Detecção de Pontos de Interesse e Predição de Próximo Local de Visita de Usuários Móveis com Base em Dados Esparsos

Cláudio Gustavo S. Capanema[1], **Fabrício A. Silva**[1] **(Orientador)**

[1] Departamento de Informática - Universidade Federal de Viçosa, Viçosa/MG – Brasil

{claudio.capanema, fabricio.asilva}@ufv.br

Resumo. *O estudo sobre a mobilidade humana é uma área que tem ganhado destaque recentemente, tanto no meio acadêmico quanto no corporativo. Pesquisadores buscam entender o comportamento de indivíduos para avançar em propostas inovadores de soluções de mobilidade. Por outro lado, empresas estão interessadas em conhecer melhor os seus usuários para oferecer melhores e mais personalizados serviços. Identificar pontos de interesse (PoIs), classificá-los semanticamente e prever o deslocamento de indivíduos são tarefas relevantes para o estudo da mobilidade humana. Este trabalho apresenta uma abordagem capaz de realizar a identificação e classificação de PoIs de indivíduos que possuem rotinas diferentes. Adicionalmente, uma nova abordagem para a predição semântica do próximo PoI a ser visitado é apresentada, englobando as principais técnicas do estado da arte. Diferente das soluções existentes, ambas as propostas têm o foco em dados esparsos (i.e., coletas com frequências menores), mais apropriados para a utilização em ambiente de produção em larga escala.*

Abstract. *The study on human mobility is an area that has recently gained prominence, both in academia and in the corporate world. Researchers seek to understand the behavior of individuals to advance innovative proposals for mobility solutions. On the other hand, companies are interested in knowing their users to offer better and more personalized services. Identifying points of interest (POIs), classifying them semantically and predicting the displacement of individuals are relevant tasks for the study of human mobility. This work presents an approach capable of performing the identification and classification of PoIs of individuals who have different routines. Additionally, a new approach to the semantic prediction of the next PoI to be visited is presented, encompassing the main state-of-the-art techniques. Unlike existing solutions, both proposals focus on sparse data (i.e., collected at low frequencies), which are more suitable for use in a large-scale production environment.*

1. Introdução

Diversos estudos sobre a mobilidade humana têm sido conduzidos com a utilização em massa dos dispositivos móveis. A possibilidade de se coletar a localização geográfica de milhares de usuários móveis tem contribuído com a solução de problemas em diversas áreas, como previsão de tráfego rodoviário [Gao et al. 2016], identificação de padrões de mobilidade [Yao et al. 2016], planejamento urbano [Rathore et al. 2016], dentre outros. Em comum, essas contribuições estão relacionadas ao conceito de Cidades Inteligentes, onde a tecnologia atua como agente de melhorias constantes.

Dois dos principais problemas relacionados com a mobilidade humana são: a identificação e classificação de pontos de interesse (PoIs) e a previsão da categoria do próximo local a ser visitado. Ambos os problemas são complementares, e juntos são capazes de descrever padrões de mobilidade individuais e coletivos. Porém, grande parte das soluções existentes na literatura necessitam que os dados sejam coletados de forma intensiva, o que leva a um alto custo em um cenário real de larga escala, em que os dispositivos móveis seriam principalmente afetados pelo alto consumo energético. Dados esparsos, no entanto, são mais apropriados a cenários reais, uma vez que são coletados entre intervalos maiores e, por isso, demandam menos recursos (i.e., comunicação, armazenamento e processamento) dos dispositivos móveis e dos servidores.

A hipótese da dissertação resumida neste artigo é que é possível resolver tais problemas de forma eficiente utilizando dados esparsos. Com isso, o objetivo do trabalho é investigar e propor soluções considerando a premissa que os dados são esparsos, comparando-as com trabalhos da literatura, para validar ou refutar essa hipótese. Os resultados mostraram que as propostas, tanto para identificação e classificação de PoIs, quanto para a previsão da categoria do próximo local de visita, superaram as soluções existentes em métricas importantes.

1.1. Contribuições e Publicações

A dissertação [Capanema et al. 2020a] a qual este resumo se refere, apresenta cinco principais contribuições, que estão sumarizadas a seguir:

1. Identificação de PoIs: foi proposto um método de identificação de pontos de interesse adequado a dados esparsos, com a utilização de um algoritmo de agrupamento e de filtros específicos para a seleção de locais relevantes ao usuário.
2. Classificação de PoIs: foi criada uma solução para classificar os PoIs em *Casa*, *Trabalho* ou *Outro*. O algoritmo proposto é ciente do perfil individual de cada usuário, e ajusta seus parâmetros de acordo com a rotina de cada um. Assim, é possível assinalar qual o local de *Casa* e de *Trabalho*, por exemplo, para indivíduos que possuam rotinas alternativas (e.g., trabalham durante a noite). As soluções da literatura não possuem essa capacidade de adaptação, e definem parâmetros fixos, assumindo um comportamento similar entre todos os indivíduos.
3. Disponibilização das soluções: As soluções mencionadas acima foram implementadas em uma extensão da ferramenta *DCluster* [Capanema et al. 2017]. O sistema está disponível para ser utilizado por meio de uma imagem *Docker*[1][2].
4. Análise de rotina: foi apresentado um estudo que mostra o quanto a rotina humana varia ao longo dos finais de semana e dias de semana, correlacionando também essa informação com as categorias dos locais visitados, o que justifica a importância da utilização da informação sobre "tipo do dia" como entrada para modelos preditivos de mobilidade.
5. Previsão da categoria do próximo local de visita: foi proposta a solução *MFA-RNN (Multi-Factor Attention Recurrent Neural Network)* que, diferentemente de outros métodos da literatura, é capaz de englobar diferentes atributos de entrada (localização, tempo, tipo do dia, identificador do usuário), com as camadas *Embedding*, *GRU (Gated Recurrent Unit)* e a camada *Multi-Head Self-Attention*. Os

[1]https://nesped.caf.ufv.br/producao-cientifica/

[2]https://github.com/claudiocapanema/dcluster-docker

resultados indicam que a rede neural apresentada alcança melhorias em relação às soluções da literatura para a categoria de PoI que é menos visitada e, teoricamente, mais difícil de ser interpretada.

Os resultados alcançados durante o desenvolvimento do mestrado foram publicados em duas edições do Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC) (Qualis A4) nos artigos "Identificação e classificação de pontos de interesse com base em dados esparsos" [Capanema et al. 2019] e "MFA-RNN: Uma Rede Neural Recorrente para predição de próximo local de visita com base em dados esparsos" [Capanema et al. 2020b]. Além disso, foi produzido o trabalho intitulado "APPEL: Uma extensão do Kepler para enriquecimento de dados geoespaciais" [Coimbra et al. 2019], em colaboração com outro estudante, que foi publicado no *Brazilian Symposium on Geoinformatics* (GeoInfo) (Qualis B1).

2. Identificação e Classificação de PoIs

Nesta seção, são apresentadas as técnicas usadas neste trabalho para a detecção de pontos de interesse. Os experimentos foram realizados com dados reais de 194 usuários de dispositivos móveis, o que representa uma das maiores bases de dados esparsos onde as tarefas de identificação e classificação já foram estudadas na literatura.

2.1. Identificação de PoIs

Com base em dados esparsos (i.e., longos períodos entre a geração de registros consecutivos) de localização provenientes de usuários móveis, o objetivo da tarefa de identificação de pontos de interesse é detectar quais locais são relevantes para um indivíduo. A solução proposta começa com a utilização do algoritmo *DBSCAN* para agrupar coordenadas geradas por cada usuário. Como os dados utilizados são esparsos, não é necessário realizar um pré-processamento para encontrar os chamados *stay points* (i.e., locais em que se permaneceu por certo tempo). Ao contrário, o *DBSCAN* é capaz de detectar pontos irrelevantes, retirando coordenadas que foram geradas durante um deslocamento, e mantendo aquelas que possivelmente pertencem a um ponto de interesse. Em seguida, os grupos gerados são filtrados com base em duas restrições:

1. A primeira restrição é adaptativa, ou seja, ela se ajusta com base na característica dos dados. Ela assume que cada grupo encontrado pelo *DBSCAN* só pode ser considerado um ponto de interesse se o local em questão foi visitado em uma certa quantidade de dias, que pode ser calculada empiricamente, considerando o período de amostragem dos dados gerados por cada indivíduo.
2. A segunda restrição tem como base a seguinte intuição: de modo geral, quando um estabelecimento é relevante para uma determinada pessoa, ele tende a ser visitado em diferentes períodos do dia. Dessa forma, essa restrição define uma quantidade mínima de horas do dia em que um local deve ter sido visitado para ser considerado um PoI. Esse valor também pode ser ajustado de acordo com os dados.

Os resultados obtidos indicam que o algoritmo proposto alcançou melhorias de pelo menos 13% na precisão para a tarefa de identificação de pontos de interesse, como mostra a Figura 1. Além disso, os resultados mostram que as melhorias foram obtidas com significância estatística, e em diferentes cenários: independentemente da distância limite entre o PoI identificado e o real, representada no eixo-x da Figura 1, para se considerar um acerto, o modelo proposto possui uma precisão maior do que as soluções base.

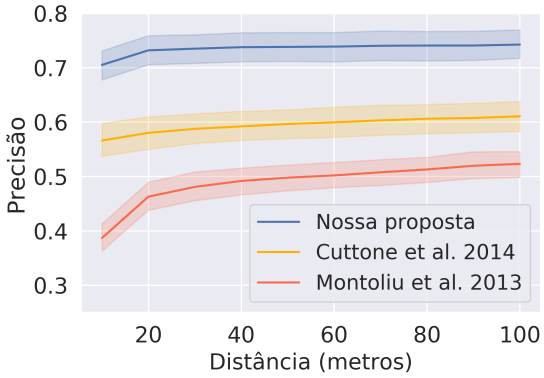

Figura 1. Precisão da identificação.

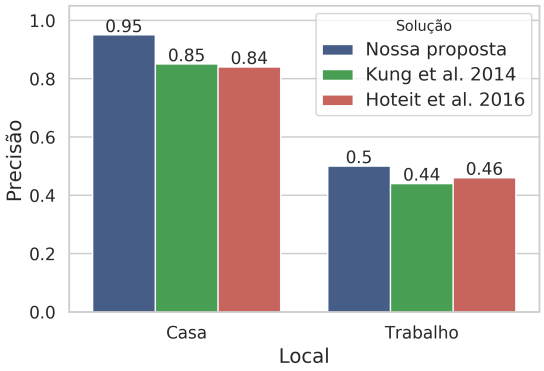

Figura 2. Precisão da classificação.

2.2. Classificação de PoIs

Os pontos de interesse identificados são então classificados em *Casa*, *Trabalho* e *Outro*. Trabalhos anteriores da literatura consideram que o local de *Casa* é aquele que tem mais registros gerados ou no qual o usuário permaneceu por mais tempo em um intervalo fixo de horário, por exemplo entre 22:00h e 08:00h. Analogamente, também é definido um intervalo fixo para se classificar o PoI correspondente ao *Trabalho*. No entanto, essa abordagem impede que seja possível detectar as categorias dos PoIs de indivíduos que possuem uma rotina alternativa, como as que trabalham à noite e descansam durante o dia.

Neste sentido, o trabalho desenvolvido apresenta uma nova abordagem para o problema. Os horários para classificar *Casa* e *Trabalho* não são mais fixos, mas variam de usuário para usuário com base no conceito proposto na dissertação nomeado como *intervalo de inatividade*. Considerando o método de coleta de certas bases de dados esparsos, no qual os registros só são gerados quando há uma certa movimentação em relação à localização anterior, é plausível assumir que a ausência de eventos indique que o usuário não esteja se deslocando por distâncias significativas. Assim, pode-se associar o maior intervalo de inatividade de cada pessoa ao horário em que ela está em sua casa. Dessa forma, o horário para classificar/definir o local de *Casa* é aquele próximo ao intervalo de inatividade, incluindo duas horas antes e duas horas depois. Já o horário para contabilizar registros gerados no local de *Trabalho* é o intervalo de horas oposto ao horário definido para *Casa*.

Os resultados obtidos apresentam melhorias para a classificação dos PoIs *Casa* e *Trabalho* em 10% e 4% (Figura 2), respectivamente. Foi possível utilizar o conceito de intervalo de inatividade em 57% dos usuários, o que demonstra o quão importante é ajustar os parâmetros do modelo de acordo com as características individuais dos usuários. Além disso, diferentemente das soluções base, foi possível classificar corretamente a *Casa* e o *Trabalho* de 4 usuários com rotinas invertidas, ou seja, que trabalham durante a noite.

3. Predição Semântica da Próxima Visita

Prever a categoria do próximo local que um usuário irá visitar permite que provedores de serviços móveis direcionem seus anúncios de forma personalizada em momentos mais

adequados. Por exemplo, sabendo que um indivíduo está mais propenso a estar em deslocamento, é um indicativo de que, talvez aquele não seja o momento ideal para iniciar um anúncio. Por outro lado, se o indivíduo está mais propenso a ir para Casa após o trabalho, um gatilho potencialmente efetivo é oferecer serviços de entrega de comida. Neste sentido, inicialmente é conduzida uma análise sobre rotina humana, de forma a compreender melhor os dados utilizados (Seção 3.1). Em seguida, na Seção 3.2, a proposta da rede neural *MFA-RNN* é apresentada, destacando as suas principais características e a utilização do mecanismo do estado da arte *Multi-Head Self-Attention*.

3.1. Análise da Rotina

Compreender a mobilidade humana e os fatores relacionados é um importante passo para ajudar na previsão de deslocamento. A partir dos PoIs de cada indivíduo, é possível calcular as probabilidades de se estar em *Casa*, em *Outro* local ou em *Deslocamento* (i.e., registro fora de um PoI) para cada hora do dia. Com essas probabilidades em mãos, a entropia de *Shannon* [Shannon 1948] pode ser usada para medir o nível de incerteza sobre a rotina de cada indivíduo. Neste sentindo, quanto maior o valor da entropia, mais incerta/variável é a rotina de uma determinada pessoa.

Neste trabalho, foram utilizados dados de 5.272 usuários de dispositivos móveis coletados durante 62 dias. A Figura 3 mostra que, a rotina dos usuários analisados é mais bem definida (menor valor de entropia) durante os dias de semana, e mais incerta aos finais de semana (maior valor de entropia).

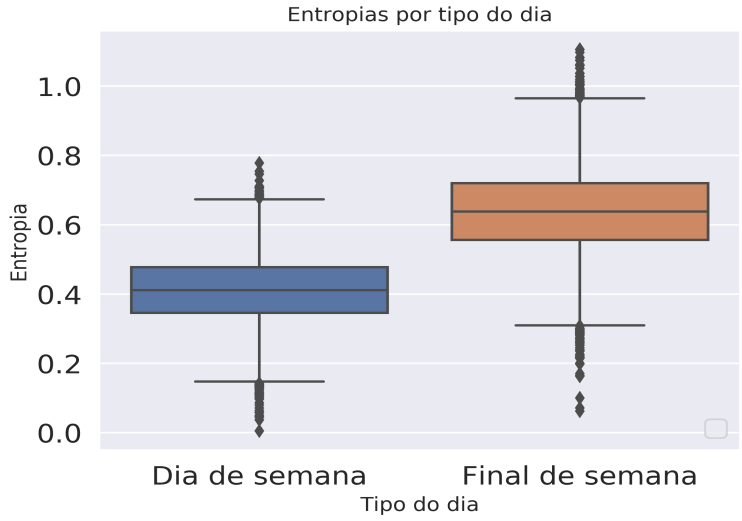

Figura 3. Distribuição das entropias médias de todos os usuários. Quanto menor o valor, mais previsível é a rotina.

Outro fator importante é compreender a taxa de registros gerados em cada tipo ao longo da semana. A Figura 4 indica que existem proporcionalmente mais eventos em *Casa* nos finais de semana do que nos demais dias. Por outro lado, os pontos de interesse do tipo *Outro* têm uma menor relevância no finais de semana. Similarmente, os usuários têm uma tendência menor de estar em *Deslocamento* durante os finais de semana, em relação ao demais dias.

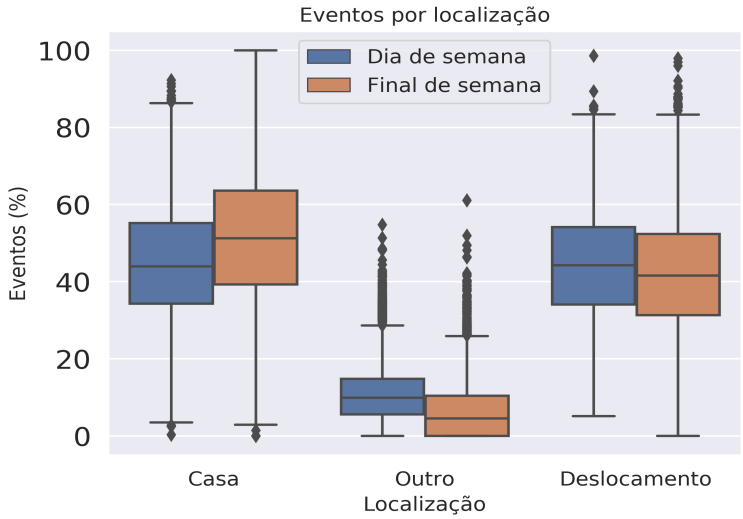

Figura 4. Comparação entre a porcentagem de registros de cada localização gerados em dia de semana e final de semana.

Portanto, indicar o dia da semana em que um registro foi gerado tende a ser uma informação relevante para a previsão do próximo local de visita de um usuário. A variação da entropia, e a troca de proporção de eventos entre *Casa* e *Outro* durante a semana, sugerem as seguintes hipóteses, que são levadas em consideração na proposta descrita na próxima seção:

1. Existe uma probabilidade maior para que o próximo local de visita seja *Casa* nos finais de semana do que nos demais dias. Por outro lado, de segunda-feira à sexta-feira os usuários tendem a estar em *Deslocamento* e visitar mais o PoI *Outro*.
2. Em geral, durante dias de semana, menos localizações diferentes são visitadas em um mesmo horário, o que reduz a entropia. Assim, a tendência é que cada horário esteja associado a uma localização predominante apenas.

3.2. Rede Neural *MFA-RNN*

A rede neural proposta, chamada *MFA-RNN (Multi-Factor Attention-Recurrent Neural Network)*, é uma rede recorrente utilizada para prever a categoria do próximo ponto de interesse a ser visitado. Ao contrário de outras abordagens da literatura, a solução proposta tem a capacidade de trabalhar com diferentes entradas em uma arquitetura recorrente, além de possuir o mecanismo do estado da arte *Multi-Head Self-Attention*. Após a identificação e classificação de pontos de interesse, é possível descrever o histórico de mobilidade de cada indivíduo com base nas sequências de PoIs visitados. Cada sequência $S(u)$ do usuário u contém N eventos $e = (l, t, id, td)$ que denotam, respectivamente, a categoria da localização visitada (*Casa*, *Outro* ou *Deslocamento*), a hora da visita (0-23), o identificador do usuário, e o tipo do dia (dia de semana ou final de semana). Se um dado indivíduo gerou algum registro fora dos seus PoIs, considera-se que o mesmo estava em *Deslocamento*.

A Figura 5 mostra a arquitetura da rede neural proposta *MFA-RNN*. Camadas *Embeddings* são utilizadas para cada entrada do modelo com o objetivo de criar uma representação densa e treinável de cada elemento das entradas. As novas representações

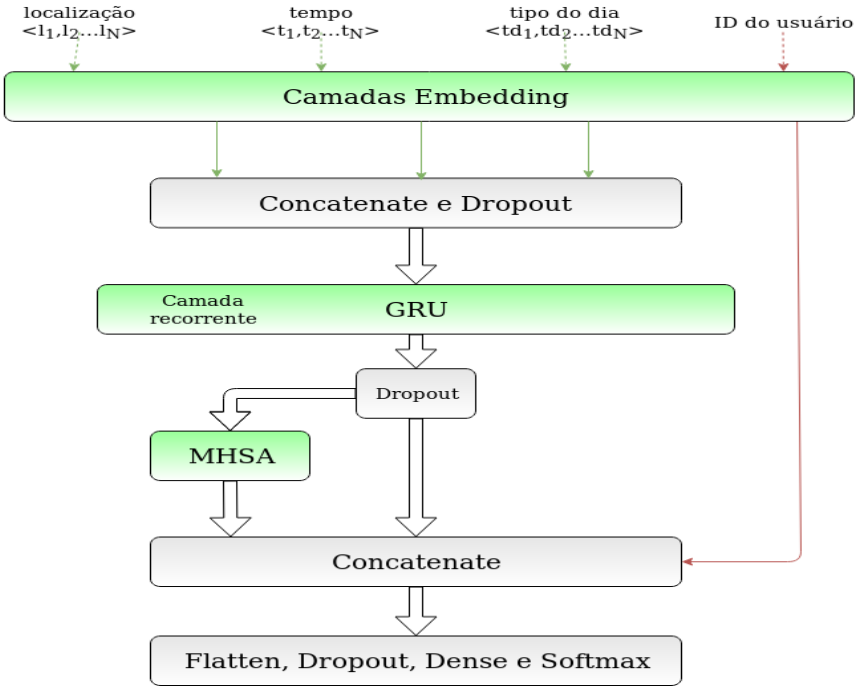

Figura 5. Arquitetura MFA-RNN.

geradas de localização, tempo e tipo do dia são concatenadas e uma camada *Dropout* é aplicada para evitar *overfiting*. Posteriormente, a camada *GRU (Gated Recurrent Unit)* é utilizada para estabelecer correlações entre os elementos da sequência com base na ordem em que estão apresentados. Em seguida, *Dropout* é novamente aplicado, e a técnica de *MHSA (Multi-Head Self-Attention)* tem o papel de estabelecer correlações sobre os elementos de diferentes partes da sequência de entrada. Por último, as saídas das camadas *MHSA* e *GRU* são concatenadas com o *Embedding* do identificador do usuário, e a predição da categoria do próximo local a ser visitado é predita utilizando as camadas *Flatten*, *Dropout*, *Dense* e *Softmax*. Agregar o identificador do usuário por último é uma prática importante para permitir que uma mesma rede neural possa ser treinada para aprender sobre a mobilidade de diferentes indivíduos.

Os resultados alcançados indicam que o modelo *MFA-RNN* supera o *f-score* de quatro soluções base em aproximadamente 5% para prever *Casa* (veja a Figura 3.8 da dissertação) e em cerca de 4% para *Outro* (veja as Figuras 3.8 e 3.11 da dissertação). Para se prever quando um indivíduo estará em *Deslocamento*, o desempenho é similar ao das quatro soluções base usadas na comparação.

4. Conclusões e Trabalhos Futuros

A dissertação apresentou novas contribuições para a área de mineração de dados de mobilidade humana, com o foco em dados de localização esparsos. Os resultados obtidos pelas soluções de identificação e classificação de PoIs trouxeram melhorias significativas em relação às soluções conhecidas da literatura, destacando-se a proposta do conceito de intervalo de inatividade de cada usuário como fator crucial para que o algoritmo de classificação seja capaz de se adaptar aos dados. A rede neural *MFA-RNN* se destacou perante modelos relevantes da literatura ao prever melhor quando o próximo PoI a ser

visitado pertence à categoria *Outro*, de menor suporte, e teoricamente mais difícil de ser compreendida. Uma possível e relevante contribuição futura é classificar mais categorias de PoIs (e.g., Compras, Transporte, Lazer, dentre outros) e utilizá-las para a tarefa de previsão do tipo do próximo local a ser visitado.

5. Agradecimento

Este trabalho contou com o apoio da CAPES.

Referências

Capanema, C., Silva, and Aguiar, F. (2020a). Detecção de pontos de interesse e predição de próximo local de visita de usuários móveis com base em dados esparsos. Master's thesis. Defendida em 20/03/2020.

Capanema, C., Silva, F. A., and Braga, T. M. (2019). Identificacão e classificacão de pontos de interesse individuais com base em dados esparsos. In *Anais do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 15–28. SBC.

Capanema, C. G. S., Silva, F., and Silva, T. (2017). Dcluster: Um sistema para análise exploratória de grandes volumes de dados georreferenciados. In *Satellite Events of the 32nd Brazilian Symposium on Databases (SBBD)*.

Capanema, C. G. S., Silva, F. A., and Silva, T. R. d. M. B. (2020b). Mfa-rnn: Uma rede neural recorrente para predição de próximo local de visita com base em dados esparsos. In *Anais do XXXVIII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos*, pages 127–140. SBC.

Coimbra, G. T., Capanema, C. G. S., Silva, F. A., and Silva, T. R. B. (2019). Appel: Uma extensão do kepler para enriquecimento de dados geoespaciais. In *GEOINFO*, pages 176–181.

Gao, J., Sun, Y., Liu, W., and Yang, S. (2016). Predicting traffic congestions with global signatures discovered by frequent pattern mining. In *2016 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*, pages 554–560. IEEE.

Rathore, M. M., Ahmad, A., Paul, A., and Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks*, 101:63–80.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

Yao, Z., Fu, Y., Liu, B., Liu, Y., and Xiong, H. (2016). Poi recommendation: A temporal matching between poi popularity and user regularity. In *2016 IEEE 16th international conference on data mining (ICDM)*, pages 549–558. IEEE.