman 1992], os autores concluíram que incorporar os dados de *check-ins* oferece uma melhora notável na previsão de falha de negócios empresariais.

Outra rede social *online* georreferenciada abordada em alguns trabalhos é o Gowalla. Allamanis et al. [2012] apresentaram um estudo da evolução temporal da rede e aplicaram diferentes modelos probabilísticos para caracterizá-la. Os autores demonstraram que a distância geográfica desempenha um importante papel na rede de amizade, pois os usuários tendem a manter uma rede de amizade com pessoas mais próximas fisicamente. Além disso, há tendências dos usuários em visitar lugares em comum entre os seus amigos e estabelecer novos laços de amizade com amigos dos amigos.

O trabalho desenvolvido nesta dissertação complementa os esforços anteriores discutidos acima por focar na inferência da localização de *residência* do usuário. Trabalhos anteriores específicos neste tema são discutidos na Seção 2.3.

2.2 Privacidade em Redes Sociais *Online*

Mesmo com todos os alertas sobre segurança e privacidade do usuário, as redes sociais *online* continuam a atrair o interesse das pessoas, fomentando o paradoxo da privacidade [Barnes 2006], segundo o qual participar de uma rede social implica em algum grau de exposição (tópico abordado na Seção 2.2.1). A exposição exagerada nas redes sociais *online* e até o desconhecimento de configurações de privacidade podem contribuir para ataques e ações maliciosas, conforme apresentamos na Seção 2.2.2.

2.2.1 Vulnerabilidade e Exposição

A maior parte dos usuários não possui consciência da exibição de suas informações pessoais, tais como nome, sobrenome, lista de amigos, fotos e preferências, em redes sociais *online*. Estas informações podem ser facilmente utilizadas de forma indevida, ocasionando situações diversas, desde a eventual exibição não autorizada até ações publicitárias dirigidas ao seu perfil [Danah & Marwick 2011]. Alguns casos mais sérios de violação de privacidade podem expor a pessoa até mesmo a situações de risco físico, como assaltos e sequestros, principalmente devido à exposição de publicações e *check-ins* com localização geográfica, ou até mesmo *check-ins* em sua própria residência[13].

utilizados em aprendizado de máquina e reconhecimento de padrões

[13]No Foursquare existe a possibilidade de atribuir a um *venue* a categoria *Residence*

Segundo Quercia et al. [2012], traços da personalidade do usuário refletem a forma de relacionamento nas aplicações de redes sociais *online*. A exibição das informações do usuário pode ser prevista através de um modelo que aborda o quanto a exposição da informação está relacionada com os traços de personalidade e outras variáveis, tais como idade, sexo e número de contatos.

Rauber et al. [2011] analisaram e compararam a percepção de privacidade usando experimentos coletados através de uma aplicação do Facebook. Foram analisados três atributos individuais (data de nascimento, cidade atual e gênero) e dois compartilhados com a rede (álbuns e *links*). Os autores concluíram que, enquanto a exposição de álbuns de fotos para o nível de visualização público (*everyone*) chega a 35% no geral, a estatística equivalente para *links* chega a 56%. A situação é mais acentuada no Brasil, onde esse *gap* atinge 27% (42% para os álbuns e 69% para *links*). Isso reforça a hipótese dos autores de que álbuns de fotos e *links* são duas características que mais proporcionam exposição do usuário. A respeito da percepção de privacidade dos usuários, os mesmos autores mostraram que homens e mulheres partilham a mesma quantidade de informações pessoais. No entanto, as mulheres tendem a ser mais cautelosas e tornam a informação menos visível do que os homens [Rauber et al. 2011].

Um dos grandes problemas relatados por usuários do Facebook tem relação com a privacidade das informações. Vale ressaltar que o padrão adotado das configurações relativas à privacidade assume a visibilidade também para usuários desconhecidos. Esse tópico foi abordado por Junior et al. [2014], em pesquisa experimental com usuários reais do Facebook. O estudo mostra que usuários têm dificuldade em antecipar os efeitos de alteração dos parâmetros associados à privacidade. Os autores criaram um simulador que permite ao usuário avaliar o impacto do efeito das configurações de privacidade.

Assim como o preenchimento de atributos sensíveis, ou seja, atributos que são capazes de identificar o usuário e ocasionar maior vulnerabilidade, o uso de *tags* em imagens já foi caracterizado como recurso potencial para a invasão de privacidade e ataques pessoais. Segundo Pesce et al. [2012], a marcação de fotos tem sido alvo de muitas críticas, principalmente considerando que o padrão de configuração de privacidade do Facebook permite a identificação e marcação do usuário em uma foto, além de sua publicação sem autorização prévia.

Para Ahn et al. [2011] a vulnerabilidade refere-se à perda de controle de dados pessoais. Na pesquisa realizada por Luo et al. [2009], os autores mostraram que se um usuário assume que uma rede social *online* é confiável, ele exibe suas informações no seu

perfil, incluindo dados pessoais, o que pode levar a ataques, tanto de engenharia social[14] [EC-Council 2009] quanto por *hackers*. Por exemplo, uma pessoa pode utilizar técnicas de engenharia social para tornar-se amigo de um usuário em uma rede social *online*, elogiar, agradar e até persuadi-lo a fornecer informações sigilosas como endereço, conta bancária e dados do cartão de crédito

A vulnerabilidade de exposição de dados pessoais também já foi observada na rede social *online* Foursquare. Ao realizar um *check-in* em um determinado *venue*, o usuário está revelando implicitamente o local em que se encontra, bem como o seu tipo ou categoria[15]. Em Bilogrevic et al. [2015], os autores analisaram os padrões de atividades dos *check-in* dos usuários e desenvolveram um modelo para prever o próximo *check-in* dos mesmos. Os autores também propuseram uma ferramenta para ser utilizada em redes sociais *online* com o objetivo de ofuscar a localização exata de um local e até mesmo o seu tipo ou categoria.

Segundo Jin et al. [2012], 36,58% dos usuários marcam suas residências como *venues*, enquanto 15,61% realizam *check-ins* públicos, sem nenhuma preocupação com a exposição das informações. Os autores mostraram que, embora a maioria dos usuários esteja ciente da sensibilidade dos endereços residenciais e evite se expor por completo no Foursquare, é possível obter as coordenadas de latitude e longitude dos *venues* (particularmente os *venues* da categoria *Residence*) através de algum sistema de API pública, por exemplo, no Google Maps é possível obter o endereço (com certa precisão) através de uma entrada com coordenadas geográficas. Ainda segundo Jin et al. [2012], explorando as coordenadas dos *venues* da categoria *Residence*, torna-se possível extrair essa informação e mapear o endereço do usuário, com uma acurácia de até 800 metros de distância.

Em outro trabalho referente ao Foursquare, Rossi et al. [2015] investigaram o conteúdo semântico dos *venues* visando a identificação de usuários, com foco na detecção de possíveis agentes maliciosos. Os autores analisaram mais de 1 milhão de *check-ins* em 17 regiões urbanas dos Estados Unidos e revelaram que diferentes tipos de *venues* podem chamar mais a atenção de usuários maliciosos. Por exemplo, cerca de 80% dos usuários informam o nome do local ao realizarem um *check-in* em um *venue* da categoria *Shop*. Os autores também mostraram que usuários que possuem a maior parte de *check-ins* em *venues* mais populares são os mais fáceis de serem identificados. Isso, por sua vez, sugere que o comportamento coletivo da população leva a um maior

[14]Refere-se a manipulação psicológica de pessoas para a execução de ações e/ou a extração de informações confidenciais.

[15]No Foursquare, os *venues* são organizados em nove macro categorias, como exemplo *Arts & Entertainment, Colleges & Universities, Food e Residence*

risco de identificação individual a partir de dados georreferenciados.

2.2.2 Ataques e Ações Maliciosas

Grandes redes sociais *online* sofrem com situações que envolvem a vulnerabilidade e a privacidade de dados de seus usuários. O Twitter, por exemplo, já foi vítima de usuários oportunistas e maliciosos que enviam *spam*[16], aplicam golpes através de *phishing*[17], espalham *malwares* e outras atividades ilícitas [Ghosh et al. 2012].

Em 2010 a conta do Twitter do Presidente dos EUA, Barack Obama, supostamente foi invadida [BBC 2010]. Em seguida, o site *Wikileaks*[18] publicou grande quantidade de documentos confidenciais do Exército e do Governo dos Estados Unidos, com forte repercussão mundial [Times 2010]. Vasculhar as redes sociais *online* tornou-se uma das principais tarefas para as agências de inteligência em todo o mundo, buscando desde ações de terroristas até possíveis informações para suporte a ações de Estado.

Diversas estratégias são utilizadas por usuários maliciosos no Twitter para evitar a detecção e a suspensão da conta. Uma estratégia é seguir e ser seguido por usuários reais. Neste contexto, Yang et al. [2012] abordaram a estrutura topológica das relações sociais entre as contas dos usuários maliciosos, apresentando as principais características dos seguidores dessas contas, além de explorar as diversas táticas utilizadas pelos mesmos. Uma estratégia semelhante foi utilizada para detectar usuários maliciosos que postam vídeos na rede YouTube [Benevenuto et al. 2008]. Os autores coletaram vídeos postados no YouTube e, através de inspeção manual, criaram uma coleção de usuários classificados como legítimos, *spammers* e promotores de vídeos. Os usuários promotores de vídeo são aqueles que tentam ganhar visibilidade, associando a um vídeo seu um grande número de vídeo-respostas[19], normalmente não relacionados ao vídeo alvo, visando colocá-lo nas *top* listas mantidas pelo YouTube. Já usuários considerados *spammers* pelos autores são aqueles que associam seus vídeos, como vídeo-respostas, a vídeos muito populares, visando atrair atenção para os mesmos. No trabalho, os autores apresentaram uma caracterização de atributos que podem ser utilizados para diferenciar usuários nas três classes. Os autores ainda exploraram estes atributos como entrada para um algoritmo de classificação capaz de identificar corretamente 97% de

[16] Spam é o envio de mensagens não solicitadas em massa

[17] Phishing é uma forma de fraude eletrônica caracterizada por tentativas de adquirir dados pessoais de diversos tipos.

[18] https://wikileaks.org/

[19] O YouTube oferecia aos usuários a possibilidade de associar respostas, no formato de vídeos, a um vídeo alvo. Vídeos com um grande número de vídeo-respostas, ganhavam maior visibilidade no sistema.

usuários promotores de vídeos e 54% de *spammers*, errando apenas 5,4% de usuários legítimos.

A maior rede social da atualidade, o Facebook, também já foi alvo de usuários maliciosos, roubos de informação e grande exposição de dados pessoais. Para analisar a vulnerabilidade do usuário no Facebook, Gundecha et al. [2011] propuseram uma estratégia para estimar o índice de exposição de dados de um usuário. A estratégia leva em consideração a quantidade de dados públicos expostos pelo usuário e sua rede de amizade. Já Hanne et al. [2012] consideraram o peso de cada atributo com base na popularidade da comunidade analisada. Quanto menos disponível for um atributo, maior seria o seu peso. O estudo concluiu que os usuários tendem a preencher os atributos menos sensíveis no seu perfil e manter um número pequeno de usuários em sua rede de amizade, entre 1 e 250 amigos.

Outra forma de tratar a privacidade nas redes sociais *online* leva em consideração a utilização de aplicativos de terceiros. Esse tipo de aplicativo normalmente pode acessar informações do seu perfil, porém o usuário pode não possuir o controle sobre esse acesso, nem tampouco sabe como essas informações podem ser usadas. O mercado de aplicativos móveis é o que mais explora o acesso a informações do perfil do usuário, impulsionado principalmente por anúncios que dependem dessa informação para a divulgação precisa de produtos. Leontiadis et al. [2012] desenvolveram um arcabouço que busca o equilíbrio entre proteger a privacidade do usuário e as receitas do desenvolvedor com anúncios. Já Anthonysamy et al. [2012], desenvolveram um arcabouço chamado *Collaborative Privacy Management* (CPM), cujo objetivo é servir como uma camada entre as redes sociais *online* e os aplicativos de terceiros instalados em um perfil, visando mediar o acesso às informações privadas.

Segundo Krishnamurthy [2009], a Informação de Identificação Pessoal (*Personal Identification Information*, PII) é definida como uma informação que pode ser usada para distinguir e detectar um usuário em uma rede social. Os autores mostraram em seu estudo como essas informações estão sendo utilizadas através de aplicativos de terceiros nas redes sociais *online*, principalmente para fins comerciais, visando descobrir hábitos dos usuários e identificá-los. Funções simples associadas à combinação de informações em cabeçalhos HTTP (Hypertext Transfer Protocol) e em *cookies*[20] são repassadas para os aplicativos normalmente utilizados nas redes sociais *online* com a permissão do próprio usuário. Malin [2005] também mostrou que é possível identificar, somente com três atributos (data de nascimento, sexo e CEP), 87% da população dos EUA.

No contexto específico da privacidade do usuário nas redes sociais *online*, nós

[20]Dados enviados por um site web e armazenados em um arquivo do tipo texto no computador do usuário.

avaliamos, nessa dissertação, a combinação de atributos na construção de métodos de inferência de localização de residência de um usuário no Foursquare. Nosso objetivo não é limitar o acesso a estes atributos, mas sim avaliar até que ponto esta inferência pode ser feita com precisão, usando apenas dados públicos compartilhados pelos próprios usuários.

2.3 Inferência de Localização em Redes Sociais *Online*

As redes sociais *online* são fontes valiosas de informação. Prever com precisão a localização de um usuário em um instante de tempo específico ou o seu local de origem (como a sua residência), pode ser bastante útil para várias tarefas. Como exemplos, tais previsões podem auxiliar na identificação de contas de usuários maliciosos, personalizar recursos de recomendação e contribuir para que ferramentas de monitoramento de eventos em tempo real possam localizar um maior número de usuários com maior exatidão. Esta seção discute os principais trabalhos de recomendação de lugares e eventos que utilizam da inferência de localização (Seção 2.3.1) e os relevantes trabalhos em duas áreas relacionadas a este tópico: previsão da mobilidade humana (Seção 2.3.2) e inferência do local de residência de um usuário (Seção 2.3.3).

2.3.1 Recomendação de Lugares e Eventos

A recomendação de lugares e eventos é uma das tendências das redes sociais georreferenciadas, que aumenta com o crescimento e difusão de aplicativos móveis [Leontiadis et al. 2012]. Os usuários estão interessados em saber se um local é bom ou não, e querem receber notificações do que está acontecendo ao seu redor. Em 2011, Scellato et al. [2011] apresentaram o *NextPlace*, um modelo preditivo para locais baseado na análise de séries temporais não-lineares. O *NextPlace* utiliza dados de GPS e da rede *Wifi* para estimar o tempo de futuras visitas e o tempo de permanência em locais.

Um dos primeiros estudos com foco em recomendação de *venues* em redes sociais georreferenciadas foi o de Noulas et al. [2012b]. Nesse estudo, os autores exploraram padrões de mobilidade dos usuários, analisando os lugares visitados, a frequência e a regularidade de padrões de locomoção, em especial das redes Foursquare e Gowalla. Eles observaram que entre 60 e 80% dos *check-ins* dos usuários ocorrem em locais que não foram visitados anteriormente. A conclusão dos autores foi que a combinação de diferentes fontes de dados, incluindo dados das preferências dos usuários extraídos de

redes sociais georreferenciadas, contribuem para melhorar sistemas de recomendação.

Outro estudo que explora a recomendação de *venues* é o de Long et al. [2012]. Os autores construíram um modelo preditivo baseado nas trajetórias dos usuários, inferidas a partir de seus *check-ins*. Os autores realizaram uma análise temporal dos *check-ins*, considerando a granularidade de dias e semanas, agrupando-os em tópicos de temas específicos, como compras, lazer, alimentação e outros. A pesquisa revelou em quais situações os usuários estariam mais interessados em receber recomendações, por exemplo, a quais restaurantes as pessoas costumam ir após fazerem compras em um *shopping*; qual café é mais popular em torno de suas posições atuais.

Brown et al. [2013] analisaram o comportamento dos usuários de 5 cidades dos Estados Unidos e sua rede de amizade com base em seus *check-ins* no Foursquare. Eles apresentaram um modelo que pode ser aplicado na recomendação de locais que o usuário poderá visitar e gerar laços de amizades.

Mapear áreas de uma cidade é um foco de estudo na linha de recomendação de locais e eventos, com o propósito de gerar regiões de interesse coletivo [Noulas et al. 2011b], [Silva et al. 2014b]. Noulas et al. [2011b] propuseram uma abordagem para classificar áreas e os usuários de uma cidade usando as categorias dos *venues* do Foursquare. A proposta consiste em agrupar os usuários em comunidades de interesse em padrões de visitação similares. Já em Silva et al. [2014b] uma técnica (*City Image*) desenvolvida pelos autores permite uma melhor compreensão da dinâmica e rotina das cidades. Ambas as técnicas poderiam ser aplicadas como entrada para sistemas de recomendação de lugares que exploram os padrões de deslocamento mais típicos.

Outra caracterização dos espaços urbanos foi abordada por Zhang et al. [2013]. Os autores utilizaram as informações obtidas a partir de *check-ins* de usuários do Foursquare para caracterização de espaços urbanos e organização dos mesmos em bairros. Os autores também propuseram a ferramenta *Hoodsquare*, para inferência do bairro de residência do usuário, visando uma aplicação em recomendação de lugares.

Enfim, a recomendação de lugares e eventos é uma das principais aplicações originadas da inferência de localização. A proposta deste trabalho não tem como foco específico desenvolver métodos de recomendação de lugares e eventos. Porém, os métodos propostos e apresentados no Capítulo 4 podem subsidiar o desenvolvimento de aplicações neste contexto, no futuro.

2.3.2 Previsão de Mobilidade

Uma das linhas de pesquisa relacionadas ao tema de inferência de localização explora o padrão de mobilidade do usuário visando prever a sua posição em um dado instante

de tempo. Alguns trabalhos que abordaram este tema são discutidos a seguir.

Gao et al. [2012] propuseram um modelo preditivo no Foursquare baseado no histórico de *check-ins* e na lista de amigos, para prever possíveis locais de *check-ins*. Sua principal conclusão é que os usuários tendem a visitar locais semelhantes aos dos seus amigos, bem como visitar o mesmo lugar mais de uma vez. Cho et al. [2011] descreveram um modelo da previsão de mobilidade a partir do histórico dos *check-ins* de usuários da rede social *Gowalla*. O resultados da avaliação do modelo proposto sugerem que há um comportamento periódico constante ao longo do dia, como exemplo, sair de casa para o trabalho. Desse modo, é possível prever quando o usuário está em casa ou está no trabalho.

Noulas et al. [2012a] estudaram os padrões de mobilidade urbana de usuários do Foursquare em 34 grandes metrópoles. Os autores exploraram os locais visitados pelos usuários através dos *check-ins*, bem como as distâncias de locomoção dentro de uma mesma região para prever a localização do usuário em um momento futuro. Complementando o trabalho anterior, Noulas et al. [2013] utilizaram um conjunto de dados de telefonia celular e *venues* georreferenciadas do Foursquare, criando uma estrutura de aprendizado supervisionado capaz de inferir atividades em centros urbanos. Nesse trabalho, os autores exploraram os padrões de comunicação dos usuários a fim de prever os *check-ins* em locais próximos, considerando um conjunto de tipos de atividades na área geográfica. Por exemplo, ao visitar um *shopping*, um usuário pode também fazer compras, ir ao cinema ou a um restaurante. Os autores mostraram que prever locais de entretenimento e vida noturna é mais fácil que áreas comerciais e escolas.

Silveira et al. [2015] propuseram um modelo de previsão de mobilidade humana (MobDatU), projetado para utilizar dados de fontes heterogêneas, notadamente dados de telefonia móvel e aplicativos georreferenciados. Os autores compararam a eficácia do modelo proposto com a de modelos considerados estado da arte que consideram uma única fonte (dados de telefonia móvel ou georreferenciados), concluindo que o novo modelo MobDatU é tão bom quanto ou mesmo superior ao melhor modelo alternativo em todos os cenários testados. Mais ainda, diferentemente do MobDatU, os modelos estado da arte não se mostraram robustos ao tipo de dado de entrada, com uma queda significativa no desempenho quando configurados com dados de tipos diferentes daqueles para os quais foram projetados.

Outro trabalho cujo foco é a mobilidade dos usuários do Foursquare é o de Karamshuk et al. [2013]. Os autores exploraram como a mobilidade dos usuários pode contribuir para identificar locais com bom potencial para abertura de um novo negócio. Vários outros estudos buscaram predizer a localização de usuários com base em dados de telefonia móvel, visando sobretudo otimizar os recursos da rede, selecionando de

forma inteligente as estações base mais adequadas para o tráfego da rede no momento [Bui et al. 2014], [Dong et al. 2013].

Ao contrário dos demais, nosso trabalho trata do problema da inferência de localização de residência propriamente dito. Porém, os modelos apresentados poderiam ser aperfeiçoados visando adequar a questão temporal, como por exemplo, acrescentar as informações de *check-ins* com data e hora, visando inferir a localização de um usuário em um dado momento.

2.3.3 Inferência de Localização de Residência

Nesta seção apresentamos os trabalhos no campo de inferência de localização que buscam descobrir o local de origem do usuário, considerando diferentes granularidades como país, estado, bairro, residência e até mesmo coordenadas geográficas. Vários autores abordaram essa linha de pesquisa, utilizando diferentes redes sociais *online*, como o Twitter, Facebook e Foursquare. O trabalho apresentado nessa dissertação enquadra-se nessa linha de pesquisa.

Chang et al. [2012] utilizaram modelos probabilísticos para prever as cidades de origem de usuários do Twitter com base no conteúdo dos *tweets*. Os autores propuseram um método para estimar a distribuição de probabilidade de uso de diferentes palavras em diferentes regiões usando modelos mistos gaussianos (*Gaussian Mixture Model* - GMM[21]). Eles também propuseram estratégias não supervisionadas para ordenar as palavras de uma dada região, visando remover ruído. A avaliação feita mostrou que a solução proposta superou o estado da arte na época, uma vez que atingiu um desempenho comparável ou até melhor usando apenas 250 palavras por região (versus 3.183 palavras usadas pela solução estado da arte). Diferentemente de Chang et al. [2012], o nosso trabalho não explora o conteúdo textual dos atributos, mas leva em consideração apenas as evidências georreferenciadas.

Outro trabalho que também analisou o conteúdo dos *tweets* foi o de Cheng et al. [2010]. Os autores propuseram estimar a localização (especificamente o bairro de residência) de um usuário explorando somente o conteúdo do seus *tweets*. As principais características da proposta são: (i) a dependência somente do conteúdo do *tweet* do usuário, sem a necessidade de outras informações, dados pessoais ou conhecimentos externos de outros sistemas; (ii) um componente de classificação do conteúdo dos *tweets*, explorando como entrada palavras referentes a locais e lugares georreferenciados; e (iii) um modelo de inferência da localização do usuário, com granularidade

[21]Modelo probabilístico que assume que todos os pontos de dados são gerados a partir de uma mistura de número finito de distribuições gaussianas com parâmetros desconhecidos.

do bairro. O sistema calcula k possíveis locais para cada usuário e classifica por ordem decrescente de confiança. Os melhores resultados foram alcançados com k igual a 5. Neste cenário, a estratégia proposta conseguiu inferir corretamente pelo menos 1 dos 5 lugares listados para 51% dos usuários, em uma base de aproximadamente 130 mil. Em nosso trabalho, utilizamos as granularidades de cidade, bairro e coordenadas geográficas para inferir a localização do usuário.

Em Mahmud et al. [2012] e Mahmud et al. [2014], os autores inferiram os locais de origem dos usuários do Twitter em diferentes granularidades, como cidade, estado, ou fuso horário, utilizando o conteúdo dos *tweets* e evidências do comportamento do usuário como o volume e a periodicidade de publicações. Utilizando um conjunto de classificadores, treinados a partir de *tweets* originados de 100 cidades dos Estados Unidos, os autores desenvolveram um algoritmo para inferir a localização de residência dos usuários. Os autores observaram que os melhores resultados foram obtidos quando o usuário possuía no mínimo 200 *tweets* e eliminando os usuários viajantes, ou seja, aqueles usuários que tinham publicações em diferentes locais, distantes geograficamente. Os melhores resultados consideraram uma combinação de classificadores (*Ensemble*), atingindo uma acurácia de 58% para as cidades, 66% para os estados e 78% para os fusos horários. Assim como Mahmud et al. [2014], tambem propomos um modelo que explora um conjunto de classificadores, conforme será apresentado no Capítulo 4.

Outros autores utilizaram dados extraídos do Twitter para inferir a localização de usuários, utilizando estratégias diversas [Rout et al. 2013], [Kong et al. 2014] e [Ribeiro 2015]. Rout et al. [2013] propuseram uma abordagem para inferir as cidades com base no perfil dos usuários e nas redes de amizade, produzindo como resultado uma ordenação (*ranking*) das cidades pela probabilidade de cada uma ser o local de residência do usuário. Kong et al. [2014] desenvolveram um sistema denominado SPOT, cuja finalidade é inferir a localização do usuário com base na rede de amizade. O trabalho foi inicialmente desenvolvido para a rede de menções do Twitter, assumindo pesos diferentes para os amigos de um usuário. Já Ribeiro [2015] propôs diferentes métodos para inferir a localização de um usuário, explorando as redes de amizade (seguidor e seguido) e de menções. Rodrigues et al. [2015] utilizaram um modelo probabilístico para inferir a localização de usuários, explorando características dos usuários do Twitter, como a frequência de suas publicações, a rede de amizade e o conteúdo dos *tweets*. Em um dos modelos propostos nesta dissertação, exploramos além das evidências do usuário (*mayorships*, *tips* e *likes*) a rede de amizade (lista de amigos), o que contribuiu para melhorar a acurácia da inferência.

Outra abordagem de inferência de localização foi proposta por Rossi & Musolesi

[2014]. A partir de caracterização das trajetórias temporais dos usuários, inferidas a partir da frequência dos *check-ins* feitas por eles e de suas redes de amizade, os autores propuseram estratégias para inferir a cidade de origem de usuários de diferentes redes sociais *online* geolocalizadas (Brightkite, Gowalla e Foursquare). No Foursquare especificamente, as inferências produzidas foram corretas para uma fração entre 30% a 50% dos usuários. Diferentemente de Rossi & Musolesi [2014] não utilizamos as trajetórias temporais dos usuários, e sim todas as evidências em nossa base de dados.

Um dos estudos utilizados como referência para essa dissertação foi o de Pontes et al. [2012b]. Nesse trabalho, os autores caracterizaram e analisaram os principais atributos dos usuários do Foursquare, como fontes de evidências para a inferência da localização de residência do usuário, e propuseram um método de votação pela maioria (*Majority Voting Scheme* - MVS) para inferir a cidade de origem do usuário. Os mesmos autores posteriormente estenderam o trabalho para o Google+ e Twitter [Pontes et al. 2012a]. Os resultados alcançados evidenciaram que foi possível descobrir a cidade de origem do usuário com uma acurácia de 67% no Foursquare, 72% no Google+ e 82% no Twitter. Especificamente no Foursquare, foi possível prever corretamente 53% das inferências com um raio inferior a 5 *Km*, e 77% em um raio de até 20 *Km*. Em seu trabalho de dissertação, Pontes [2013] propôs outros modelos de inferência, variantes do *MVS*, que exploram restrições como o número mínimo de votos, o número mínimo de evidências e a distância da cidade inferida. Esses modelos são tratados como referência para as nossas propostas, sendo detalhadas no Capítulo 4.

Complementando os trabalhos anteriores, em especial o de Pontes [2013], esta dissertação busca analisar em profundidade a inferência de localização de residência de usuários, utilizando a rede social *online* Foursquare. Apresentamos modelos variantes ao MVS com foco principal na rede de amizade do usuário, bem como uma abordagem que combina diferentes modelos, a qual levou a resultados melhores em comparação aos anteriormente apresentados na literatura.

Capítulo 3

Contextualização: Foursquare e Coleção de Dados

O Foursquare, uma das maiores redes sociais georreferenciadas, permite aos usuários efetuar *check-ins, tips* e *likes* em lugares visitados. O Foursquare possui integração com outras redes sociais *online*, como o Twitter, Facebook e Google+, favorecendo a relação de amizade. As principais características, o funcionamento e o histórico do Foursquare estão descritos na Seção 3.1. Na Seção 3.2 apresentamos uma visão geral da coleção de dados extraída do Foursquare utilizada nessa dissertação.

3.1 Conceitos Principais do Foursquare

O Foursquare, criado em 2008 e lançado em 2009, popularizou o serviço de compartilhamento de localização em redes sociais *online*. O acesso ao Foursquare normalmente ocorre através de aplicativos móveis com auxílio de GPS (*Global Positioning System*) para localização e marcação de lugares (*check-ins*). Os *check-ins* são compartilhamentos de localização em tempo real realizados em locais especiais, denominados *venues*. Os *venues* correspondem a locais existentes no mundo real, por exemplo, um restaurante, um *shopping* ou um parque, criados pelos próprios usuários, proprietários ou não do local. Caso o proprietário do local comprove a sua veracidade, o status do *venue* é marcado para verificado. Os *venues* são classificados em nove macrocategorias: *Arts & Entertainment, Colleges & Universities, Food, Nightlife Spots, Outdoors & Recreation, Professional & Other Places, Residence, Shop & Service* e *Travel & Transport*. Essas categorias possuem uma hierarquia de subcategorias que detalham com maior precisão o tipo do *venue*.

No Foursquare o usuário acumula pontos que são trocados por benefícios ou

concessões. Em particular, *check-ins* podem ser acumulados e trocados por *badges* e *mayorships*. *Badges* são como medalhas dadas a usuários que realizam *check-ins* em *venues* específicos ou alcançam um certo número de *check-ins* predefinidos. Uma *mayorship*, por sua vez, é dada ao usuário que efetuar o maior número de *check-ins* em um mesmo local em um intervalo de 60 dias. Além disso, os usuários também podem postar comentários curtos ou *tips* (conteúdo com até 200 caracteres) sobre os *venues*. Após ler uma *tip*, um usuário pode marcá-la com um *like* em sinal de aprovação de seu conteúdo [Vasconcelos et al. 2012]. Além disso, algumas empresas usam o Foursquare como um "cartão de fidelidade digital", proporcionando descontos aos usuários que realizam *check-ins*.[1]

Os usuários do Foursquare são categorizados em cinco tipos: *users*, *brands*, *chain*, *celebrity* e *venuePage*. Um usuário do tipo *users* pode estabelecer relacionamentos de amizade mútua e pode seguir usuários do tipo *celebrity* ou *brands*. Usuários caracterizados como *brands* representam marcas comerciais, como por exemplo a Bravo[2]. Já usuários do tipo *chain* representam uma cadeia de lojas de uma mesma marca, por exemplo, o restaurante *Outback Steakhouse*[3]. Os usuários do tipo *celebrity* estabelecem relações de seguidor e seguidos, e podem postar dicas, por exemplo a página da cantora Madonna[4]. Os usuários do tipo *venuePages* representam oficialmente a página de um local, possuem conteúdo e podem receber visitas de outros usuários, por exemplo, a página do BH Shopping[5].

Em 2012, o Foursquare criou medalhas (*badges*) para três níveis de superusuários: nivel 1 - Bronze, nível 2 - Prata e nível 3 - Ouro. Esses emblemas podem ser atribuídos a usuários comuns que se dedicam a organizar o Foursquare, denunciando conteúdo impróprio, indicando locais duplicados e com erro de localização, entre outras tarefas. No final do mesmo ano, o Foursquare lançou o recurso de dar notas aos *venues*, que vão de um a dez. As notas são utilizadas em um sistema de recomendação de locais, lançado em 2014, como pode ser visto na Figura 3.1.

Além dos diferentes tipos de usuários disponíveis no Foursquare, o desenvolvedor pode utilizar os recursos da API para criar aplicativos capazes de integrar recursos do próprio Foursquare. É possível interagir com *venues*, *check-ins*, usuários e amigos, extrair dados, procurar lugares e gerenciar através de aplicativo móvel toda a movimentação de um *venuePage*. Vários aplicativos móveis utilizam a API do Foursquare, que está disponível para as plataformas Android e Iphone, como por exemplo, o *Venue Map*

[1] http://aboutfoursquare.com/foursquare-101/
[2] https://foursquare.com/bravo
[3] https://foursquare.com/outback
[4] https://foursquare.com/madonna
[5] https://pt.foursquare.com/v/bh-shopping/4b4ba7f6f964a52006a326e3

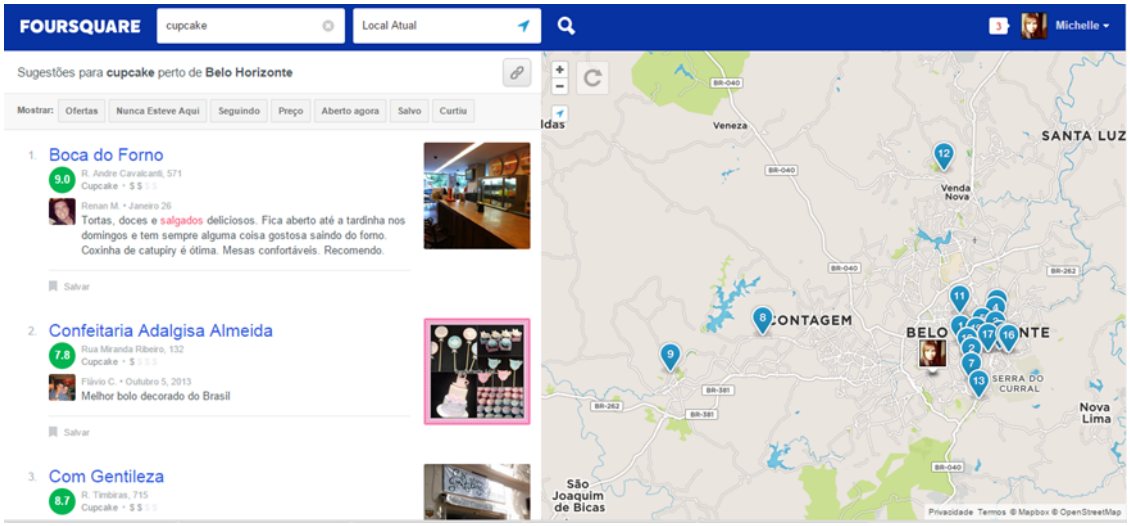


Figura 3.1. Resultado de busca no Foursquare, exibindo a recomendação de locais com notas

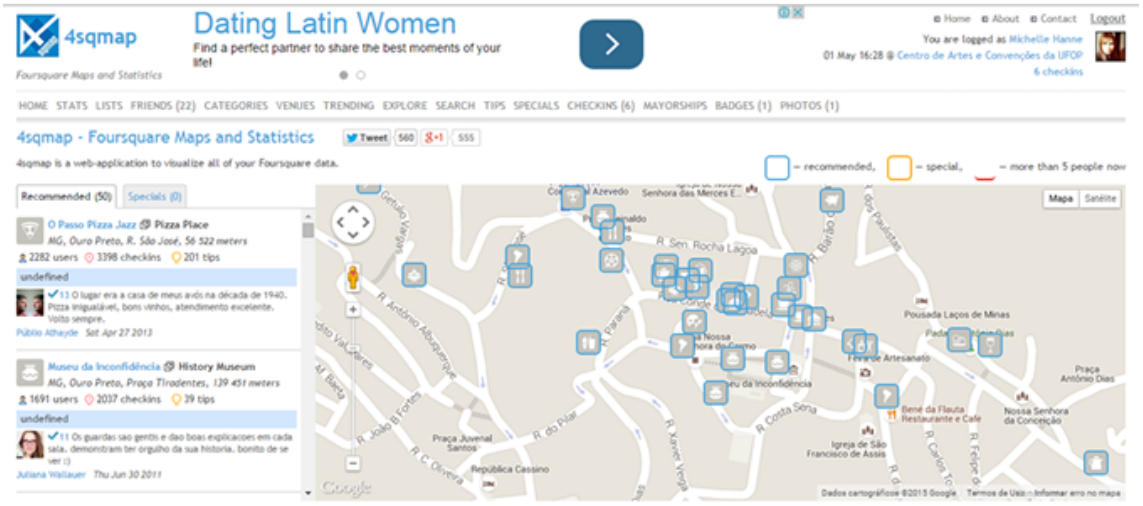


Figura 3.2. *Foursquare Maps and Statistics* (4sqmap) - Aplicativo Web para visualização de dados do Foursquare

for Foursquare - find venues around the world[6], capaz de encontrar locais em qualquer lugar ao redor do mundo. O ponto forte do aplicativo é a interação do usuário com os recursos de mapa, como agrupar informações e até realizar *check-ins*. Outro exemplo de aplicação da API do Foursquare é a ferramenta *Foursquare Maps and Statistics* (4sqmap)[7], onde o usuário pode visualizar em um mapa do Google Maps todos os seus dados extraídos do Foursquare, conforme visto na Figura 3.2.

Em 2014, o Fousquare lançou o aplicativo móvel *Swarm* para efetuar o serviço de

[6]https://itunes.apple.com/br/app/venue-map-for-foursquare-find/id406174482?mt=8
[7]http://www.4sqmap.com.

check-in, e assumiu a função de serviço de recomendação. O *Swarm* é um aplicativo próprio para os *check-ins*, que visa ser uma maneira mais rápida e fácil de acompanhar e encontrar amigos. Os *check-ins* efetuados pelo Foursquare foram transferidos para o aplicativo *Swarm*. O Foursquare foi repaginado para oferecer recomendações personalizadas aos usuários. A versão de aplicativo do Foursquare aprende sobre os hábitos do usuário e oferece sugestões por região. O usuário também pode escolher palavras-chave para configurar o seu perfil e receber sugestões e dicas. Ao clicar sobre um estabelecimento, o botão de *check-in* é exibido para os usuários que tiverem o *Swarm* instalado.

Estatísticas de março de 2016 do Foursquare revelam uma comunidade de aproximadamente 55 milhões de usuários em todo mundo, com mais de 65 milhões de *venues*, 70 milhões de *tips* e mais de 8 bilhões de *check-ins* realizados. Mais de 2 milhões de empresas estão cadastradas no Foursquare e possuem a sua localização confirmada para se conectarem com os seus clientes[8].

3.2 Coleção de Dados

Nesta seção apresentamos a coleção de dados utilizada na avaliação dos métodos de inferência propostos assim como os atributos usados por estes métodos. É importante ressaltar que esses dados foram coletados antes do Foursquare ser transformado em um sistema de recomendação de lugares, como atualmente se apresenta. Na Seção 3.2.1 descrevemos brevemente o processo de coleta dos dados do Foursquare e apresentamos uma visão geral dos mesmos. Na Seção 3.2.2 discutimos os atributos geograficamente referenciados extraídos da base de dados e na Seção 3.2.3 apresentamos uma breve análise desses atributos.

3.2.1 Coleta e Visão Geral dos Dados

A base de dados utilizada na avaliação dos métodos propostos foi coletada em 2011 por Pontes [2013], através da API do Foursquare. Foram coletados somente dados públicos e utilizados por pesquisadores [Vasconcelos et al. 2012], [Pontes 2013], [Pontes et al. 2012b], [Pontes et al. 2012a]. A seguir descrevemos brevemente o processo de coleta adotado.

O coletor partiu do princípio que cada usuário tem um ID (número identificador) único e sequencial. Então, considerando uma estimativa N do maior ID já atribuído a algum usuário, o coletor sorteava, de acordo com uma distribuição uniforme entre 1

[8] https://pt.foursquare.com/about

Tabela 3.1. Visão Geral da Coleção de Dados do Foursquare [Pontes 2013].

Item	*Número*
Usuários	13.570.060
Venues	15.898.484
Mayorships	15.149.981
Tips	10.618.411
Likes	9.989.325
Usuários com amigos	6.973.727

e N, um ID a ser coletado. Uma requisição à API do Foursquare era então enviada para recuperar os metadados associados àquele usuário como cidade de origem, *tips, likes*, lista de amigos e demais dados públicos. Também foram coletadas informações associadas aos *venues* sinalizados nos *mayorships, tips* e *likes* do usuário, tais como as suas categorias e as coordenadas geográficas.

O maior ID para o qual se obteve uma resposta de página válida foi de 20 milhões, valor usado para N. Pontes [2013] conjectura que no período da coleta esse era o maior ID de usuário cadastrado no sistema. O coletor foi executado de agosto a outubro de 2011. Foram coletados mais de 13 milhões de usuários, mais de 15 milhões de *venues*, além de atributos como *mayorships, tips e likes*, associados a cada usuário. Mais de 6 milhões (51,4%) de usuários na coleção possuem relação de amizade, cerca de 2.873.883 possuem pelo menos um *mayorship* e 1.802.997 possuem alguma *tip* ou *like*. A Tabela 3.1 sumariza os dados públicos utilizados neste trabalho.

3.2.2 Informações Geograficamente Referenciadas

Os atributos considerados nos modelos de inferência de localização de residência propostos nesta dissertação são geograficamente referenciados. Exemplos incluem a cidade onde o usuário reside (*home city*) e a lista de *venues* aos quais os *mayorships*, *tips* e *likes* do usuário estão associados.

Como o campo cidade de residência é livre, porém limitado a 100 caracteres, o usuário pode informar qualquer texto sem nenhuma validação do Foursquare. Alguns usuários colocam frases, *e-mails* ou mesmo lugares inexistentes. Já o *venue* é preenchido no momento da criação e possui uma marcação (*pin*), informando a sua localização, como por exemplo, endereço e cidade. No entanto, pode haver marcações arbitrárias ou até mesmo endereços incompletos e inválidos.

A autora Pontes [2013], utilizou o *Yahoo! PlaceFinder*[9] para padronizar e desambiguar as localizações dos *venues* e o nome das cidades fornecidas pelos usuários, identificando locais inválidos. O *Yahoo! PlaceFinder* permitiu identificar uma cidade, mesmo com variações do mesmo nome, como por exemplo, São Paulo, Sampa e SP. O processo funciona basicamente através de um campo textual de consulta, retornando uma resposta válida ou uma mensagem de erro. As respostas válidas consistem em um indicador entre 0 e 99, que representa a granularidade (rua, bairro, cidade, país e outros) do local consultado. Por exemplo, para a consulta "Belo Horizonte", o indicador retornou 40, com isso, foi possível obter o resultado das coordenadas geográficas (latitude: -19.945360 e longitude: -43.932678) da cidade de Belo Horizonte, o nome do estado (Minas Gerais) e o país (Brazil).

Como mostrado na Tabela 3.1, a coleção de dados contém um total de 13.570.060 usuários e 15.898.484 *venues*. A partir do uso do *Yahoo! PlaceFinder* foi possível obter a informação geográfica de cidade válida para 10.354.058 (76%) usuários e 6.937.523 (44%) *venues*. Ou seja, 24% das cidades de residência dos usuários não são válidas, estão em branco ou geraram ambiguidade. Isso revela que os usuários do Foursquare, em sua grande maioria, tendem a preencher o atributo de cidade de residência de forma correta.

A informação geográfica obtida através do *Yahoo! PlaceFinder* permitiu extrair granularidades diferentes, referentes à qualidade da informação, como visto na Tabela 3.2. Observa-se que a granularidade varia de continente até coordenadas geográficas exatas, porém, grande parte dos usuários estabelece uma relação municipal, informando a cidade e até mesmo o bairro onde reside no atributo *home city*.

Portanto, assim como em Pontes [2013], a principal granularidade utilizada nos experimentos dessa dissertação foi a cidade de residência do usuário, totalizando 10.354.058 usuários válidos e 6.937.523 *venues*.

3.2.3 Análise dos Atributos dos Usuários

Essa seção apresenta uma breve análise dos atributos coletados, visando avaliar potenciais combinações de atributos que possivelmente abrangeriam uma maior fração de usuários na inferência de localização de residência. Focamos nos atributos: *mayorship, tip, like* e a lista de amigos.

Temos que dos 10.354.058 usuários, cerca de 30%, ou quase 4,2 milhões, possuem pelo menos um *mayorship*, ou uma *tip* ou um *like*. Destes, mais de 1 milhão e 800 mil possuem pelo menos um *mayorship*, cerca de 1 milhão e 500 mil pelo menos uma *tip* e 1

[9] http://developer.yahoo.com/geo/placefinder/

Tabela 3.2. Números de Locais de Residência Válidos na Base de Dados do Foursquare [Pontes 2013].

Granularidade	*#Usuários*	*#Venues*
Continente	107	61
País	602.932	294.596
Estado	390.224	93.513
Cidade	10.354.058	6.937.523
Bairro	981.139	1.060.124
Área de Interesse	27.307	47.896
Rua	326.751	95.543
POI	5.607	9.792
Coordenadas Geográficas	61	32

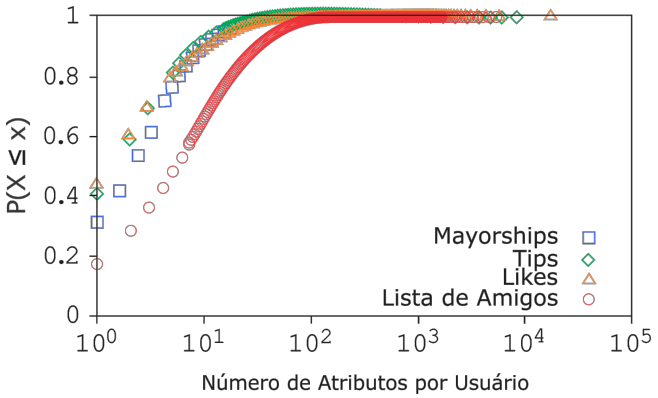


Figura 3.3. Distribuição dos atributos por usuários do Foursquare

milhão e 100 mil um *like*. Com relação à rede de amizade o número de usuários é mais expressivo, temos que quase 7 milhões de usuários possuem pelo menos um amigo.

O gráfico da Figura 3.3 mostra as distribuições acumuladas dos números de *mayorships*, *tips*, *likes* e amigos por usuário, indicando caudas longas, o que significa que alguns usuários têm muitos atributos (*mayorships*, *tips*, *likes* e *friends*), porém a grande maioria tem poucos atributos. Essa constatação já foi feita por outros autores [Noulas et al. 2011a], [Vasconcelos et al. 2012] e [Pontes 2013].

Analisando o número de atributos por cidade, descobrimos que a distribuição continua sendo muito enviesada: poucas cidades têm mais de 100 *mayorships*, *tips*, *likes* ou *friends*(lista de amigos). Agrupamos os atributos (*mayorships, tips, likes* e lista de amigos) por cidade, em uma distribuição acumulada, conforme visualizado na Figura 3.4.

Buscando descobrir se há relação entre os atributos (*mayorships, tips, likes* e lista

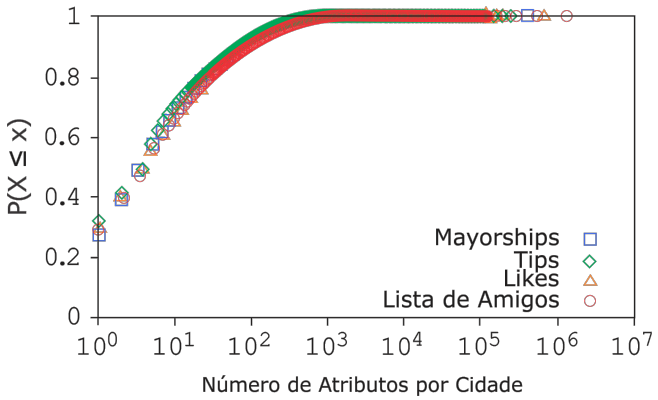


Figura 3.4. Distribuição dos atributos por cidade do Foursquare

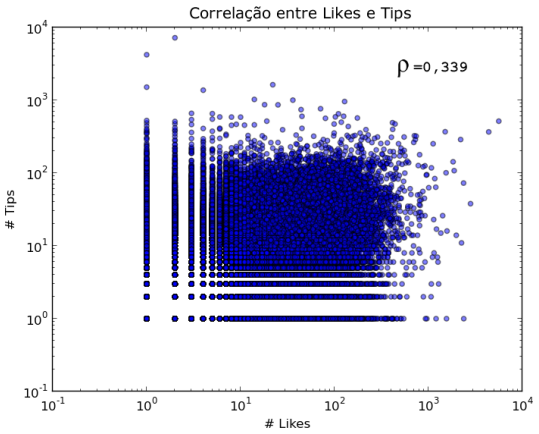
de amigos), fizemos a análise de cada par de atributos associados a um mesmo usuário utilizando o coeficiente de correlação de Spearman (ρ). O coeficiente de Spearman utiliza ao invés do valor observado, apenas a ordem das observações, com isso não faz suposições sobre a distribuição de frequência das variáveis e não requer que as mesmas sejam lineares. Desse modo, o coeficiente de Spearman não é sensível a assimetrias na distribuição, nem à presença de *outliers*, portanto, não exige que os dados sejam de duas populações normais. O coeficiente ρ de Spearman é dado pela fórmula:

$$\rho = 1 - \frac{6 \sum {d_i}^2}{(n^3 - n)}$$

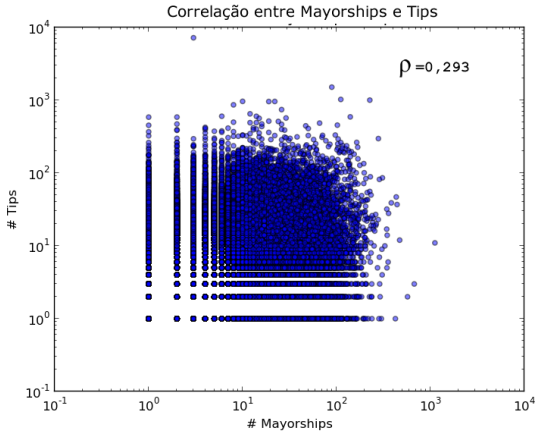
onde n é o número de observações e $d_i = rg(X_i) - rg(Y_i)$ é a diferença entre dois *ranks* (posições) de cada observação.

Os valores do coeficiente de Spearman variam entre -1 e +1, sendo que 0 implica que não há correlação entre as variáveis e valores próximos ou iguais a +1 e -1 indicam que há uma correlação que pode ser crescente ou decrescente. O sinal negativo do coeficiente indica que as variáveis variam em sentido contrário, isto é, os valores mais elevados de uma variável estão associados a valores mais baixos da outra variável. A Figura 3.5 mostra que existe correlação entre os atributos (*mayorships, tips, likes* e lista de amigos), porém fraca, ρ entre 0,23 a 0,34 .

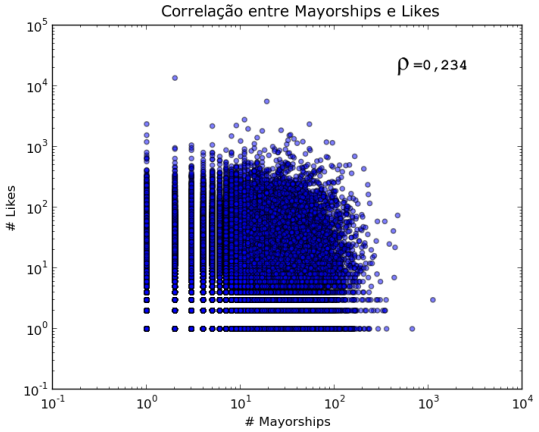
Os gráficos indicam uma correlação positiva entre os pares de variáveis analisadas, porém, percebemos que para pequenos valores de X temos grandes valores de Y. Por exemplo, o gráfico 3.5(d) que mostra a correlação entre *Mayorships* e *Friends*, temos usuários que possuem poucos *mayorships* e muitos amigos. As correlações tipicamente fracas entre os atributos sugerem que combinar diferentes atributos na tarefa de inferência de localização pode ser uma estratégia promissora no sentido que pode cobrir um número maior de usuário. Nós abordaremos esta tarefa no próximo capítulo.



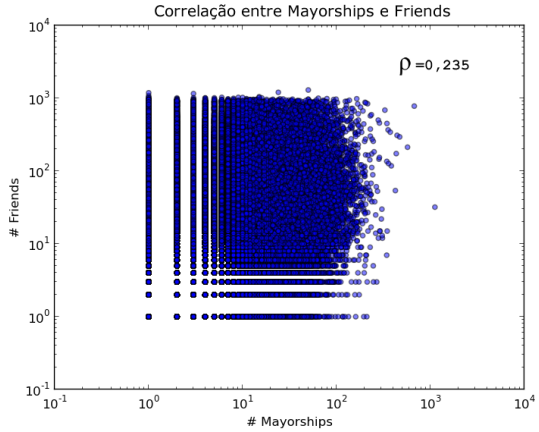
(a) Correlação LK e TP



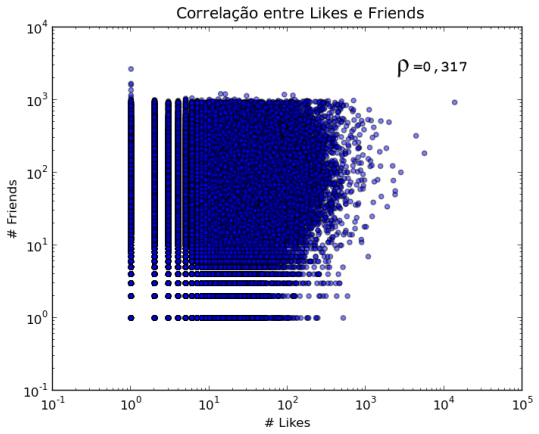
(b) Correlação MY e TP



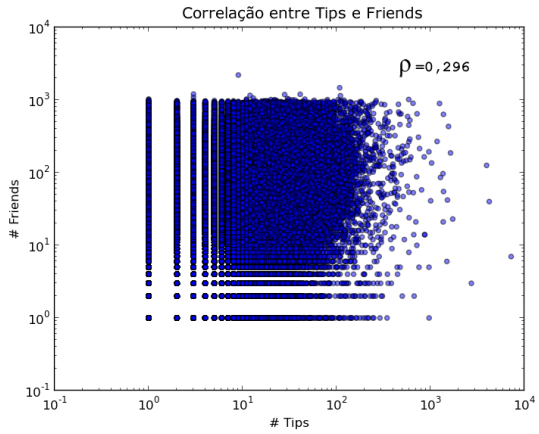
(c) Correlação MY e LK



(d) Correlação MY e FR



(e) Correlação LK e FR



(f) Correlação TP e FR

Figura 3.5. Correlação entre os atributos *mayorships, tips, likes* e *friends*

Capítulo 4

Modelos de Inferência de Localização de Residência

Nesta dissertação, buscamos propor métodos para inferir a localização de residência de um usuário, usando apenas dados públicos compartilhados no Foursquare. Embora a localização de residência possa ser obtida em diferentes granularidades, por exemplo, cidade, país e até mesmo coordenadas geográficas, exploramos em particular a inferência de cidade de residência do usuário, encontrada no atributo *home city*. Os modelos de inferência propostos utilizam as informações públicas de um usuário, disponíveis nos atributos, *mayorship, tips, likes e lista de amigos*.

Este capítulo apresenta na Seção 4.1 o problema de inferência abordado nesta dissertação e discute na Seção 4.2 os modelos de inferência de localização de residência que foram utilizados como referência (*baseline*) dessa pesquisa. A Seção 4.3 apresenta os novos modelos propostos.

4.1 Problema de Inferência

Formalmente, o problema de inferência alvo dessa dissertação pode ser definido como segue. Seja um usuário u, caracterizado por conjuntos de *mayorships* M_u, *tips* T_u, *likes* L_u e amigos F_u. Deseja-se inferir a localização de residência de u, $R_g(u)$ definida com uma granularidade g (exemplo, g = cidade, país, etc). Note que cada usuário no conjunto de amigos F_u também é caracterizado pelos mesmos atributos. Assume-se que os atributos são combinados em uma função f, que atribui pesos (potencialmente diferentes) α_i a cada um deles. Em outras palavras, tem-se expresso conforme a equação 4.1:

$$R_g(u) = f(w_m M_u, w_t T_u, w_l L_u, w_f F_u) \tag{4.1}$$

onde $w_m + w_t + w_l + w_f = 1$.

A principal premissa explorada nessa dissertação é que os usuários tendem a se locomover em regiões mais próximas de sua residência. Mais ainda, assume-se também que as movimentações dos amigos também tendem a se concentrar próximo de onde o usuário reside (conforme discutido em [Rossi & Musolesi 2014] e [Pontes et al. 2012b]). Por isso, as evidências de *mayorship*, *tip* e *like* dos amigos foram consideradas. Em princípio, a lista de *mayorship* poderia ser considerada a evidência mais forte para a inferência, já que os *venues* associados a estes *mayorships* são locais frequentemente visitados pelo usuario. Porém, como observado em [Pontes et al. 2012a], somente esse atributo não é garantia de inferências corretas para todos os usuários. Em particular, usuários sem nenhum *mayorship* não são elegíveis para essa inferência. Por isso, a combinação de atributos utilizada nos modelos visa ampliar a *cobertura* da inferência em termos do número de usuários elegíveis para inferência (exemplo, usuários para os quais os atributos explorados foram localizados geograficamente). Neste contexto, duas métricas devem ser consideradas, a acurácia (dada pela razão do número de inferências corretas pelo número total de inferências feitas) e a cobertura (dada pelo número de usuários para os quais a inferência foi feita). O modelo de inferência utilizado como referência, assim como os propostos, são abordados nas próximas seções.

4.2 Modelos de Referência

Conforme discutido na Seção 2.3.3, a literatura apresenta diferentes métodos de inferência de localização de residência de usuários em redes sociais *online*. Dentre os modelos apresentados naquela seção escolhemos as soluções apresentadas por Pontes [2013] como referência para nosso trabalho, uma vez que elas abordam o mesmo problema e o mesmo sistema alvo (Foursquare). A autora propôs diferentes modelos baseados no algoritmo *Majority Voting Scheme* (MVS), apresentando resultados robustos e significativos na inferência de localização da cidade de residência do usuário do Foursquare.

A técnica denominada *Majority Voting Scheme - MVS* consiste em uma regra de decisão que seleciona uma dentre várias alternativas por meio de uma *votação*: a alternativa com maior número de votos é selecionada. A regra de votos pela maioria possui algumas propriedades como: (i) trata cada alternativa elegível de forma idêntica, permitindo a igualdade de opções; (ii) conta os pontos para cada alternativa elegível e (iii) seleciona um vencedor, eleito pela maioria dos votos. A regra de decisão de

votos pela maioria é utilizada em várias áreas do conhecimento, como por exemplo, administração e finanças [Fujioka et al. 1993], [Sun & Li 2008] e [Li & Sun 2009].

No contexto apresentado por Pontes [2013], o MVS foi aplicado para inferir a localização de residência de usuário no Foursquare. O método consiste em associar para cada atributo do usuário, um local elegível (dado pela informação geográfica associada ao atributo). Em seguida, somam-se os votos para cada local elegível. Por fim, os votos são apurados e se uma maioria é alcançada temos o vencedor.

A autora utilizou o modelo MVS com 15 combinações diferentes dos 4 atributos: *mayorship, tip, like* e lista de amigos (*friends*). Exemplos de combinações consideradas são, *Mayorship+Tip*, *Mayorship+Tip+Like* e *Mayorship+Tip+Like+Friend*. Os resultados alcançados na aplicação do modelo MVS foram significativos, pois em todas as combinações a acurácia foi acima de 50%, com destaque para os modelos que consideraram somente o *Mayorship* (58,03%), *Mayorship+Tip* (58,34%) e *Mayorship+Tip+Like+Friend* (54,93%). Em particular, a estratégia que considerou os 4 atributos alcancou uma cobertura de 7.153.077 usuários, superior a todos os outros métodos. Esses resultados evidenciam que o atributo *mayorship* é um bom candidato para inferir a cidade onde o usuário reside, mas a combinação com outros atributos garante uma maior cobertura (possivelmente com pequena penalização na acurácia)

Pontes [2013] também propôs variantes do modelo MVS para tratar os casos de empate, ou seja, quando múltiplos locais elegíveis recebem o mesmo número de votos. Uma das sugestões adotadas por Pontes [2013] foi aplicar um filtro ao modelo (*Filtered_MVS*), que consiste em dois parâmetros: (i) *min_evidence* - número mínimo de atributos explorados pelo modelo; e (ii) *min_votes* - número mínimo de votos para que um local seja elegível para inferência. Desse modo, com a utilização do parâmetro *min_evidence*, usuários que possuem poucos atributos são excluidos do modelo, por não possuírem evidências suficientes para a inferência. O parâmetro *min_votes* ressalta que locais eleitos com um pequeno número de votos não representam uma forte evidência de inferência. O impacto do uso dos parâmetros no modelo *Filtered_MVS* é a redução na cobertura de usuários, porém em todos os cenários, há ganhos significativos da acurácia. Por exemplo, *min_votes* fixado em 20 apresenta uma acurácia acima de 70%, porém, com uma redução de mais de 90% em cobertura. Para valores de *min_votes* acima de 100 não há melhorias evidentes da acurácia. Para a autora, um bom compromisso entre cobertura e acurácia seria a escolha dos valores para *min_votes* e *min_evidence* iguais a 5. Como exemplo, o modelo que utiliza os 4 atributos (*mayorships, tips, likes* e lista de amigos) parametrizado com *min_evidence*=5 apresenta uma acurácia de 60,47% e uma cobertura de 4.241.262 usuários. Já quando o mesmo modelo é parametrizado com *min_votes*=5, a acurácia aumenta para 66,99%, porém, o número de usuários

reduz para 2.710.606.

Outra variação do modelo apresentada por Pontes [2013] para tratar as situações de empate foi o *Iterative_MVS*, que consiste em uma abordagem iterativa do modelo MVS, no qual os locais candidatos empatados em uma iteração se tornam os (únicos) locais elegíveis como candidatos para a iteração seguinte. Cada atributo que não está localizado nos locais elegíveis na iteração corrente entra nos cálculos dos votos computados para o local mais próximo, dentre os elegíveis da iteração atual. O modelo visa minimizar o número de empates dentre os locais empatados inicialmente, o vencedor é o que apresenta o mais denso conjunto de atributos localizados nas suas proximidades. A autora adotou o parâmetro α para controlar o voto dos atributos não elegíveis, evitando que distâncias muito grandes gerem possíveis ruídos na inferência. Por exemplo, se α=100, somente será considerado um voto para um local elegível na inferência se a distância de localização for no máximo 100 Km do local candidato. Experimentos variando o parâmetro α em 100, 200 e ilimitado (∞) indicaram que o mesmo possuiu pouco impacto sobre os resultados.

Em todas as técnicas propostas em [Pontes 2013], a autora considerou pesos idênticos para todos os atributos, por exemplo, $\alpha_m = \alpha_t = \alpha_l = \alpha_f$. Nesta dissertação, exploramos os benefícios de atribuir pesos diferentes para os atributos.

4.3 Novos Modelos de Inferência de Localização de Residência

Esta seção apresenta os novos modelos de inferência de localização de residência propostos nesta dissertação. Inicialmente, na Seção 4.3.1, apresentamos extensões dos modelos propostos por Pontes [2013] para considerar a atribuição de pesos diferentes para os quatro atributos. A seguir, na Seção 4.3.2, apresentamos uma solução que explora os padrões de mobilidade do usuário e de seus amigos, denominada *MOB_User_Friends*. Por fim, na Seção 4.3.3 apresentamos uma solução híbrida baseada na combinação de múltiplos métodos.

4.3.1 MVS Ponderado

O modelo MVS Ponderado é uma extensão da estratégia MVS básica, explorada por Pontes [2013]. A autora parte da premissa que nem todos os atributos envolvidos na votação possuem a mesma importância e atribui pesos diferentes a cada um deles. Esse tipo de sistema de votação ponderada é utilizado, por exemplo, em reuniões de

Tabela 4.1. Exemplo de Apuração de Votos para Inferência de Cidade

Cidade	Mayorship	Tip	Like	Apuração	Inferência
Belo Horizonte	1	2		3	MVS
São Paulo		2	2	4	São Paulo
Belo Horizonte	w_m*1	w_t*2		(0,55*1)+(0,27*2)=1,09	MVS Ponderado
São Paulo		w_t*2	w_l*2	(0,27*2)+(0,18*2)=0,9	Belo Horizonte
$w_m = 0,55$ $w_t = 0,27$ $w_l = 0,18$, sendo que $\sum_w = 1$					

acionistas, onde os votos são ponderados pelo número de ações que cada acionista detém. O modelo MVS Ponderado pode ser caracterizado por jogadores, pesos e quota: (i) jogadores - representam as opções possíveis de votos, $P1, P2...PN$, sendo N o número total de opções; (ii) pesos - o peso (w) de cada alternativa (jogador), dado pelo número de votos que ele controla; e (iii) quota - representa o número mínimo de votos necessários para aprovar uma das opções.

Trazendo para o contexto do nosso trabalho, temos que no modelo MVS Ponderado cada opção distinta de localização de residência do usuário consiste em um possível jogador, ao qual foram atribuídos pesos, de acordo com o grau de relevância de cada atributo. A quota de votos para aprovação consiste na apuração do maior valor atribuído a um jogador (opção). Por exemplo: temos que um usuário u possui 1 *mayorship* na cidade de Belo Horizonte, 2 *likes* na cidade de São Paulo e 2 *tips* na cidade de Belo Horizonte e 2 em São Paulo. Se o método utilizado no exemplo fosse o MVS não ponderado [Pontes 2013], o resultado para a cidade de inferência seria São Paulo (4 votos). Já aplicando o modelo MVS Ponderado, os pesos são atribuídos a cada opção de voto. Considerando os pesos de *mayorship, tip e like* iguais a 0,55, 0,27 e 0,18 respectivamente, o resultado da inferência será a cidade de Belo Horizonte, conforme visto na Tabela 4.1. Neste exemplo de combinação de atributos, não utilizamos a lista de amigos (*friends*).

4.3.1.1 Pesos dos Atributos

Nossa proposta para o Modelo MVS Ponderado é descobrir os pesos para cada atributo envolvido na inferência de modo a maximizar o número de acertos a partir de um conjunto de treinamento[1]. Nós adotamos uma estratégia de força bruta para buscar pelos melhores valores para os pesos dos atributos. Para cada combinação de atributos

[1]Utilizamos o processo de aprendizado supervisionado, em que a base de dados é divida em treino e teste

explorada (exemplo: *Mayorship+Tip+Like*), nós realizamos uma busca pelo peso de cada atributo no intervalo de 0 a 1, com incrementos de 0,1. Em outras palavras, em cada iteração da busca, o valor de um atributo variou de 0,1 até 0,9, enquanto os demais pesos ficaram fixos, ao final o somatório dos pesos foi normalizado em 1.

Os pesos dos atributos foram obtidos a partir de um conjunto de treinamento para cada combinação proposta, desse modo, testamos todas as possibilidades de valores w para cada atributo e selecionamos a combinação que melhor obteve resultados.

Aplicamos o modelo MVS Ponderado nas 11 combinações: *Mayorship+Tip, Mayorship+Like, Mayorship+Friend, Tip+Like, Tip+Friend, Friend+Like, Friend+Tip, Mayorship+Tip+Like, Mayorship+Tip+Friend, Tip+Like+Friend* e *Mayorship+Tip+Like+Friend*. Em cada combinação o peso dos atributos foi obtido no conjunto de treinamento. No capítulo 5 discutiremos a metodologia de avaliação e os valores dos pesos obtidos nos modelos.

4.3.2 Modelo MVS Ponderado Filtrado

Assim como em [Pontes 2013], também propomos variações do Modelo MVS Ponderado, visando tratar os casos de empate. A versão MVS Ponderado Filtrado consiste na filtragem de 2 parâmetros, o número de atributos (*min_evidence*) e o número de votos (*min_votesweight*). O *min_evidence* corresponde ao número mínimo de atributos considerados no modelo. Já o *min_votesweight* consiste na menor soma (ponderada) dos votos para que uma evidência seja considerada para a inferência. Exemplos da aplicação do modelo MVS Ponderado Filtrado:

1. ***min_evidence*:** Na combinação *Mayorship+Tip* com o parâmetro *min_evidence*=1, temos a seleção de usuários que possuem no mínimo 1 evidência no total, independente do atributo. Em nossa base de dados temos poucos usuários muito ativos e muitos usuários pouco ativos. Isso significa que quanto maior o valor de *min_evidence*, menor o número de usuários para os quais a inferência poderá ser aplicada.

2. ***min_votesweight*:** Na combinação *Mayorship+Tip* com o parâmetro *min_votesweight*=3, temos usuários com $R_g(u) = 3,2$ e $R_g(u) = 4,3$, entre outros. Semelhante ao parâmetro *min_evidence*, à medida que a soma mínima de votos ponderados necessários para que uma evidência seja eleita aumenta, o número de usuários torna-se bem reduzido.

4.3.3 Modelo MVS Ponderado Iterativo

A versão Iterativa do modelo MVS Ponderado representa uma alternativa para tratar a questão de empate dos candidatos à inferência. Assim como a proposta original de Pontes [2013], essa estratégia utiliza n iterações até que reste apenas um candidato para a inferência. A técnica iterativa consiste nos seguintes passos:

1. Separar as inferências candidatas (com maior número de votos ponderados) e que estejam em situação de empate (grupo A), daquelas que receberam a maioria dos votos (ponderados) (Grupo B);

2. Para cada opção existente no grupo B, calcula-se a distância entre ela e cada uma das inferências candidatas no grupo A. Utilizamos nesta tarefa o cálculo da distância entre as coordenadas geográficas de duas cidades com base na fórmula de Haversine[2], que considera a Terra uma esfera não perfeita (raio igual 6371 km).

3. Os votos (ponderados) associados a cada opção no grupo B são migrados para a inferência do grupo A que estiver mais próxima (menor distância). Para evitar associar locais muito distantes àqueles do grupo A, um limiar α é adotado como a distância máxima para que os votos de uma inferência do grupo B possam ser migrados para um local do grupo A. Como em Pontes [2013], variamos o parâmetro α em 3 faixas: $100km$, $200km$ e ∞km. O valor ∞ estabelece que não há uma distância máxima definida;

4. Após associar uma opção do grupo B ao grupo A, a soma ponderada de votos de cada inferência do grupo A é recalculada, definindo assim um novo grupo A de inferências candidatas;

5. O processo iterativo continua até que reste apenas uma inferência no grupo A. A estratégia de associação, em caso de empate, de uma opção do grupo B ao grupo A, leva em consideração que a inferência resultante da iteração apresente um conjunto ponderado mais denso de atributos. Desse modo, torna-se mais provável inferir corretamente a localização de residência de um usuário.

A Figura 4.1 ilustra um exemplo da aplicação do Modelo MVS Ponderado Iterativo para a combinação *Mayorship+Tip+Like* na inferência da cidade de residência

[2]A fórmula de Haversine é uma importante equação usada em navegação, fornecendo distâncias entre dois pontos de uma esfera a partir de suas latitudes e longitudes. https://en.wikipedia.org/wiki/Haversine_formula acesso em Agosto de 2016

Distribuição dos Atributos

Cidades	Mayorship	Tip	Like
Belo Horizonte	1	2	4
Rio de Janeiro	1	2	4
Sete Lagoas	0	1	3
São Paulo	0	3	4
Salvador	0	3	3

Apuração dos Pesos

Cidades	$R_g(u)$
Belo Horizonte	=(0,55*1)+(0,27*2)+(0,18*4)
Rio de Janeiro	=(0,55*1)+(0,27*2)+(0,18*4)
Sete Lagoas	=(0,55*0)+(0,27*1)+(0,18*3)
São Paulo	=(0,55*0)+(0,27*3)+(0,18*4)
Salvador	=(0,55*0)+(0,27*3)+(0,18*3)

w_m=0,55 w_t=0,27 w_l=0,18

1ª Iteração

Cidades	$R_g(u)$
Belo Horizonte	1,81
Rio de Janeiro	1,81
Sete Lagoas	0,81
São Paulo	1,53
Salvador	1,35

2ª Iteração

Cidades	$R_g(u)$
Belo Horizonte	3,97
Rio de Janeiro	3,34

Figura 4.1. Exemplo da aplicação do Modelo MVS Ponderado Iterativo

de um usuário u. Como mostrado na primeira tabela da Figura 4.1, o usuário u tem 1 *mayorship*, 2 *tips* e 4 *likes* em Belo Horizonte, 1 *mayorship*, 2 *tips* e 4 *likes* no Rio de Janeiro, 0 *mayorship*, 1 *tip* e 3 *likes* em Sete Lagoas, 0 *mayorship*, 3 *tips* e 4 *likes* em São Paulo e 0 *mayorship*, 3 *tips* e 3 *likes* em Salvador. O modelo utiliza os pesos $w_m = 0,55$ $w_t = 0,27$ $w_l = 0,18$. Após a primeira iteração, a aplicação do modelo em um usuário u, o que resultou no empate dos votos ponderados para as cidades Belo Horizonte e Rio de Janeiro com $R_g(u) = 1,81$. Na segunda iteração, após o cálculo da distância entre as cidades não selecionadas e as candidatas à inferência, os votos de Sete Lagoas e Salvador foram migrados para Belo Horizonte (cidade candidata mais próxima), enquanto os votos de São Paulo foram migrados para o Rio de Janeiro. A cidade de Belo Horizonte foi então eleita na inferência com $R_g(u) = 3,97$, enquanto que Rio de Janeiro ficou com o $R_g(u) = 3,34$. Neste exemplo, adotamos o parâmetro $\alpha = \infty$, por isso, aceitamos todas as distâncias entre as cidades não selecionadas e as candidatas à inferência.

No Capítulo 5, avaliaremos a eficácia do modelo MVS Ponderado e suas 2 variantes, Filtrado e Iterativo, para as 11 combinações de atributos.

4.3.4 Modelo Baseado na Mobilidade do Usuário e de seus Amigos (*MOB_ User_ Friends*)

A premissa principal do modelo que explora os padrões de mobilidade do usuário e de seus amigos é que eles tendem a visitar locais próximos ao local de residência do usuário. O método proposto, denominado *MOB_ User_ Friends*, consiste nos seguintes passos. Inicialmente os padrões de mobilidade do usuário u, alvo da inferência, são estimados a partir das localizações dos *venues* associados aos seus *mayorships, tips e likes*, através das coordenadas geográficas de cada um. O mesmo é feito para cada amigo do usuário u. O passo seguinte consiste em delimitar a área de mobilidade do usuário u, calculada através da interseção A^* entre as áreas de mobilidade de u e de seus amigos, obtendo, desse modo, os *venues* correspondentes aos *mayorships, tips e likes* comuns entre eles. Por fim, um dos algoritmos de inferência apresentados nas seções anteriores (exemplo: MVS, MVS Ponderado) é aplicado considerando apenas os *venues* localizados dentro na área de interseção A^*. Logo, o algoritmo *MOB_ User_ Friends* é uma extensão dos métodos anteriores que consiste em primeiramente reduzir a área de inferência para uma região comum, visitada tanto pelo usuário u quanto por seus amigos. O Algoritmo 1 mostra a visão geral do método *MOB_ User_ Friends*). Note que o método não pode ser aplicado quando a área de interseção A^* é nula. Neste caso, o usuário alvo é descartado (linha 6).

Algorithm 1 *MOB_ User_ Friends*

1: **for all** $i \in \{u, F(u)\}$ **do**
2: $\quad A_i \leftarrow CalculaAreaMobilidade(i, my(i), tp(i), lk(i))$
3: **end for**
4: $A^* \leftarrow \bigcap_{i \in \{u, F(u)\}} A_i$
5: **if** $A^* = 0$ **then**
6: $\quad R_g(u) \leftarrow Null$ ▷ usuario descartado
7: **end if**
8: $R_g(u) \leftarrow M(A^*)$ ▷ onde M = MVS ou MVS Ponderado

A parte principal do algoritmo *MOB_ User_ Friends* é o cálculo da área de mobilidade do usuário u e de seus amigos (linha 2 do Algoritmo 1). Este procedimento é mostrado no Algoritmo 2. Inicialmente, buscam-se as coordenadas geográficas dos *venues* associados aos *mayorships, tips e likes* do usuário. O passo seguinte consiste em delimitar a área formada pelos pontos, buscando os extremos superior direito e inferior esquerdo, como referência para criar uma área retangular (retângulo envolvente mínimo). Caso todos os pontos sejam coincidentes, a área delimitada é nula. Nestes

casos, o método não se aplica e o usuário é descartado. Os próximos passos consistem em reduzir a área inicialmente delimitada a partir de um processo de agrupamento (*clustering*) dos pontos associados aos *venues* do usuário. O objetivo é terminar com uma área menor, porém, mais densamente povoada de pontos visitados pelo usuário. Especificamente, é criada uma matriz de distâncias entre pares de pontos. Os pares de pontos mais próximos são sucessivamente agrupados, desde que a distância entre eles não ultrapasse um limiar β_{max}. Consideramos β_{max}=2.000 Km. A cada agrupamento um centróide é computado como representante dos pontos do *cluster*. O processo termina quando a distância entre os pontos mais próximos que pertencem a *clusters* diferentes for superior a β_{max}, ou quando todos os pontos estiverem em um mesmo *cluster*.

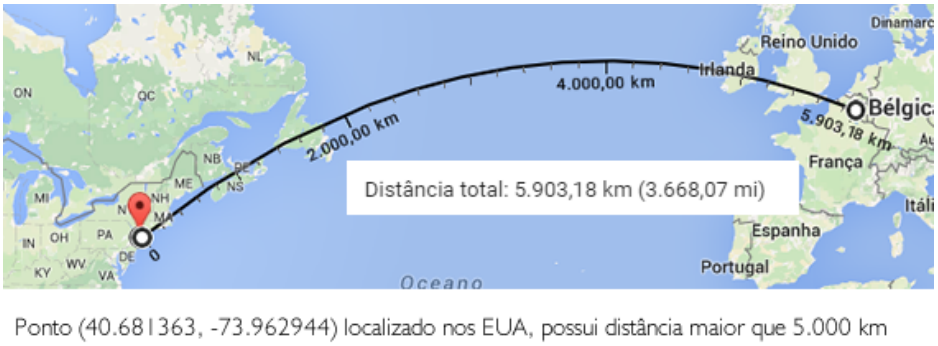
Caso o processo termine com mais de um *cluster*, propomos duas variações do algoritmo: (i) Densidade - que seleciona o *cluster* de maior densidade de pontos; e (ii) Densidade Ponderada - que seleciona o *cluster* que possui a maior soma ponderada, considerando os pesos dos atributos *mayorships, tips e likes*. Nesta situação, foram considerados os pesos de cada atributo obtidos nas instâncias de treinamento. Após a seleção do cluster a área formada pelos pontos é recalculada. O objetivo é terminar com uma área que não ultrapasse um limiar α_{max}=10.000 km^2. Caso a nova área calculada ultrapasse α_{max}, pontos são removidos do *cluster* na ordem reversa à sua inserção, um a um, até que o limiar α_{max} seja atingido. Em nossos experimentos consideramos α_{max} = 10.000 km^2. Ambos limiares α_{max} e β_{max} foram adotados para restringir a área de mobilidade dos usuários, tendo em vista o objetivo de focar em uma área reduzida, mais frequentemente visitada pelo usuário e por seus amigos.

A Figura 4.2 ilustra um exemplo da aplicação do Algoritmo 2 para um usuário u. O primeiro mapa da Figura 4.2(a) mostra a disposição dos *venues* do usuário u, que possui um *tip*, um *mayorship* e um *like* localizados na Bélgica e outro *like* nos Estados Unidos. A distância entre todos os pares de pontos (*venues*) do usuário u é calculada em uma matriz, conforme visto na Figura 4.2(c). Neste exemplo, temos 2 *clusters*, o primeiro com os pontos 1, 2 e 4 e o segundo com o ponto 3. O primeiro *cluster* é escolhido por ter a maior densidade (ponderada ou não) de pontos.

A área de mobilidade do usuário u é então calculada e o mesmo procedimento é seguido para todos os seus amigos. Neste exemplo, o usuário u possui apenas um amigo e sua área de mobilidade possui 5 pontos (*venues*), localizados todos na Bélgica (Figura 4.3(b)). Por fim, temos a área final de mobilidade do usuário u, correspondente a interseção das áreas que contêm três *venues*, conforme visto na Figura 4.3.

Após a obtenção da área de mobilidade do usuário u, utilizamos duas versões do modelo *MOB_ User_ Friends* para a inferência de localização de residência: (i) MVS e

(a) Mapa de *venues* do usuário u



(b) *Venue* com distância superior a 2.000 *Km*

Matriz de Distância entre Pontos

Ponto	1	2	3	4
1	0	12,15	5.903,32	15,63
2		0	5.905,55	19,77
3			0	5.903,32
4				0

1º Agrupamento

Ponto	(1-2)	3	4
(1-2)	0	5.904,44	16,75
3		0	5.903,32
4			0

2º Agrupamento

Ponto	(1-2-4)	3
(1-2-4)	0	5.898,94
3		0

Pontos

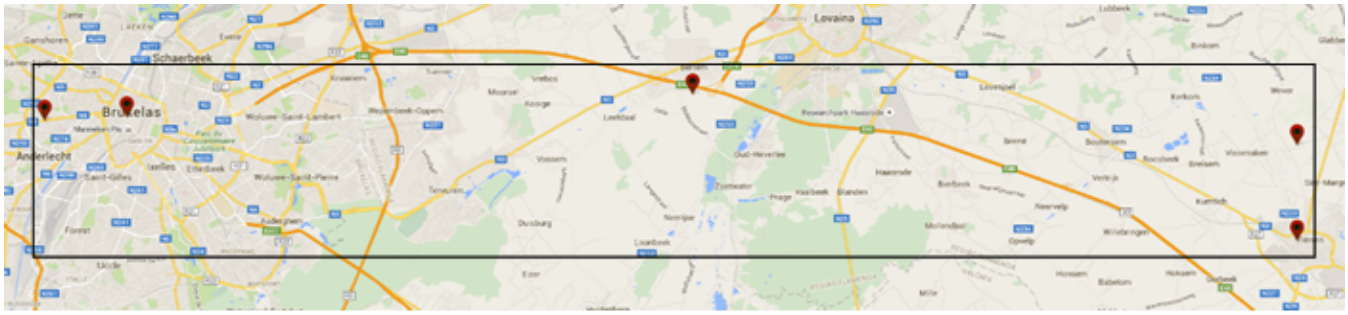
1 tp (50.773187, 4.537702)
2 my (50.865674, 4.629935)
3 Ik (40.681363, -73.962944)
4 Ik (50.848385, 4.349685)

*Distâncias maiores que 5.000 não selecionadas

(c) Matriz de Distância

Figura 4.2. Aplicação do modelo *MOB_ User_ Friends* no agrupamento de pontos

(a) Área do usuário u



(b) Área de $F(u)$

(c) Área final do usuário u

Figura 4.3. Aplicação do modelo *MOB_ User_ Friends* na obtenção da área de mobilidade

Algorithm 2 Calcula a área de mobilidade do usuário

```
1: function CalculaAreaMobilidade(u, my, tp, lk)
2:     β_max ← 2.000
3:     α_max ← 10.000
4:     A(u) ← Area
5:     if A(u) = 0 then
6:         u ← Null                                   ▷ usuario descartado
7:     else
8:         D ← Matriz
9:         d[(i),(j)] : medida da distância entre os pontos i, j
10:        mindist ← 0
11:        nclusters =número de pontos
12:        while (mindist < β_max) or (nclusters > 1) do
13:            d[(r),(s)] ← min d[(i),(j)]
14:            mindist ← min
15:            cluster ← merge[(r),(s)]
16:            NP[(i),(j)] ← centroide(r, s)   ▷ calcula o centroide, média dos pontos
17:            D ← Atualiza matriz de distância
18:            nclusters = nclusters − 1
19:        end while
20:        if nclusters > 1 then
21:            C_u ← getClusterMaxDensity                ▷ Algoritmo Density
22:            or
23:            C_u ← getClusterWeightedDensity      ▷ Algoritmo Weighted Density
24:        end if
25:        A(u) ← Área de C_u                              ▷ Recalcula a area
26:        while A(u) > α_max do
27:            C_u ← removeLastPoint          ▷ Remove os últimos pontos inseridos
28:            A(u) ← Área de C_u                          ▷ Recalcula a area
29:        end while
30:    end if
31:    return A(u)
32: end function
```

(ii) MVS Ponderado.

4.3.5 Modelo Híbrido

Propomos a utilização de uma combinação de classificadores, semelhante ao modelo *Ensemble*, proposto em Mahmud et al. [2014], ao qual denominamos modelo Híbrido. O modelo Híbrido busca combinar os resultados obtidos pelos métodos apresentados anteriormente, possibilitando uma maior cobertura e acurácia na inferência de localização de residência de usuários.

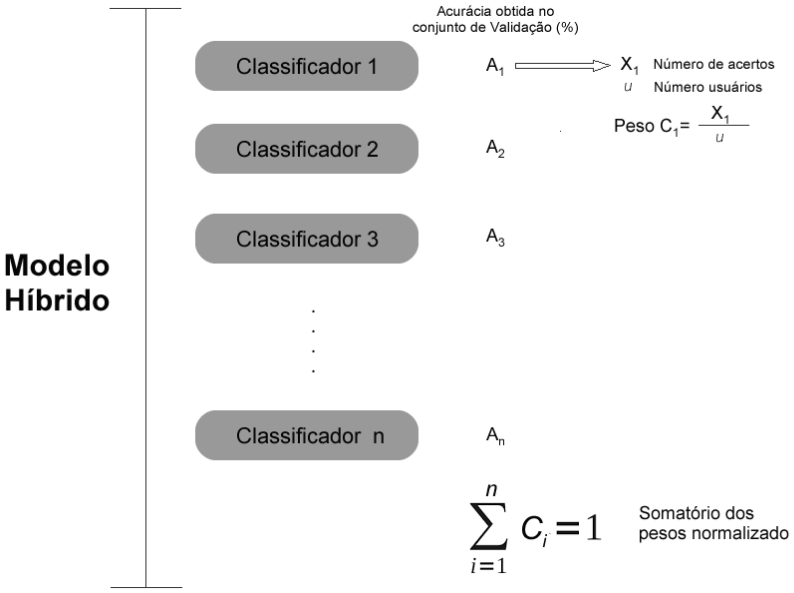


Figura 4.4. Esquema do Modelo Híbrido

Os resultados foram combinados considerando a votação ponderada de cada modelo (*Stacking*). A votação ponderada atribui a cada modelo um voto ajustado. Em nosso trabalho, utilizamos o número de acertos obtido no conjunto de validação como critério para ajustar os pesos. Utilizamos no modelo Híbrido as combinações dos melhores resultados das seguintes técnicas: MVS Ponderado, MVS Ponderado Filtrado, MVS Ponderado Iterativo e Modelo *MOB_User_Friends* (Densidade Ponderada). A Figura 4.4 ilustra o esquema de funcionamento do modelo Hibrido, com n classificadores, cada um com um percentual de acurácia obtida no conjunto de validação e o respectivo número de acertos. O peso é então calculado para cada classificador, considerando o número total de usuários u_t avaliados. E, por fim, os pesos são normalizados para que o somatório seja igual a 1.

4.4 Sumário

Neste capítulo, definimos o problema de inferência e apresentamos o modelo de Pontes [2013], adotado como referência para este trabalho. Apresentamos os modelos de localização de residência de usuários propostos nessa dissertação: MVS Ponderado e suas variações, *MOB_User_Friends* e Híbrido. O modelo MVS Ponderado é uma versão do MVS que trata os pesos de cada atributo utilizados na inferência. Mostramos duas variações do modelo MVS Ponderado, as versões Filtrado e Iterativo. Apresentamos

dois parâmetros com a versão MVS Ponderado Filtrado, são eles: *min_evidence* e *min_votesweight*. O parâmetro *min_evidence* trata o número mínimo de atributos utilizados na inferência. Já o *min_votesweight* trata a soma mínima ponderada para que uma evidência seja considerada para a inferência. O modelo *MOB_User_Friends* foi desenvolvido com base na mobilidade do usuário e sua rede de amizade. Mapeamos, através de coordenadas geográficas, os atributos de *mayorship, tip* e *like* vinculados aos *venues* do usuário. Obtivemos uma possível área de mobilidade para cada usuário, e fizemos a interseção com as áreas de mobilidade com seus amigos. A inferência foi gerada com os atributos contidos na interseção das áreas. O modelo Híbrido combinou os melhores classificadores de inferência de localização de residência. Utilizamos a técnica de votação ponderada para atribuir o peso a cada classificador. A Figura 4.5 apresenta o esquema de inferência de localização abordado no próximo capítulo, considerando as granularidades e os modelos propostos.

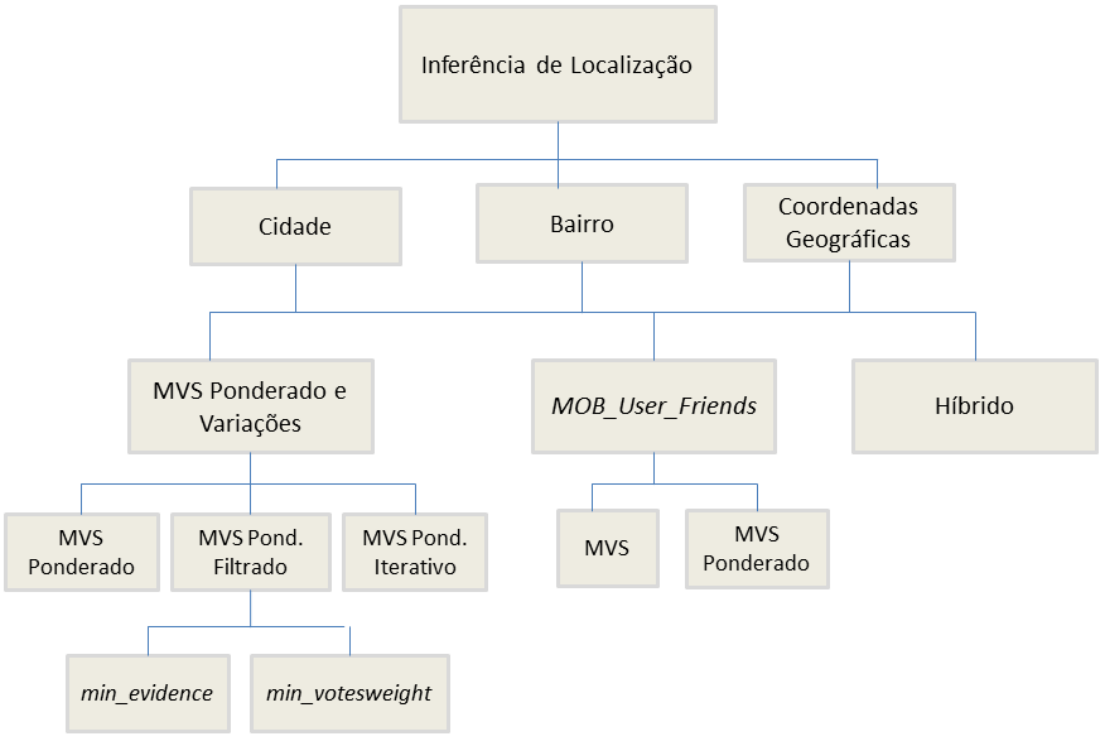


Figura 4.5. Esquema dos Modelos de inferência de localização de residência

Capítulo 5

Análise Experimental

Este Capítulo apresenta os resultados experimentais referentes aos modelos de inferência de localização de residência de usuários propostos no Capítulo 4. Discutimos na Seção 5.1 a metodologia adotada na avaliação dos modelos apresentados. Apresentamos na Seção 5.2 os resultados dos novos métodos de inferência para a granularidade cidade, além da análise de cada modelo separadamente. Na Seção 5.3 mostramos e analisamos os resultados do modelo de inferência para a granularidade de bairro e na Seção 5.4 de coordenadas geográficas. A Seção 5.5 apresenta uma discussão geral sobre os resultados e por fim, a Seção 5.6 apresenta as limitações encontradas nesta pesquisa.

5.1 Metodologia de Avaliação

Nesta seção apresentamos os principais aspectos relacionados à metodologia adotada para avaliar os modelos propostos. Discutimos o *ground truth* e os usuários elegíveis para a análise experimental. Apresentamos também as métricas utilizadas para avaliar a eficácia dos modelos de inferência de localização de residência de usuários.

A informação obtida no atributo *home city* de cada usuário de nossa base de dados foi considerada como o local de residência do usuário para a granularidade de cidade. Desse modo, adotamos esse atributo como sendo o nosso *ground truth* e consideramos apenas os usuários que possuíam locais válidos no nível de cidade. A validação do atributo *home city* foi feita através do *Yahoo! PlaceFinder*, conforme descrito na Seção 3.2.2. O número de usuários que possuem o atributo *home city* na granularidade cidade chega a mais de 10 milhões. Desses usuários, descartamos os que não possuem nenhum atributo associado. Conseguimos selecionar mais de 7 milhões de usuários, por exemplo, com a combinação dos atributos *Mayorship+Tip+Like+Friend*. Os usuários elegíveis na granularidade cidade foram divididos em duas classes, como feito por Pontes [2013]:

(i) Classe 0 - são aqueles usuários que possuem uma única evidência para a inferência, exemplo, um único *mayorship* ou uma única *tip*; e (ii) Classe 1 - são os usuários que possuem múltiplas evidências para a inferência, por exemplo, 1 *mayorship* e 2 *tips*.

Também aplicamos os nossos métodos para a inferência de local de residência nas granularidades de bairro e coordenadas geográficas. Nessas granularidades selecionamos somente usuários que possuem *mayorship* no *venue* de categoria *Residence*. Para estes usuários, as coordenadas geográficas do *venue Residence* foram consideradas o *ground truth*. O número de usuários avaliados na granularidade de coordenadas geográficas é de 832.191, cerca de 6,13% da nossa base de dados. Quanto à granularidade de bairro, selecionamos 6.344 usuários elegíveis da cidade de São Paulo, utilizamos a API do *GoogleMaps* para obter os nomes correspondentes aos bairros.

Quanto à metologia de avaliação dos experimentos utilizamos a técnica de validação cruzada com 5 *folds*. Esta técnica que consiste em dividir aleatoriamente a base de dados em duas partes: D_1 e D_2, instanciadas ora no conjunto de treinamento e ora no conjunto de teste. Primeiramente, executamos a técnica de força bruta para obter a melhor combinação de pesos, ou seja, selecionamos a combinação que obteve o maior número de acertos de inferência no conjunto de treinamento (D_1). Em seguida, executamos nos dados de teste (D_2), a mesma combinação de pesos apurada em D_1, obtendo então a acurácia no conjunto de teste (D_2). Posteriormente, realizamos o treinamento nos dados em D_2 e avaliamos o método, com os pesos aprendidos, em D_1. O processo foi repetido 5 vezes, posteriormente computamos a acurácia do modelo como a média das acurácias (D_1 e D_2), obtida dos testes de cada parte dos dados. As mesmas divisões de treino e teste foram aplicadas para todos os modelos.

Semelhante à descrição anterior, utilizamos no modelo Híbrido a técnica de validação cruzada para obtenção dos pesos de cada classificador. Porém, dividimos os dados em 3 conjuntos: (i) treinamento (D_1) com 40% dos dados; (ii) validação (D_2) com 10% dos dados; e (iii) teste (D_3) com 50%. Em D_1 os pesos são apreendidos para cada classificador, em D_2 os classificadores são aplicados e avaliados. Em D_3, o método Híbrido é avaliado com os classificadores ponderados pelas acurácias obtidas em D_2. O processo é repetido uma segunda vez, utilizando D_3 como conjuntos de treino (40%) e validação (10%) e o restante dos dados como teste.

Quanto à parametrização dos métodos, adotamos as seguintes configurações dos parâmetros para as variações dos modelos MVS e MVS Ponderado, conforme apresentado em [Pontes 2013].

1. ***min_ evidence*** em 11 faixas: 1, 2, 3, 4, 5, 10, 20, 50, 100, 150 e 200.
2. ***min_ votesweight*** em 8 faixas: 1, 2, 3, 5, 10, 50, 100 e 200.

3. α em 3 faixas: $100km$, $200km$ e ∞km.

Para o método *Mob_ User_ Friends*, fixamos os parâmetros β_{max} e α_{max} em 2.000 Km e 10.000 km^2, respectivamente, conforme discutido no Capítulo 4.

A avaliação da eficácia de nossos modelos de inferência foi feita em torno de duas métricas principais, cobertura e acurácia. A acurácia é dada pela razão entre o número de usuários inferidos corretamente e o total de inferências realizadas pelo método. Já a métrica cobertura representa o total de usuários elegíveis para a inferência pelo método. A cobertura é computada pela soma dos usuários nas duas instâncias de teste (D_1 e D_2).

Uma das limitações da métrica acurácia é que ela não considera, no caso de um erro, a distância entre a localização inferida da localização real do usuário (*ground truth*). Por isso, nas granularidades apresentadas (cidade, bairro e coordenadas geográficas) comparamos a distância das inferências com o *ground truth*.

5.2 Inferência da Cidade de Residência de Usuários

Apresentamos e discutimos nas próximas seções os experimentos realizados com o modelo MVS Ponderado e suas duas variações, Filtrado e Iterativo, assim como com os métodos *Mob_ User_ Friends* e Híbrido. Também comparamos os métodos apresentados com o modelo de referência [Pontes 2013].

5.2.1 Inferência com o Modelo MVS Ponderado e Variações

Nesta seção discutimos os resultados de cada variação do modelo MVS Ponderado, comparando-os com os resultados correspondentes obtidos com os métodos não ponderados propostos por Pontes [2013].

5.2.1.1 MVS Ponderado

Avaliamos o modelo MVS Ponderado nas 11 combinações: *Mayorship+Tip, Mayorship+Like, Mayorship+Friend, Tip+Like, Tip+Friend, Friend+Like, Friend+Tip, Mayorship+Tip+Like, Mayorship+Tip+Friend, Tip+Like+Friend* e *Mayorship+Tip+Like+Friend*. Os resultados são apresentados na Tabela 5.1. A Tabela mostra resultados de acurácia e cobertura para cada classe: classe 0, representa os usuários que possuem apenas uma evidência para a inferência e classe 1, são os usuários que possuem múltiplas evidências para a inferência.

Tabela 5.1. Resultados Obtidos com o Modelo MVS Ponderado na Inferência de Cidade de Residência de Usuários

Atributos	**Cobertura**	**Distribuição dos usuários (%)**		**Acurácia**		
		Classe 0	Classe 1	Classe 0	Classe 1	IC
Mayorship+Tip	2.521.337	35,63	64,37	49,93	63,30	59,70 ± 0,52
Mayorship+Like	2.309.900	35,72	64,28	49.35	62,40	59,41 ± 0,68
Mayorship+Friend	7.013.106	16,79	83,21	32.87	69,78	57,94 ± 0,60
Tip+Like	2.093.119	39,74	60,26	49.50	66,78	58,40 ± 0,52
Tip+Friend	7.082.095	17,16	82,84	33.95	69,47	56,28 ± 0,51
Like+Friend	7.027.402	17,11	82,89	33.04	69,28	55,79 ± 0,55
Mayorship+Tip+Like	2.823.403	33,29	66,71	49.93	70,95	60,19 ± 0,53
Mayorship+Tip+Friend	7.112.548	16,79	83,21	35.50	70,28	60,30 ± 0,51
Mayorship+Like+Friend	7.062.524	16,70	83,30	37.80	70,87	60,41 ± 0,54
Tip+Like+Friend	7.124.687	17,03	82,98	39,84	70,42	60,04 ± 0,67
Mayorship+Tip+Like+Friend	**7.153.077**	16,69	83,31	36.81	71,20	**61,56 ± 0,46**
Baselines MVS [Pontes 2013]						
Mayorship+Tip	2.521.337	35,63	64,37	49,93	65.01	**58,85 ± 0,24**
Mayorship+Tip+like+Friend	**7.153.077**	16,69	83,31	35,60	61,67	**55,67 ± 0,27**

Analisando os resultados apresentados na Tabela 5.1 observamos que o percentual de usuários na classe 1 é sempre maior que 60%, e nas combinações em que aparece o atributo *friend* este valor ultrapassa 80%. Isso ressalta que a maioria dos usuários possui múltiplas evidências. Em todas as combinações, o modelo MVS Ponderado obteve uma acurácia superior a 60% para usuários da classe 1. A cobertura de usuários foi superior a 7 milhões nas combinações em que o atributo *friend* está presente. Especificamente, os dois melhores resultados da cobertura foram nas combinações: *Tip+Like+Friend* e *Mayorship+Tip+Like+Friend*. Já os melhores resultados quanto à acurácia foram para as combinações *Mayorship+Tip+Friend* e *Mayorship+Tip+Like+Friend*. Por isso, destacamos a combinação *Mayorship+Tip+Like+Friend* como o melhor compromisso entre cobertura (7.153.077 usuários) e acurácia (61,56%), totalizando 4.403.077 inferências corretas. Nesta combinação, temos na Tabela 5.2 os pesos obtidos nos 10 folds, respectivamente.

Na maioria dos *folds*, o peso do atributo *friends* foi superior aos demais, evidenciando a importância dos amigos no resultado da inferência. Na sequência, destacamos em ordem de relevância os atributos, *mayorship, tip e like*.

A Tabela 5.1 também mostra os resultados para o modelo de referência com 2 diferentes conjuntos de atributos. Como o modelo de referência não tem nenhum

Tabela 5.2. Pesos dos atributos na combinação *Mayorship+Tip+Like+Friend*

Mayorship (w_m)	**Tip** (w_t)	**Like** (w_l)	**Friend** (w_f)
0,31	0,12	0,13	0,44
0,29	0,19	0,15	0,37
0,37	0,11	0,10	0,62
0,32	0,18	0,12	0,38
0,33	0,17	0,09	0,41
0,35	0,12	0,10	0,43
0,39	0,12	0,11	0,38
0,25	0,14	0,13	0,48
0,35	0,18	0,10	0,37
0,34	0,19	0,09	0,38

parâmetro, não há necessidade de treinamento. Entretanto, para fins de comparação com os nossos modelos, adotamos uma metodologia semelhante, executando o método em cada um dos 10 *folds*, reportando a média da acurácia nos dois conjuntos. Foram feitos experimentos com cada um dos 11 conjuntos de atributos. Verificamos que a melhor acurácia média do modelo MVS foi de 58,85% na combinação dos atributos *Mayorship+Tip* e a melhor cobertura foi obtida com a combinação *Mayorship+Tip+Like+Friend*, resultados mostrados na Tabela 5.1. A combinação *Mayorship+Tip+Like+Friend* também mostrou melhor compromisso entre cobertura e acurácia no modelo MVS, inferindo corretamente 3.981.903 usuários.

Comparando os métodos MVS Ponderado e MVS, observamos que o primeiro resulta em ganhos superiores a 9% em termos de acurácia média quando a combinação *Mayorship+Tip+Like+Friend* é usada. Este resultado demonstra os benefícios de ponderar os diferentes atributos de forma diferenciada.

5.2.1.2 MVS Ponderado Filtrado

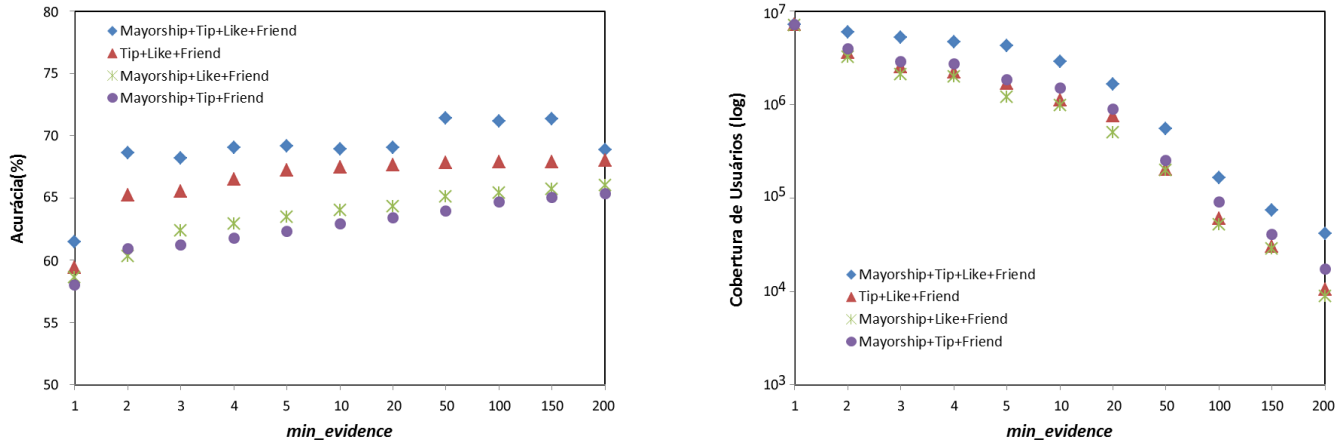
Lembramos que a variação do modelo MVS Ponderado Filtrado é aplicável apenas para usuários com múltiplas evidências para inferência, ou seja, nos usuários da classe 1. Utilizamos as mesmas 11 combinações de atributos nos dois parâmetros, *min_ evidence* e *min_ votesweight*. Para cada uma das 11 combinações de atributos, avaliamos a acurácia e a cobertura do método à medida que variamos os parâmetros *min_ evidence* e *min_ votesweight*. Embora o método seja aplicado apenas aos usuários com múltiplas evidências, nesta seção apresentamos resultados agregados para ambas classes (para usuários da classe 0, a inferência é feita diretamente a partir da única evidência

disponível).

O parâmetro *min_evidence* representa o número mínimo de atributos usados na inferência. Por exemplo, *min_evidence*=20 na combinação *Mayorship+Tip+like+Friend*, significa que consideramos todos os usuários que possuem no mínimo 20 atributos no total.

Os 4 melhores resultados do modelo MVS Ponderado Filtrado com a variação do parâmetro *min_evidence* foram nas combinações: *Mayorship+Tip+Like+Friend*, *Tip+Like+Friend*, *Mayorship+Like+Friend* e *Mayorship+Tip+Friend*. A Figura 5.1(a) mostra a acurácia média das 4 combinações de atributos em função do parâmetro *min_evidence*. À medida que *min_evidence* aumenta, a acurácia da inferência também aumenta, chegando a 71,43% para o cenário *Mayorship+Tip+Like+Friend* com *min_evidence*=50. Logo, a tarefa de inferência torna-se mais precisa quanto maior for o número de atributos existentes. Já a cobertura do método, mostrada na Figura 5.1(b), cai à medida que *min_evidence* decresce. Note a escala logarítmica no eixo Y da Figura 5.1(b). De fato, a cobertura chega a apenas 8.900 usuários no cenário *Mayorship+Like+Friend* quando *min_evidence* é aumentado para 200.

Em nossa avaliação, o cenário *Mayorship+Tip+like+Friend* mostrou ter o melhor compromisso entre cobertura e acurácia, principalmente para *min_evidence=2*.



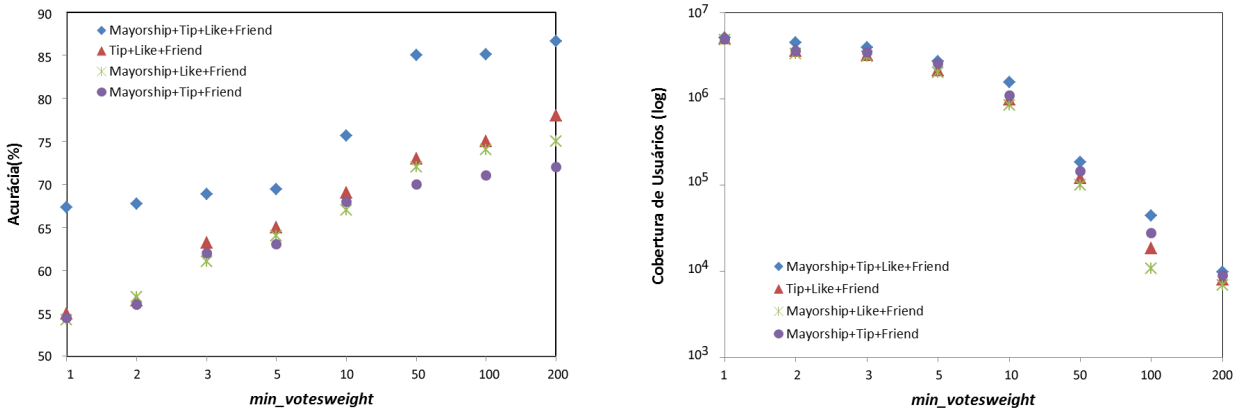
(a) Acurácia *min_evidence* (b) Cobertura *min_evidence*

Figura 5.1. Acurácia e Cobertura dos resultados do Modelo MVS Ponderado Filtrado com o parâmetro *min_evidence*

O parâmetro *min_votesweight* representa a soma total ponderada de votos mínima necessária para eleger uma cidade candidata à inferência. Por exemplo, *min_votesweight*=10, significa que para a cidade ser escolhida no modelo, ela deve possuir uma soma ponderada mínima de votos igual a 10.

Os 4 melhores resultados do modelo MVS Ponderado Filtrado com a variação do parâmetro *min_votesweight* foram os mesmos do parâmetro *min_evidence*: *Mayorship+Tip+Like+Friend*, *Tip+Like+Friend*, *Mayorship+Like+Friend* e *Mayorship+Tip+Friend*. A acurácia média e a cobertura destes métodos em função do parâmetro *min_votesweight* são mostrados nas Figuras 5.2(a) e 5.2(b), respectivamente. Para facilitar a apresentação, o eixo X de cada gráfico apresenta as faixas de valores mínimas de *min_votesweight*. Note, mais uma vez, a escala logarítmica no eixo Y da Figura 5.2(b).

A Figura 5.2(a) mostra que, para as 4 combinações de atributos, a acurácia tende a aumentar à medida em que *min_votesweight* cresce, chegando a 86,65% para o cenário *Mayorship+Tip+Like+Friend* e *min_votesweight*=200. Em contrapartida, como era de se esperar, a cobertura diminui sensivelmente à medida que *min_votesweight* aumenta, restringindo os usuários elegíveis, chegando a apenas 9.780 usuários para *min_votesweight*=200.



(a) Acurácia min_votesweight (b) Cobertura min_votesweight

Figura 5.2. Acurácia e Cobertura dos resultados do Modelo MVS Ponderado Filtrado com o parâmetro *min_votesweight*

A melhor combinação, considerando o compromisso entre cobertura e acurácia foi para a combinação *Mayorship+Tip+Like+Friend* com *min_votesweight*=1.

Assim, a Tabela 5.3 sintetiza os resultados para a combinação de atributos *Mayorship+Tip+Like+Friend* com as variações dos parâmetros *min_evidence* e *min_votesweight*. Percebe-se a relação entre os dois parâmetros, uma vez que ao aumentarmos o valor dos votos ponderados (*min_votesweight*), temos um aumento da acurácia média e quando aumentamos o número de evidências (*min_evidence*), também temos um aumento da acurácia. A acurácia média do modelo MVS Ponderado com o parâmetro *min_votesweight* foi a melhor encontrada, chegando a 86,65%, porém,

Tabela 5.3. Resultados obtidos com o método MVS Ponterado Filtrado na Inferência de Cidade Residência de Usuários - Cenário: *Mayorship+Tip+Like+Friend*

Parâmetro	**Cobertura**	**IC**
min_evidence=1	7.153.077	61,48 ± 0,64
min_evidence=2	5.959.581	68,60 ± 0,61
min_evidence=3	5.224.049	68,19 ± 0,58
min_evidence=4	4.674.879	69,05 ± 0,66
min_evidence=5	4.241.262	69,15 ± 0,62
min_evidence=10	2.893.156	68,95 ± 0,68
min_evidence=20	1.664.757	69,06 ± 0,63
min_evidence=50	546.434	**71,43 ± 0,71**
min_evidence=100	165.584	71,17 ± 0,62
min_evidence=150	74.149	77,35 ± 0,71
min_evidence=200	41.634	68,89 ± 0,66
min_votesweight=1	5.187.320	67,29 ± 0,67
min_votesweight=2	4.540.840	67,75 ± 0,62
min_votesweight=3	3.990.653	68,87 ± 0,63
min_votesweight=5	2.710.606	69,40 ± 0,66
min_votesweight=10	1.565.713	75,67 ± 0,62
min_votesweight=50	185.472	84,97 ± 0,96
min_votesweight=100	44.126	85,11 ± 0,58
min_votesweight=200	9.780	**86,65 ± 0,79**
Baselines [Pontes 2013]		
min_evidence=2	5.959.581	61,22 ± 1,13
min_evidence=50	546.434	**65,52 ± 0,98**
min_votes=2	3.716.122	60,78 ± 0,66
min_votes=200	8.296	**82,86 ± 0,57**

uma das menores coberturas, cerca de 9.780 usuários. A mesma relação ocorre com a cobertura do parâmetro *min_evidence*, onde o número de usuários diminui consideravelmente, iniciamos com 7.153.077 usuários e chegando a 41.634 usuários com *min_evidence=200*, ou seja, menos de 1% do total de usuários. A redução significativa de usuários é esperada, uma vez que não temos muitos usuários com grandes volumes de atividades em nossa base de dados.

A Tabela 5.3 também mostra os resultados obtidos com o método MVS Filtrado original proposto por Pontes [2013], considerando a combinação *Mayorship+Tip+Like+Friend* e os parâmetros *min_evidence=2* e *min_votesweight=2*, temos um aumento superior a 10% e 9% de acurácia média, respectivamente, comparando com os mesmos parâmetros do modelo MVS Ponderado Filtrado.

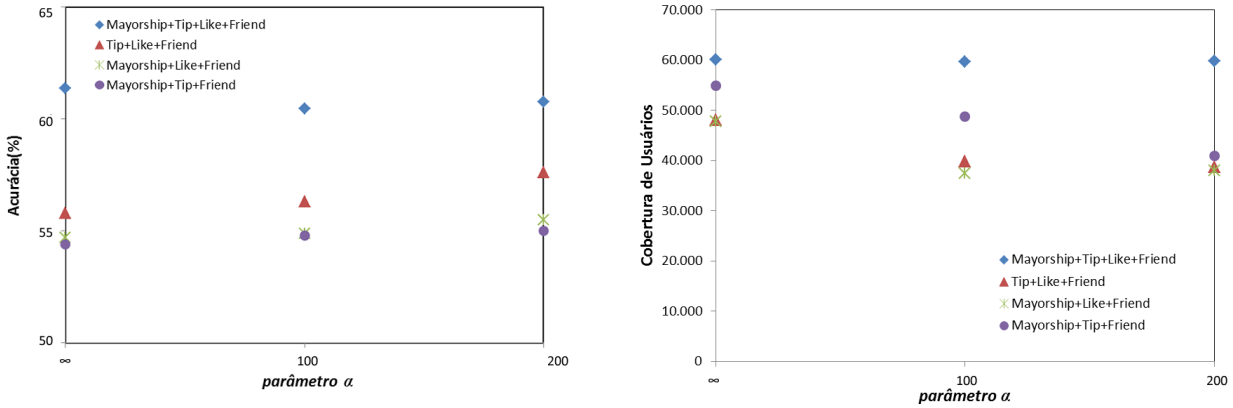
Já, comparando os métodos MVS Ponderado (melhor resultado) com sua versão Filtrada, notamos um aumento de 10% de acurácia média da inferência, ao custo de uma redução (até 17%) de cobertura, quando usamos o parâmetro *min_evidence*=2. Quando usamos o paramêmtro *min_votesweight*=1, os ganhos em acurácia são maiores que 8%, com impacto na redução da cobertura de até 27%.

5.2.1.3 MVS Ponderado Iterativo

Outra variação do modelo MVS Ponderado é a versão Iterativa. O modelo Iterativo vai agregando as evidências (atributos) relacionadas as cidades não selecionadas na inferência às cidades candidatas mais próximas (dado que a distância entre elas não ultrapasse um limiar α, até que, por fim, reste apenas uma alternativa para a inferência. Semelhante ao modelo de referência (Pontes [2013]), em quase todas os experimentos, o resultado também foi obtido com 2 ou no máximo 3 iterações.

Os 4 melhores resultados do modelo MVS Ponderado Iterativo foram com as combinações dos atributos: *Mayorship+Tip+Like+Friend*, *Tip+Like+Friend*, *Mayorship+Like+Friend* e *Mayorship+Tip+Friend*. Os gráficos da Figura 5.3 mostram a acurácia média e a cobertura de cada combinação para os três valores diferentes de α.

Destacamos a combinação dos atributos *Mayorship+Tip+Like+Friend* como o melhor resultado (61,37%), tendo o parâmetro α=100. Identificamos que a variação do parâmetro α não influencia diretamente na cobertura dos usuários, já que para os valores de α=100 *Km*, 200 *Km* ou qualquer valor (∞), quase não houve variação do número de usuários cobertos pelo método.



(a) Acurácia parâmetro α

(b) Cobertura parâmetro α

Figura 5.3. Acurácia e Cobertura dos resultados do Modelo MVS Ponderado Iterativo com variação do parâmetro α

A Tabela 5.4 sumariza os resultados do modelo MVS Ponderado Iterativo para

Tabela 5.4. Resultados obtidos com o modelo MVS Ponderado Iterativo na Inferência de Cidade de Residência de Usuários - Cenário: *Mayorship+Tip+Like+Friend*

Parâmetro α	**Cobertura**	**IC**
100 km	5.971.388	**61,37 ± 0,87**
200 km	5.984.306	60,45 ± 0,70
∞ Km	6.019.011	60,76 ± 1,39
Baseline [Pontes 2013]		
$\alpha = 100$	5.971.388	**56,65 ± 3,35**

a combinação de atributos *Mayorship+Tip+Like+Friend*, comparando-o com a maior acurácia obtida com o modelo de referência MVS Iterativo com $\alpha = 100$ [Pontes 2013]. Note que a estratégia de ponderação dos atributos garante um ganho superior a 7% em acurácia média, além de um aumento modesto na cobertura.

Assim como observado para o modelo de referência MVS [Pontes 2013], o MVS Ponderado Iterativo não leva a ganhos de acurácia sobre as outras duas variações do MVS Ponderado, com a consideração do método Iterativo ser aplicado somente aos usuários da classe 1. A melhor acurácia média do modelo MVS Ponderado Iterativo foi de 61,56% com o parâmetro $\alpha = 100$ e também o melhor compromisso entre cobertura e acurácia. Analisando a acurácia obtida na variação do parâmetro $\alpha = \infty$, podemos observar que as evidências estão em sua maioria a uma distância de até $100km$ da área de residência do usuário.

Destacamos como os melhores resultados do modelo MVS Ponderado e suas variações, aqueles que apresentaram um bom compromisso entre cobertura e acurácia: (i) MVS Ponderado com a combinação dos atributos *Mayorship+Tip+Like+Friend*, com um total de 4.403.077 inferências corretas (melhor resultado); e (ii) MVS Ponderado Filtrado, *min_evidence=2* com um total de 4.088.392 inferências corretas e *min_votesweight=1* com 3.490.651 inferências corretas. (iii) MVS Ponderado Iterativo, $\alpha = 100$ com 3.664.342 inferências corretas.

5.2.2 Inferência com o Modelo *MOB_User_Friends*

O modelo *MOB_User_Friends*, baseado nas áreas de mobilidade do usuário, foi aplicável a 2.304.831 usuários elegíveis, ou seja, cujas áreas de mobilidade são maiores que 0 e menores que 10.000 km^2. Para compreender melhor essa área, buscamos avaliar se usuários que tem amigos que residem na mesma cidade possuem áreas de mobili-

dade semelhantes. Fizemos a análise da sobreposição das áreas de mobilidade para cada par usuário-amigo. Mais de 80% dos pares em que tanto o usuário alvo da inferência quanto seu amigo residem na mesma cidade possuem uma interseção não nula entre suas áreas de mobilidade. Para os casos em que o usuário e seu amigo residem em cidades diferentes, ainda sim 20% dos pares têm uma interseção entre as áreas de mobilidade.

O *MOB_ User_ Friends* utiliza técnicas de agrupamento hierárquico, cuja complexidade espacial normalmente é quadrática $O(N_2)$, porém, utilizamos em sua implementação em Python a biblioteca *fastcluster* ([Müllner 2011]) que otimiza a tarefa de agrupamento de dados vetoriais, possibilitando maior desempenho, principalmente nos casos em que o número de pontos é reduzido, o que ocorre com a maioria dos usuários. A complexidade de tempo do algoritmo *MOB_ User_ Friends* é $O(n^2 \log n)$, sendo uma variação do algoritmo de *cluster* hierárquico com cortes, principalmente na escolha do *cluster* resultante.

A análise experimental do modelo *MOB_ User_ Friends* considerou somente os usuários com mais de uma evidência (*mayorship, tip ou like*) e foram descartados os seguintes casos: (i) usuários que possuíam apenas uma evidência (24,87%) e área igual a 0 (4,14%); (ii) usuários que não possuíam nenhuma interseção com a área de mobilidade de seus amigos (42,46%) e (iii) usuários cuja área de interseção não possuía nenhuma evidência (28,53%).

Aplicamos 2 variações para a inferência com o modelo *MOB_ User_ Friends*: (i) *MVS*: este método consiste na aplicação do critério densidade para escolha do *cluster* representativo da área de mobilidade de um usuário (veja linha 18 no Algoritmo 2) e no uso do modelo MVS original [Pontes 2013] sobre as evidências existentes na área de mobilidade comum entre usuário e seus amigos; e (ii) *MVS Ponderado*: esta técnica consiste no uso do critério densidade ponderada para escolha do *cluster* representativo da área de mobilidade e da técnica MVS Ponderado para inferência na área resultante. Em ambos casos, foi considerada a melhor combinação de atributos, ou seja, *Mayorship+Tip+Like+Friend*. Para a estratégia (ii) utilizamos a técnica de força bruta para atribuir os pesos aos atributos, conforme descrito na Seção 4.3.1. A combinação dos pesos obtidos nas instâncias de treinamento para o modelo foram: $w_m = 0,47$ $w_t = 0,28$ $w_l = 0,25$ e $w_m = 0,40$ $w_t = 0,33$ $w_l = 0,37$.

A tabela 5.5 sumariza os resultados obtidos com as 2 variações do modelo *MOB_ User_ Friends* e ao final o melhor resultado do método de referência [Pontes 2013]. Com oa variação MVS do modelo *MOB_ User_ Friends* a acurácia média foi de 65,71%, cerca de 1.512.487 usuários inferidos corretamente. Já com a variação MVS Ponderado, a acurácia foi de 84,81%, cerca de 1.952.275 usuários inferidos correta-

Tabela 5.5. Resultados obtidos com o Modelo *MOB_ User_ Friends* na Inferência de Cidade de Residência de Usuários

Técnica	**Cobertura**	**IC**
MVS	2.304.831	65,71 ± 0,56
MVS Ponderado	**2.304.831**	**84,81 ± 0,42**
Baselines [Pontes 2013]		
MVS Filtrado (min_ votes=200)	8.296	**82,86 ± 0,57**
MVS (Mayorship+Tip+like+Friend)	**7.153.077**	55,67 ± 0,19

mente.

Comparando a acurácia do modelo *MOB_ User_ Friends* com o melhor resultado da acurácia obtido por Pontes [2013], vimos que a melhoria em acurácia média foi modesta (2%). Porém, analisando a cobertura, o modelo *MOB_ User_ Friends* superou significativamente o modelo de referência, foram mais de 1 milhão de usuário inferidos corretamente.

Comparando os modelos MVS Ponderado (melhor resultado) com o modelo *MOB_ User_ Friends*, notamos aumento de 27% de acurácia média da inferência, ao custo de uma redução de cobertura de até 33%. Apesar do aumento significativo da acurácia do modelo *MOB_ User_ Friends*, houve redução de 45% do percentual de usuários inferidos corretamente em comparação com o modelo MVS Ponderado.

5.2.3 Inferência com o Modelo Híbrido

Os resultados apresentados nas seções anteriores já apresentam ganhos, principalmente comparados ao modelo de referência de Pontes [2013]. Apresentamos o modelo Híbrido como uma alternativa para melhorar o compromisso entre cobertura e acurácia, consequentemente, aumentar o número de usuários inferidos corretamente. Para isso, combinamos os melhores resultados dos modelos propostos, o modelo MVS Ponderado, suas duas variações - Filtrado e Iterativo e o modelo *MOB_ User_ Friends*.

Inicialmente, selecionamos os melhores parâmetros de cada modelo proposto, seguindo o critério de melhor compromisso entre cobertura e acurácia, são eles: (i) MVS Ponderado com a combinação *Mayorship+Tip+Like+Friend*; (ii) MVS Ponderado Filtrado com o parâmetro *min_ evidence=2*; (iii) MVS Ponderado Filtrado com o parâmetro *min_ votesweight=1*; (iv) MVS Ponderado Iterativo com o parâmetro $\alpha = 100km$; e (v) *MOB_ User_ Friends* com a variação MVS Ponderado. Temos então 5 métodos para a inferência com o Modelo Híbrido.

Tabela 5.6. Resultados obtidos com o Modelo Híbrido na Inferência de Cidade de Residência de Usuários

Modelo	**Cobertura**	**IC**
Híbrido	7.153.077	**70,04 ± 0,80**
Baselines		
MVS Ponderado	7.153.077	61,56 ± 0,46
MVS [Pontes 2013]	7.153.077	55,67 ± 0,19

A proposta do modelo Híbrido considera pesos diferentes para cada um dos métodos, obtidos pelo número de acertos no conjunto de validação, assim, temos a seguinte distribuição dos pesos: (i) MVS Ponderado = 0,24; (ii) MVS Ponderado Filtrado $min\_evidence = 0{,}23$; (iii) MVS Ponderado Filtrado $min\_votesweight = 0{,}20$;(iv) MVS Ponderdo Iterativo = 0,21; e (v) $MOB\_User\_Friends = 0{,}12$. O resultado final da inferência consiste na cidade que obteve a maior soma ponderada de votos, considerando os pesos de cada método.

Temos na Tabela 5.6 o resultado do modelo Híbrido e os *baselines*: MVS Ponderado e MVS, ambos com a combinação dos atributos *Mayorship+Tip+Like+Friend*. Comparando os resultados do modelo Híbrido com o modelo MVS Ponderado, temos um ganho de acurácia média superior a 12% e um aumento de 9% do percentual de usuários inferidos corretamente, considerando a mesma cobertura. Já em comparação ao modelo MVS, temos um aumento de 21% de acurácia média e um ganho de 14% com relação ao percentual de usuários inferidos corretamente.

Realizamos a análise dos usuários cujas cidades inferidas foram divergentes das cidades declaradas como residência, para cada um dos modelos apresentados. A Figura 5.4 mostra a distribuição acumulada das distâncias entre cidades declarada e inferida em cada modelo, considerando o melhor resultado de cada um. Note que, para todos os modelos, esta distância é inferior a 20 Km para mais de 40% dos usuários e 100 km para 75% dos usuários. Logo, mesmo para uma parcela significativa dos casos de erro de inferência, a cidade inferida é relativamente próxima da cidade de residência do usuário. Observamos em alguns casos mais extremos de inferências erradas, que existem usuários com poucos atributos (2 ou 3, grande parte *likes* ou *tips*) e poucos amigos (1 ou 2). Sendo que, a maioria do seus atributos e de seus amigos não estão na cidade de residência e nem próximos a ela, o que possivelmente pode apontar que este usuário utiliza o Foursquare para fins de turismo.

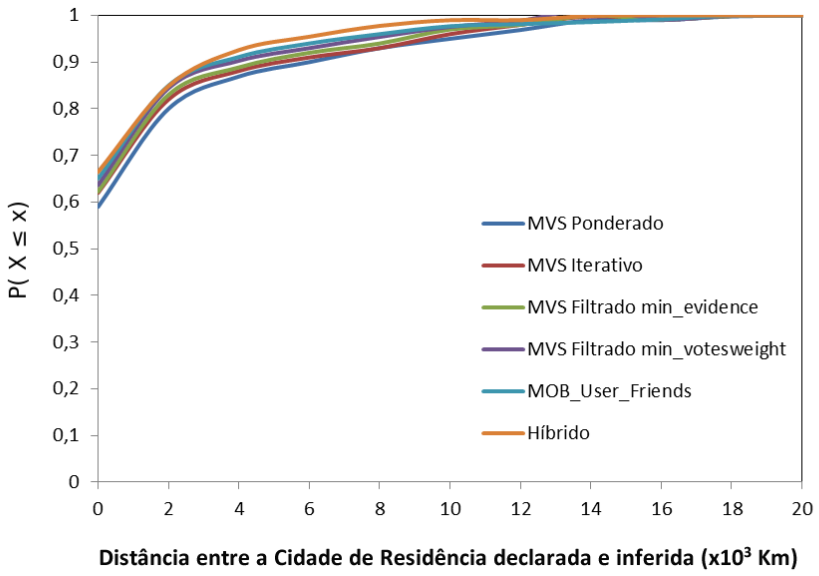


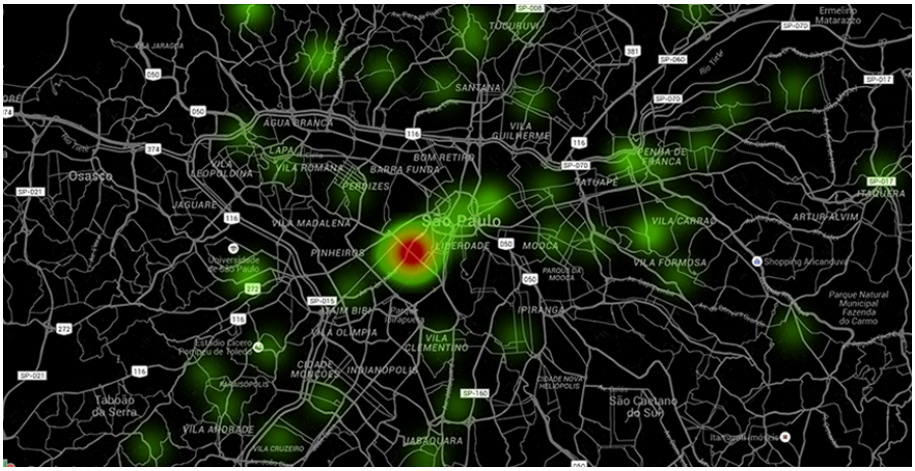
Figura 5.4. Distribuição das distâncias entre as cidades inferidas e as declaradas como local de residência dos usuários

5.3 Inferência do Bairro de Residência de Usuários

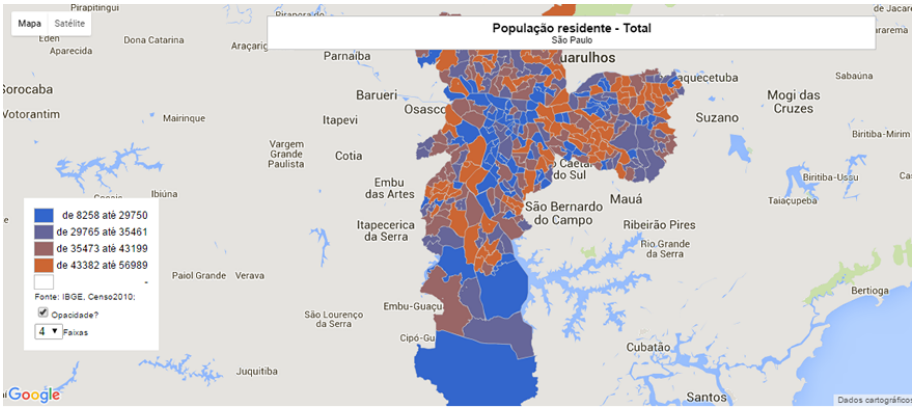
A inferência do bairro de residência dos usuários considerou somente aqueles que possuem *mayorship* associados à categoria *Residence*. Utilizamos neste experimento os usuários da cidade de São Paulo, por representar a maior fração de usuários do Brasil com essa característica. Para realizar essa tarefa, cruzamos a informação das coordenadas geográficas do atributo *home city* com a API do *Google Maps*, assim, foi possível obter o nome do bairro correspondente na cidade de São Paulo. Por exemplo, a entrada das coordenadas: latitude=-25.557906 e longitude=-46.631060, retornou o bairro Liberdade, mais precisamente a Rua Barão de Iguape, 607. O nome do bairro foi considerado o nosso *ground truth*.

Para compreender melhor a distribuição dos usuários que possuem *mayorship* na categoria *Residence* da cidade de São Paulo, as Figuras 5.5(a) e 5.5(b) mostram os mapas de densidade dos *venues* da categoria *Residence* e de densidade populacional, respectivamente, da cidade de São Paulo. O mapa de densidade populacional foi obtido no *site* do Instituto Brasileiro de Geografia e Estatística (IBGE)[1]. Não parece haver uma relação clara entre os mapas de densidade. Ao contrário, parece haver uma relação entre a concentracão de *venues* e a concentração de renda. Percebemos concentrações de *venues* em alguns bairros de regiões sabidamente com maior concentração de renda,

[1]http://www.censo2010.ibge.gov.br/apps/areaponderacao/index.html



(a) Mapa de densidade dos *Venues* da Categoria *Residence* da cidade de São Paulo



(b) Mapa de densidade da População residente na cidade de São Paulo. Fonte: IBGE

Figura 5.5. Comparação da densidade de *Venues* x População da cidade de São Paulo

tais como, Jardim Paulista, Bela Vista e Liberdade. Por outro lado, observamos que as regiões mais carentes possuem número reduzido de *venues*. Por exemplo, a região leste de São Paulo e os bairros Paraisópolis e Vila Água Funda, apresentam grande número de habitantes, porém, um número reduzido de *venues*.

Após a extração dos bairros correspondentes aos usuários que possuem *mayorship* associado à categoria *Residence* da cidade de São Paulo, tivemos um total de 6.344 usuários elegíveis. Inicialmente, aplicamos o modelo MVS Ponderado com as 11 combinações de atributos, como proposto na Seção 4.3.1. Os 2 melhores resultados foram as combinações *Mayorship+Tip+Like* e *Mayorship+Tip+Like+Friend*, apresentados na Tabela 5.7. A combinação *Mayorship+Tip+Like* obteve 60,99% de acurácia média, enquanto que a combinação *Mayorship+Tip+Like+Friend* cerca de 61,65%,

Tabela 5.7. Resultados da Inferência do Bairro de Residência de Usuários da Cidade de São Paulo - Modelo MVS Ponderado

Atributos	**Cobertura**	**IC**
Mayorship+Tip+Like	4.694	60,99 ± 2,43
Mayorship+Tip+Like+Friend	6.344	61,65 ± 0,61
Baseline [Pontes 2013]		
Mayorship+Tip+Like+Friend	6.344	53,50 ± 0,84

mostrando melhor compromisso entre cobertura e acurácia, inferindo corretamente o bairro de residência de 3.911 usuários.

Posteriormente, utilizamos a melhor combinação dos resultados do modelo MVS Ponderado (*Mayorship+Tip+Like+Friend*) para aplicarmos as variações do modelo com Filtro (parâmetros *min_ evidence* e *min_ votesweight*) e Iterativo. Seguimos os mesmos parâmetros de intervalos da inferência de cidade de localização de residência, ou seja, *min_ evidence=(1, 2, 3, 4, 5, 10, 20, 50, 100, 150)* e *min_ votesweight*=(1, 2, 3, 5, 10, 50, 100 e 200). Assim como observado para a inferência de cidade de residência, o aumento dos valores aos parâmetros *min_ evidence* e *min_ votesweight* leva a uma redução considerável na cobertura. Os melhores resultados foram com *min_ evidence*=2 e *min_ votesweight*=1, com acurácia média de 65,08% e 62,46%, respectivamente e cobertura igual a 5.739 e 5.453 usuários, respectivamente. Com o modelo MVS Ponderado Iterativo variamos o parâmetro α nas 3 faixas de valores: 100, 200 e ∞. O melhor resultado foi com $\alpha = \infty$, no qual obtivemos 2,53% de acurácia média e uma cobertura de 5.960 usuários. Já com o modelo *MOB_ User_ Friends*, mais uma vez consideramos as duas variações MVS e MVS Ponderado para realização da inferência. O melhor resultado foi obtido com a variação MVS Ponderado, com acurácia média de 63,89%, porém, esta técnica resultou na menor cobertura entre todos os modelos, cerca 3.100 usuários. O modelo Híbrido combinou os métodos e conseguiu o melhor compromisso entre cobertura e acurácia, totalizando 4.423 usuários inferidos corretamente, com acurácia média de 69,73%. A Tabela 5.8 sintetiza os melhores resultados de cobertura e acurácia apurados para cada um dos modelos.

Considerando o compromisso entre cobertura e acurácia, o modelo Híbrido apresentou melhores resultados na inferência do bairro de residência, superando o modelo MVS Ponderado em mais de 12% de acurácia média. Todos os demais resultados tiveram uma acurácia superior a 60%, chegando a 69,73% com o modelo Híbrido. Comparando os resultados do Modelo Híbrido com o modelo de referência (MVS), observamos que houve um aumento de mais de 23% na acurácia, considerando a mesma cobertura de usuários.

Tabela 5.8. Resultados obtidos na Inferência do Bairro de Residência de Usuários da Cidade de São Paulo

Modelo	**Cobertura**	**IC)**
MVS Ponderado (*Mayorship+Tip+Like+Friend*)	6.344	61,65 ± 0,61
MVS Ponderado Filtrado *min_evidence=2*	5.739	65,08 ± 0,77
MVS Ponderado Filtrado *min_votesweight=1*	5.453	62,46 ± 1,59
MVS Ponderado Iterativo $\alpha = \infty km$	5.960	62,53 ± 1,16
MOB_User_Friends - MVS Ponderado	3.100	63,89 ± 1,44
Híbrido	**6.344**	**69,73 ± 0,93**
Baseline [Pontes 2013]		
Modelo MVS (Mayorship+Tip+Like+Friend)	6.344	53,50 ± 0,84

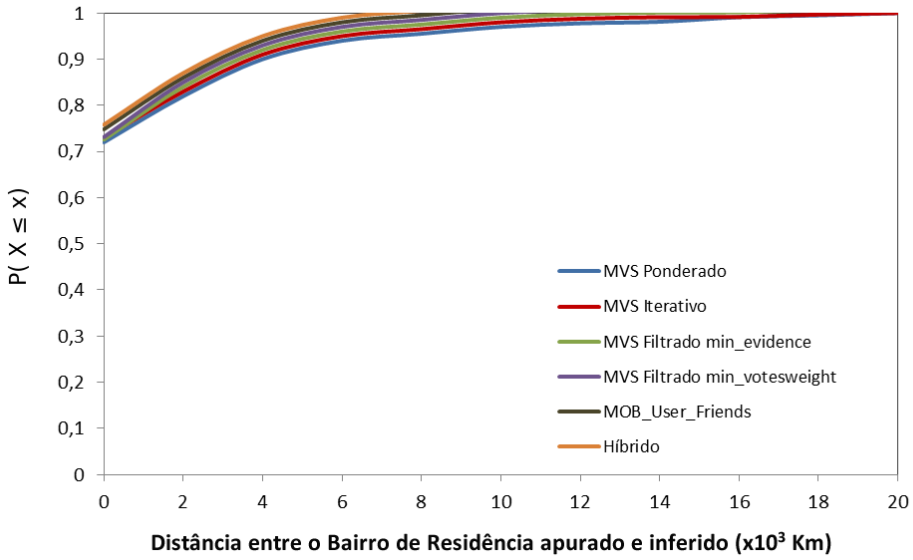


Figura 5.6. Distribuição das distâncias entre os bairros inferidos e os declarados como local de residência dos usuários

Por fim, assim como feito para as inferências de cidade, também analisamos os casos de inferências incorretas, quantificando as distâncias entre os bairros inferidos e declarados. A Figura 5.6 mostra a distribuição acumulada destas distâncias para todos os métodos. Note que, para todas as estratégias, foi possível inferir com uma precisão de até $2Km$ o bairro de residência de cerca de 60% dos usuários. Mais ainda, os modelos *MOB_User_Friends* e Híbrido realizaram inferências com uma distância de até $5Km$ (razoável para uma grande metrópole como São Paulo) para cerca de 76% e 74% dos usuários, respectivamente. Estas distribuições demonstram a eficácia dos

modelos propostos.

5.4 Inferência das Coordenadas Geográficas de Residência de Usuários

Semelhante à inferência de localização na granularidade de bairro, selecionamos para a inferência das coordenadas geográficas somente os usuários que apresentam *mayorhsip* associados à categoria *Residence*. Utilizamos a API do *Google Maps* para extrair as coordenadas geográficas dos *venues* associados à categoria *Residence*, obtendo 832.191 usuários elegíveis para a inferência.

Inicialmente, aplicamos o modelo MVS Ponderado nas 11 combinações de atributos, como proposto na Seção 4.3.1. A Tabela 5.9 mostra os resultados para as duas melhores combinações *Mayorship+Tip+Like* e *Mayorship+Tip+Like+Friend*, assim como para o modelo de referência MVS que produziu melhores resultados de inferência nesta granularidade. A Tabela 5.9 mostra a cobertura assim como a acurácia para diferentes limiares de distância máxima entre as coordenadas inferidas e o local de residência. Note que, comparando as acurácias das duas variações do MVS Ponderado, aquela que explora o atributo $friend$ é superior apenas quando consideramos distâncias maiores (maiores que 5 Km). Para distâncias até 5 Km, utilizar apenas os atributos *mayorship, tip* e *like* leva a uma acurácia ligeiramente melhor, ou seja, a inclusão dos amigos pode trazer ruído para as inferências. Entretanto, ela também contribui para uma cobertura bem maior (832 mil versus 695 mil usuários). Considerando o número de inferências corretas, temos que a combinação *Mayorship+Tip+Like+Friend* produziu melhores resultados. Comparando ambas as estratégias com o modelo de referência (MVS com a combinação dos atributos *Mayorship+Tip+Like*), mais uma vez observamos que a ponderação dos atributos trouxe ganhos significativos: o número de inferências corretas (considerando uma mesma distância máxima) aumentou em até 39% e 7% para distâncias máximas iguais a 1 e 5Km, respectivamente.

Destacamos a combinação dos atributos *Mayorship+Tip+Like+Friend* do modelo MVS Ponderado para aplicar as duas variações: com Filtro (parâmetros *min_ evidence* e *min_ votesweight*) e Iterativo. Seguimos os mesmos intervalos da inferência das granularidades de cidade e bairro, ou seja, *min_ evidence=(1, 2, 3, 4, 5, 10, 20, 50, 100, 150)* e *min_ votesweight*=(1, 2, 3, 5, 10, 50, 100 e 200). Os melhores resultados foram apurados com as combinações dos parâmetos *min_ evidence=2* e *min_ votesweight=1*. Já com o MVS Ponderado Iterativo variamos o parâmetro α nas 3 faixas de valores: 100, 200 e ∞. O melhor resultado foi obtido com $\alpha = \infty$. Com o modelo

Tabela 5.9. Resultados da Inferência das Coordenadas Geográficas - Modelo MVS Ponderado

Atributos	**Cobertura**	**IC**
Mayorship+Tip+Like	695.012	até 1 *Km* - 31,37 ± 0,97
		até 5 *Km* - 65,78 ± 0,50
		até 20 *Km* - 81,87 ± 0,43
		até 50 *Km* - 85,70 ± 0,56
		até 100 *Km* - 90,73 ± 0,43
Mayorship+Tip+Like+Friend	832.191	até 1 *Km* - 29,55 ± 0,96
		até 5 *Km* - 65,36 ± 1,07
		até 20 *Km* - 89,39 ± 0,74
		até 50 *Km* - 92,94 ± 0,43
		até 100 *Km* - 98,31 ± 1,09
Baseline [Pontes 2013]		
Mayorship+Tip+Like	695.012	até 1 *Km* - 22,63 ± 0,35
		até 5 *Km* - 60,93 ± 0,58
		até 20 *Km* - 63,48 ± 0,64
		até 50 *Km* - 68,19 ± 0,41
		até 100 *Km* - 69,83 ± 0,43

MOB_User_Friends, a versão que utiliza o modelo MVS Ponderado obteve acurácia média de 65,43% em até 5 *Km*, porém, com a menor cobertura, cerca de 65% do total de usuários elegíveis para o método. O modelo Híbrido foi o que teve melhor compromisso entre cobertura e acurácia, totalizando 558.539 usuários inferidos a uma distância de até 5 *km* das coordenadas geográficas de residência do usuário. A Tabela 5.10 sumariza os resultados obtidos nos modelos avaliados, a uma distância de até 5 *km* das coordenadas geográficas de residência dos usuários. Já a Figura 5.7 mostra as distribuições acumuladas destas distâncias para todos os métodos.

Enfim, assim como observado para a inferência da cidade de residência, destacamos nas granularidades bairro e coordenadas geográficas o modelo Híbrido, como aquele que produziu os melhores resultados do compromisso entre cobertura e acurácia.

5.5 Análise Geral dos Resultados

Ao investigar a melhor combinação dos atributos utilizando o modelo MVS Ponderado, vimos que os amigos do usuário podem ser bons indicadores a serem utilizados na inferência de localização de residência, como na combinação *Mayor-*

Tabela 5.10. Resultados da Inferência de Localização das Coordenadas Geográficas de Residência de Usuários do Foursquare

Modelo	**Cobertura**	**IC de até 5*Km***
MVS Ponderado (*Mayorship+Tip+Like+Friend*)	832.191	65,36 ± 1,07
MVS Ponderado Filtrado *min_evidence=2*	698.765	65,40 ± 0,82
MVS Ponderado Filtrado *min_votesweight=1*	697.531	65,15 ± 1,47
MVS Ponderado Iterativo $\alpha = \infty km$	743.765	66,11 ± 1,21
MOB_User_Friends	538.264	65,43 ± 0,77
Híbrido	**832.191**	**67,12 ± 1,32**
Baseline [Pontes 2013]		
Modelo MVS (Mayorship+Tip+Like)	695.012	60,93 ± 0,58

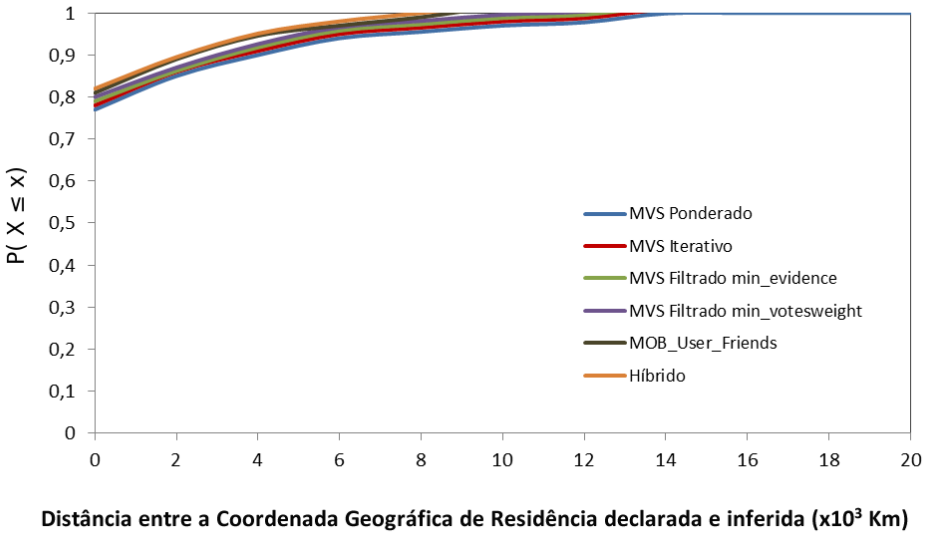


Figura 5.7. Distribuição das distâncias entre as coordenadas geográficas inferidas e a localização de residência exata dos usuários

ship+Tip+Like+Friend, o peso do atributo *friend* ficou mais evidente. Destacamos o modelo *MOB_User_Friends* que explora essa relação de amizade, através da interseção entre as áreas de mobilidade do usuário e seus amigos. Conseguimos, com isso, uma das maiores acurácias médias, 84,81% de usuários inferidos corretamente na granularidade de cidade de residência. Infelizmente este modelo também leva ao descarte de um grande número de usuários, o que impacta consideravelmente a sua cobertura.

Com o modelo Híbrido, abordamos uma solução para manter um bom compro-

Tabela 5.11. Análise Geral dos Resultados Obtidos na Inferência de Cidade de Residência

Modelo	**Cobertura**	**IC**	**# Acertos**
MVS Pond. *(Mayorship+Tip+Like+Friend)*	7.153.077	61,56 ± 0,46	4.395.923
MVS Pond. Filtrado (*min_evidence=2)*	5.959.581	68,60 ± 0,61	4.088.392
MVS Pond. Filtrado *(min_votesweight=1)*	5.987.320	67,29 ± 0,67	3.490.651
MVS Pond. Iterativo *($\alpha = 100$)*	5.971.388	61,37 ± 0,87	3.664.342
MOB_User_Friends (MVS Ponderado)	2.304.831	**84,81 ± 0,42**	1.952.275
Modelo Híbrido	7.153.077	**70,04 ± 0,57**	**5.010.280**
Baseline [Pontes 2013]			
MVS (Mayorship+Tip+like+Friend)	7.153.077	55,67 ± 0,19	3.981.903

misso entre cobertura e acurácia. Essa técnica combina todos os métodos apresentados ponderando a inferência realizada por cada um segundo o número de acerto no conjunto de validação. Buscamos, desse modo, agregar o que cada modelo explora de melhor.

A Tabela 5.11 sintetiza os melhores resultados de cada modelo de inferência da cidade de residência dos usuários da rede social *online* Foursquare. Destacamos que estes resultados apresentam os melhores compromissos entre cobertura e acurácia. Elegemos o modelo Híbrido como o melhor resultado: ele consegue inferir corretamente a cidade de residência de 5.010.280 usuários. Comparado ao modelo MVS de referência ([Pontes 2013]), o modelo Híbrido obteve um ganho na acurácia maior que 21%, com a mesma cobertura de usuários (7.153.077).

Os resultados também foram satisfatórios para as granularidades de bairro e coordenadas geográficas. O modelo Híbrido inferiu corretamente o bairro de 4.423 (69% dos usuários elegíveis) usuários da cidade de São Paulo, sendo este o melhor compromisso entre cobertura e acurácia. Já na granularidade de coordenadas geográficas, o modelo Híbrido inferiu 2.413 (29% dos usuários elegíveis) usuários a uma distância de até 1 *Km* do local exato de residência e 558.539 usuários (67% dos usuários elegíveis) a uma distância de até 5 *Km*.

5.6 Limitações

Nesta seção destacamos algumas limitações do trabalho realizado.

1. A base de dados utilizada nesta pesquisa foi coletada em 2011 e portanto representa uma fotografia do sistema na época, o que pode não refletir a realidade atual. Além disto a base de dados não inclui um atributo importante que poderia ajudar na tarefa de inferência, os *check-ins* dos usuários.

2. Os dados analisados podem apresentar falsos usuários, ou *spammers*, o que pode provocar inferências incorretas.

3. Apesar do grande volume de dados coletados e analisados nos experimentos, nas granularidades como bairro e coordenadas geográficas foi possível filtrar apenas uma pequena amostra do conjunto de dados, devido à especificidade da seleção, que exige que o usuário possua *mayorship* na categoria *Residence*.

Apesar das limitações existentes, acreditamos que os resultados são promissores e mostram que é possível executar a tarefa de inferência com dados públicos do Foursquare. Um trabalho futuro interessante consiste em avaliar soluções para esta mesma tarefa para outras redes sociais *online*.

Capítulo 6

Conclusões e Trabalhos Futuros

Ao explorar a rede social *online* Foursquare vimos que *mayorships, tips, likes* e lista de amigos representam o comportamento do usuário e podem revelar a cidade de residência do mesmo. Inferir a localização do usuário é atualmente uma tarefa primordial para sistemas que trabalham com recomendação e *marketing*, seja na granularidade de cidade, bairro, região e até coordenadas geográficas exatas. Os modelos de inferência também podem ser aplicados na previsão de mobilidade, logo inferir onde o usuário se encontra em um dado momento possui várias aplicações, por exemplo, acionar sistemas de recomendação, detecção de eventos e catástrofes.

Neste trabalho, abordamos a inferência com foco na localização de residência de usuários, utilizamos para essa tarefa a rede social *online* Foursquare. Propomos e avaliamos métodos baseados no modelo MVS, com a utilização de pesos e nas áreas de mobilidade do usuário e de sua rede de amizade. Também combinamos as técnicas propostas em um modelo Híbrido, obtendo o melhor resultado apresentado.

No modelo MVS Ponderado usamos a técnica de força bruta para obter os pesos de cada atributo em 11 combinações propostas. A combinação dos atributos *Mayorship+Tip+Like+Friend* foi a que apresentou melhor compromisso entre cobertura (7.153.077 usuário) e acurácia (70,04%), superando em 21% a acurácia média do modelo de referência [Pontes 2013].

As duas variações do modelo MVS Ponderado, Filtrado (*min_evidence* e *min_votesweight*) e Iterativo também apresentaram ganhos na acurácia média em comparação aos modelos de referência [Pontes 2013], 11%, 10% e 8% , respectivamente. A estratégia de restringir o número de evidências na inferência reduziu o número de usuários elegíveis, porém, aumentou a acurácia média, chegando a 77,43% com *min_evidence=50* e 86,65% com *min_votesweight*=200.

Já com o modelo *MOB_User_Friends* usamos a área de interseção de mobilidade

do usuário com seus amigos para inferir a localização de residência. A acurácia média do modelo foi uma das melhores (84,81%), com redução de cobertura (2.301.831 usuários). Em comparação ao melhor resultado obtido por Pontes [2013], houve um aumento de 2% de acurácia média, porém, um ganho de mais de 1 milhão de usuário inferidos corretamente e um aumento de 99% na cobertura.

O modelo Híbrido equilibrou o compromisso entre cobertura e acurácia, sendo que, o número de usuários inferidos corretamente chegou a mais de 5 milhões. Em comparação ao resultado do modelo de referência MVS, houve um aumento de 21% na acurácia média, considerando a mesma cobertura.

Além dos resultados apresentados para a granularidade de cidade, também mostramos que é possível inferir o bairro e as coordenadas geográficas exatas de onde o usuário reside. Na inferência do bairro, selecionamos os usuários da cidade de São Paulo que possuem *mayorship* na categoria *Residence*, tivemos uma acurácia média de 69,73% e uma cobertura de 6.344 usuários. Já, para a granularidade de coordenadas geográficas foi possível uma acurácia média de 67,12% em até 5 Km de distância do local exato de residência do usuário.

Portanto, os resultados mostram a possibilidade de inferir o local de residência de usuários em uma rede social *online*, utilizando dados públicos. Com isso, percebemos que a exposição de dados públicos pode levar a uma maior vulnerabilidade, como exemplo: possuir um *mayorship* em um *venue* da categoria *Residence*. Contudo, além do foco da privacidade, a inferência de localização de residência pode ser aplicada em outras áreas, como em sistemas de recomedação.

Por fim, apresentamos modelos de inferência de localização de residência de usuários, aplicados em 3 granularidades (cidade, bairro e coordenadas geográficas) com dados públicos da rede social *online* Foursquare, considerando a base de dados de nível mundial. Os resultados, comparados aos analisados na literatura, apresentam melhor compromisso entre cobertura e acurácia, destacando esses novos modelos como opções eficientes para a tarefa de inferência.

Aplicações e Trabalhos Futuros

Uma das possíveis aplicações dos modelos de inferência apresentados neste trabalho estão relacionadas ao controle da privacidade do usuário. Exemplo, uma ferramenta capaz de monitorar a rede social *online* do usuário e informar vulnerabilidades quanto à detecção do local de residência. Por outro lado, os modelos de inferência de localização de residência de usuários poderão ser aplicados na detecção e monitoramento de eventos, como catástrofes naturais, epidemias, trânsito, jogos olímpicos, eleições, entre

outros.

Uma das principais aplicações dos modelos de inferência estão relacionadas aos sistemas de recomendação, como por exemplo, sistemas de recomendação de pessoas que sejam aptas a falar sobre um determinado lugar. Uma outra abordagem sobre a recomendação são os sistemas de terceiros que fazem uso dos dados obtidos das redes sociais *online* como Foursquare, Twitter, Facebook e Instagram, podem utilizar os modelos de inferência para criar sistemas de *marketing* digital (recomendação direta de um produto ou serviço), anúncios personalizados e até mesmo recomendar pessoas para se relacionar: amizade e namoro.

As aplicações são diversas, por isso, acreditamos que há possibilidade de evolução e continuidade dessa pesquisa. Podemos adaptar os modelos de inferência para serem aplicados à previsão de mobilidade do usuário. Explorar outros atributos, como os *check-ins* e o conteúdo textual das redes sociais *online*; o que outros autores já fizeram e que contribui para a inferência de usuários que não possuem muitas evidências geo-localizadas. Outra possibilidade é incorporar a questão temporal na inferência, desse modo, seria possível separar os usuários turistas, por exemplo. A agregação temporal poderia ser útil para os ajustes de parâmetros, como o α_max e β_max, observando a concentração x espalhamento dos *venues* ao longo do tempo.

Outros parâmetros também podem ser melhor ajustados, como o α, utilizado no modelo MVS Ponderado Iterativo. Algumas ideias podem ser aplicadas neste ajuste,como considerar o tamanho da cidade e a população no cálculo do parâmetro. Os modelos apresentados também podem ser combinados, criando novas variações, como exemplo, aplicar o modelo MVS na áreas de usuários que não possuem nenhuma interseção no modelo *MOB_User_Friends*.

Por fim, pretendemos validar nossos métodos de inferência de localização de residência em outras redes sociais *online* e comprovar os bons resultados alcançados.

Referências Bibliográficas

Ahn, G.-J.; Shehab, M. & Squicciarini, A. (2011). Security and privacy in social networks. volume 15, pp. 10–12.

Allamanis, M.; Scellato, S. & Mascolo, C. (2012). Evolution of a location-based online social network: Analysis and models. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, pp. 145--158, New York, NY, USA. ACM.

Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185.

Anthonysamy, P.; Rashid, A.; Walkerdine, J.; Greenwood, P. & Larkou, G. (2012). Collaborative privacy management for third-party applications in online social networks. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pp. 5:1--5:4, New York, NY, USA. ACM.

Barnes, A. (2006). A privacy paradox: Social networking in the united states. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, New York, NY, USA. ACM.

Bauer, S.; Noulas, A.; Seaghdha, D. O.; Clark, S. & Mascolo, C. (2012). Talking places: Modelling and analysing linguistic content in foursquare. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pp. 348--357, Washington, DC, USA. IEEE Computer Society.

BBC (2010). Obama twitter account 'hacked by frenchman'.

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J. & Gonçalves, M. (2008). Detectando usuários maliciosos em interações via vídeos no youtube. In *Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia)*, Vila Velha, Brasil.

Bilogrevic, I.; Huguenin, K.; Mihaila, S.; Shokri, R. & Hubaux, J. (2015). Predicting users' motivations behind location check-ins and utility implications of privacy protection mechanisms. In *22nd Annual Network and Distributed System Security Symposium, (NDSS) 2015, San Diego, California, USA, February 8-11, 2014*.

Brown, C.; Nicosia, V.; Scellato, S.; Noulas, A. & Mascolo, C. (2012). The importance of being placefriends: Discovering location-focused online communities. In *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks*, WOSN '12, pp. 31--36, New York, NY, USA. ACM.

Brown, C.; Noulas, A.; Mascolo, C. & Blondel, V. (2013). A place-focused model for social networks in cities. In *Proceedings of the 2013 International Conference on Social Computing*, SOCIALCOM '13, pp. 75--80, Washington, DC, USA. IEEE Computer Society.

Bui, N.; Bui, N.; Michelinakis, F.; Michelinakis, F. & Widmer, J. (2014). A model for throughput prediction for mobile users. In *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, pp. 1–6.

Burke, J.; Estrin, D.; Hansen, M.; Parker, A.; Ramanathan, N.; Reddy, S. & Srivastava, M. B. (2006). Participatory sensing. In *In: Workshop on World-Sensor-Web (WSW?06): Mobile Device Centric Sensor Networks and Applications*, pp. 117--134.

Chang, H.-w.; Lee, D.; Eltaher, M. & Lee, J. (2012). @phillies tweeting from philly? predicting twitter user locations with spatial word usage. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, ASONAM '12, pp. 111--118, Washington, DC, USA. IEEE Computer Society.

Cheng, Z.; Caverlee, J. & Lee, K. (2010). You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pp. 759--768, New York, NY, USA. ACM.

Cheng, Z.; Caverlee, J.; Lee, K. & Sui, D. Z. (2011). Exploring millions of footprints in location sharing services. In Adamic, L. A.; Baeza-Yates, R. A. & Counts, S., editores, *ICWSM*. The AAAI Press.

Cho, E.; Myers, S. A. & Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining, KDD '11, pp. 1082--1090, New York, NY, USA. ACM.

Danah & Marwick, A. (2011). Social privacy in networked publics: Teens' attitudes, practices, and strategies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, New York, NY, USA. Privacy Law Scholars Conference.

Davis Jr., C. A.; Pappa, G. L.; de Oliveira, D. R. R. & de L. Arcanjo, F. (2011). Inferring the location of twitter messages based on user relationships. volume 15, pp. 735--751. Blackwell Publishing Ltd.

de Souza, L. A.; Delboni, T. M.; Borges, K. A. V.; Davis, C. A. & Laender, A. H. F. (2004). Locus: Um localizador espacial urbano. In Iochpe, C. & Câmara, G., editores, *GeoInfo*, pp. 467–478. INPE.

Dong, W.; Duffield, N. G.; Ge, Z.; Lee, S. & Pang, J. (2013). Modeling cellular user mobility using a leap graph. In Roughan, M. & Chang, R. K. C., editores, *PAM*, volume 7799 of *Lecture Notes in Computer Science*, pp. 53–62. Springer.

Doty, N. & Wilde, E. (2010). Geolocation privacy and application platforms. In *Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Security and Privacy in GIS and LBS*, SPRINGL '10, pp. 65--69, New York, NY, USA. ACM.

EC-Council (2009). *Cyber Safety*. Cengage Learning.

Facebook (2004). Facebook.

Filho, R. M.; Almeida, J. M. & Pappa, G. L. (2015). Twitter population sample bias and its impact on predictive outcomes: A case study on elections. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, ASONAM '15, pp. 1254--1261, New York, NY, USA. ACM.

Fitz-Walter, Z.; Tjondronegoro, D. & Wyeth, P. (2011). Orientation passport: using gamification to engage university students. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference*, OzCHI '11, pp. 122--125, New York, NY, USA. ACM.

Foursquare (2009). Foursquare.

Fujioka, A.; Okamoto, T. & Ohta, K. (1993). *Advances in Cryptology — AUSCRYPT 92: Workshop on the Theory and Application of Cryptographic Techniques Gold Coast, Queensland, Australia, December 13–16, 1992*, chapter A practical secret voting scheme for large scale elections.

Gao, H.; Tang, J. & Liu, H. (2012). Exploring social-historical ties on location-based social networks.

Ghosh, S.; Viswanath, B.; Kooti, F.; Sharma, N. K.; Korlam, G.; Benevenuto, F.; Ganguly, N. & Gummadi, K. P. (2012). Understanding and combating link farming in the twitter social network. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pp. 61--70, New York, NY, USA. ACM.

Gomide, J.; Veloso, A.; Meira, Jr., W.; Almeida, V.; Benevenuto, F.; Ferraz, F. & Teixeira, M. (2011). Dengue surveillance based on a computational model of spatio-temporal locality of twitter. In *Proceedings of the 3rd International Web Science Conference*, WebSci '11, pp. 3:1--3:8, New York, NY, USA. ACM.

Gundecha, P.; Barbier, G. & Liu, H. (2011). Exploiting vulnerability to secure user privacy on a social networking site. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 511--519, New York, NY, USA. ACM.

Hanne, M.; Silva, C.; Almeida, J. & Gonçalves, M. (2012). Analysis of vulnerability to facebook users. In *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web*, WebMedia '12, pp. 335--342, New York, NY, USA. ACM.

Jin, L.; Long, X. & Joshi, J. B. (2012). Towards understanding residential privacy by analyzing users' activities in foursquare. In *Proceedings of the 2012 ACM Workshop on Building Analysis Datasets and Gathering Experience Returns for Security*, BADGERS '12, pp. 25--32, New York, NY, USA. ACM.

Junior, M. P.; Xavier, S. I. d. R. & Prates, R. O. (2014). Investigating the use of a simulator to support users in anticipating impact of privacy settings in facebook. In *Proceedings of the 18th International Conference on Supporting Group Work*, GROUP '14, pp. 63--72, New York, NY, USA. ACM.

Karamshuk, D.; Noulas, A.; Scellato, S.; Nicosia, V. & Mascolo, C. (2013). Geo-spotting: Mining online location-based services for optimal retail store placement. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 793--801, New York, NY, USA. ACM.

Kong, L.; Liu, Z. & Huang, Y. (2014). Spot: Locating social media users based on social network context. *Proc. VLDB Endow.*, 7(13):1681--1684.

Kriesel, D. (2007). *A Brief Introduction to Neural Networks.*

Krishnamurthy, Balachander e Wills, C. E. (2009). On the leakage of personally identifiable information via online social networks. In *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, WOSN '09, pp. 7--12, New York, NY, USA. ACM.

Leontiadis, I.; Efstratiou, C.; Picone, M. & Mascolo, C. (2012). Don't kill my ads!: Balancing privacy in an ad-supported mobile application market. In *Proceedings of the Twelfth Workshop on Mobile Computing Systems & Applications*, HotMobile '12, pp. 2:1--2:6, New York, NY, USA. ACM.

Li, C.-T. & Hsieh, H.-P. (2015). Geo-social media analytics. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pp. 1533--1534, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Li, H. & Sun, J. (2009). Majority voting combination of multiple case-based reasoning for financial distress prediction. *Expert Systems with Applications*, 36(3, Part 1):4363 – 4373.

Lindqvist, J.; Cranshaw, J.; Wiese, J.; Hong, J. & Zimmerman, J. (2011). I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pp. 2409--2418, New York, NY, USA. ACM.

Liu, W.; Rahman, M. F.; Thirumuruganathan, S.; Zhang, N. & Das, G. (2015). Aggregate estimations over location based services. volume abs/1505.02441.

Long, X.; Jin, L. & Joshi, J. (2012). Exploring trajectory-driven local geographic topics in foursquare. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pp. 927--934, New York, NY, USA. ACM.

Luo, W.; Xie, Q. & Hengartner, U. (2009). Facecloak: An architecture for user privacy on social networking sites. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 3, pp. 26–33.

Mahmud, J.; Nichols, J. & Drews, C. (2012). Where is this tweet from? inferring home locations of twitter users.

Mahmud, J.; Nichols, J. & Drews, C. (2014). Home location identification of twitter users. volume 5, pp. 47:1--47:21, New York, NY, USA. ACM.

Malin, B. (2005). Betrayed by my shadow: learning data identity via trail matching. volume 2005.

Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. *CoRR*, abs/1109.2378.

Noulas, A.; Mascolo, C. & Frias-Martinez, E. (2013). Exploiting foursquare and cellular data to infer user activity in urban environments. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management - Volume 01*, MDM '13, pp. 167--176, Washington, DC, USA. IEEE Computer Society.

Noulas, A.; Scellato, S.; Lambiotte, R.; Pontil, M. & Mascolo, C. (2012a). A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5):e37027.

Noulas, A.; Scellato, S.; Lathia, N. & Mascolo, C. (2012b). A random walk around the city: New venue recommendation in location-based social networks. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pp. 144--153, Washington, DC, USA. IEEE Computer Society.

Noulas, A.; Scellato, S.; Mascolo, C. & Pontil, M. (2011a). An empirical study of geographic user activity patterns in foursquare. In Adamic, L. A.; Baeza-Yates, R. A. & Counts, S., editores, *ICWSM*. The AAAI Press.

Noulas, A.; Scellato, S.; Mascolo, C. & Pontil, M. (2011b). Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *The Social Mobile Web*, volume WS-11-02 of *AAAI Workshops*. AAAI.

Page, X.; Kobsa, A. & Knijnenburg, B. (2012). Don't disturb my circles! boundary preservation is at the center of location-sharing concerns.

Pesce, J. a. P.; Casas, D. L.; Rauber, G. & Almeida, V. (2012). Privacy attacks in social media using photo tagging networks: A case study with facebook. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pp. 4:1--4:8, New York, NY, USA. ACM.

Pontes, T. (2013). Inferência da localização de residência de usuários de redes sociais a partir de dados públicos. Dissertação, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais.

Pontes, T.; Magno, G.; Vasconcelos, M.; Gupta, A.; Almeida, J.; Kumaraguru, P. & Almeida, V. (2012a). Beware of what you share: Inferring home location in social networks. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops*, ICDMW '12, pp. 571--578, Washington, DC, USA. IEEE Computer Society.

Pontes, T.; Vasconcelos, M.; Almeida, J.; Kumaraguru, P. & Almeida, V. (2012b). We know where you live: privacy characterization of foursquare behavior. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, pp. 898--905, New York, NY, USA. ACM.

Quercia, D.; Casas, D. L.; Pesce, J. P.; Stillwell, D.; Kosinski, M.; Almeida, V. & Crowcroft, J. (2012). Facebook and privacy: The balancing act of personality, gender, and relationship currency.

Rauber, G.; Almeida, V. & Kumaraguru, P. (2011). Privacy Albeit Late.

Ribeiro, Jr., S. S. (2015). Integração de informações de usuários para inferência de localização geográfica no twitter. Dissertação, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais.

Ribeiro, Jr., S. S.; Davis, Jr., C. A.; Oliveira, D. R. R.; Meira, Jr., W.; Gonçalves, T. S. & Pappa, G. L. (2012). Traffic observatory: A system to detect and locate traffic events and conditions using twitter. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks*, LBSN '12, pp. 5--11, New York, NY, USA. ACM.

Rodrigues, E.; Assunção, R.; Pappa, G. L.; Renno, D. & Jr., W. M. (2015). Exploring multiple evidence to infer users' location in twitter. pp. –.

Rossi, L. & Musolesi, M. (2014). It's the way you check-in: Identifying users in location-based social networks. In *Proceedings of the Second ACM Conference on Online Social Networks*, COSN '14, pp. 215--226, New York, NY, USA. ACM.

Rossi, L.; Williams, M. J.; Stich, C. & Musolesi, M. (2015). Privacy and the City: User Identification and Location Semantics in Location-Based Social Networks. In *Proceedings of the 9th International AAAI Conference on Web and Social Media (AAAI ICWSM'15)*. AAAI.

Rout, D.; Bontcheva, K.; Preoţiuc-Pietro, D. & Cohn, T. (2013). Where's @wally?: A classification approach to geolocating users based on their social ties. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pp. 11--20, New York, NY, USA. ACM.

Sakaki, T.; Okazaki, M. & Matsuo, Y. (2010). Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pp. 851--860, New York, NY, USA. ACM.

Scellato, S.; Musolesi, M.; Mascolo, C.; Latora, V. & Campbell, A. T. (2011). Nextplace: A spatio-temporal prediction framework for pervasive systems. In *Proceedings of the 9th International Conference on Pervasive Computing*, Pervasive'11, pp. 152--169, Berlin, Heidelberg. Springer-Verlag.

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

Silva, T. H.; Vaz de Melo, P. O. S.; Almeida, J. M. & Loureiro, A. A. F. (2013a). Challenges and opportunities on the large scale study of city dynamics using participatory sensing. In *IEEE Int. Symp. on Computers and Communications (ISCC'13)*, pp. 528--534, Split, Croatia.

Silva, T. H.; Vaz de Melo, P. O. S.; Almeida, J. M.; Salles, J. & Loureiro, A. A. F. (2013b). A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *Proc. of the IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS'13)*, pp. 123--132, Cambridge, MA, USA.

Silva, T. H.; Vaz de Melo, P. O. S.; Almeida, J. M.; Salles, J. & Loureiro, A. A. F. (2014b). Revealing the city that we cannot see. *ACM Trans. Internet Technol.*, 14(4):26:1--26:23.

Silva, T. H.; Vaz de Melo, P. O. S.; Viana, A.; Almeida, J. M.; Salles, J. & Loureiro, A. A. F. (2013c). Traffic Condition is more than Colored Lines on a Map: Characterization of Waze Alerts. In *Proc. of the International Conference on Social Informatics (SocInfo'13)*, pp. 309--318, Kyoto, Japan.

Silveira, L. M.; Almeida, J. M. d.; Marques-Neto, H. T. & Ziviani, A. (2015). Mobdatu: Um novo modelo de previsão de mobilidade humana para dados heterogêneos. SBRC - XXXIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos.

Sun, J. & Li, H. (2008). Listed companies financial distress prediction based on weighted majority voting combination of multiple classifiers. *Expert Systems with Applications*, 35(3):818 – 827.

Times, T. N. Y. (2010). Leaked cables offer raw look at u.s. diplomacy.

Twaroch, F. A.; Smart, P. D. & Jones, C. B. (2008). Mining the web to detect place names. In *Proceedings of the 2Nd International Workshop on Geographic Information Retrieval*, GIR '08, pp. 43--44, New York, NY, USA. ACM.

Twitter (2006). Twitter.

Vasconcelos, M. A.; Ricci, S.; Almeida, J.; Benevenuto, F. & Almeida, V. (2012). Tips, dones and todos: uncovering user profiles in foursquare. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pp. 653--662, New York, NY, USA. ACM.

Wang, L.; Gopal, R.; Shankar, R. & Pancras, J. (2015). On the brink: Predicting business failure with mobile location-based checkins. North-Holland.

Wikipedia (2006). Wikipedia.

Yang, C.; Harkreader, R.; Zhang, J.; Shin, S. & Gu, G. (2012). Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pp. 71--80, New York, NY, USA. ACM.

YouTube (2006). Youtube.

Zhang, A. X.; Noulas, A.; Scellato, S. & Mascolo, C. (2013). Hoodsquare: Modeling and recommending neighborhoods in location-based social networks. In *Proceedings of the 2013 International Conference on Social Computing*, SOCIALCOM '13, pp. 69--74, Washington, DC, USA. IEEE Computer Society.