Utilizando dados georeferenciados para o tratamento do problema *Indoor-Outdoor Detection*

Raissa P. P. M. Souza[1], Fabrício A. Silva[1],Thais Regina M. B. Silva[1]

[1]Universidade Federal de Viçosa - *Campus* Florestal

{raissa.papini,fabricio.asilva,thais.braga}@ufv.br

Resumo. *Com o crescimento do número de usuários de dispositivos móveis, os provedores de serviços móveis estão cada vez mais preocupados com a qualidade, visando atrair novos clientes e reter os atuais. No contexto de telecomunicações, uma informação relevante para ajudar na percepção da qualidade dos serviços é se o usuário está em ambiente aberto ou fechado. Este problema, conhecido como Indoor-Outdoor Detection, já vem sendo abordado na literatura com técnicas de aprendizado supervisionado, em que são necessários dados com rótulos sobre o tipo de ambiente para se treinar um modelo. Neste artigo, é proposta uma solução não-supervisionada que utiliza dados georeferenciados e o nível de sinal para inferir o tipo de ambiente de um usuário móvel. Os resultados preliminares mostram que a solução é promissora em termos de precisão, além de ser simples e de fácil implementação.*

1. Introdução

Nos últimos anos, o número de usuários de dispositivos móveis vem crescendo significativamente. Com isso, cresce também o número de provedores de serviços a esses usuários, sejam de comunicação, de conteúdo ou de entretenimento. Com diferentes possibilidades de serviços similares sendo oferecidos, os usuários estão cada vez mais exigentes com a qualidade dos mesmos.

Em particular para empresas de telecomunicações, saber se um usuário móvel se encontra em ambiente aberto ou fechado é muito relevante para a qualidade dos serviços. Com isso, as operadoras podem identificar e entender os motivos de usuários estarem sofrendo com má qualidade dos serviços, e assim investirem em melhorias de forma mais assertiva. Por exemplo, se vários usuários estão questionando a qualidade do serviço em uma região, e esses usuários estão em ambientes abertos, pode ser por algum problema nas antenas que atendem aquela região. Por outro lado, se esses usuários estão em ambientes fechados (e.g., um Shopping), pode ser necessário a instalação de outras antenas, ou o aumento de potência das existentes.

No entanto, o problema de identificar se um dispositivo está em ambiente aberto ou fechado (conhecido como *Indoor-Outdoor Detection*) não é trivial. Primeiramente, não é possível obter em larga-escala rastros com rótulos (*fingerprints*) para que sejam utilizados [Lakmali e Dias 2008]. Para isso, seria preciso que fossem coletados dados de latitude, longitude, nível de sinal, e o rótulo indicando se em ambiente fechado ou aberto. Considerando a extensão do território nacional, essa tarefa é inviável. Em segundo lugar, a mesma antena celular atende a ambientes fechados e abertos na maioria das vezes,

dificultando a classificação de ambientes com base apenas na antena utilizada. Por fim, é inviável a implantação de transmissores estáticos em locais estratégicos de ambientes internos para indicar quando um usuário está em ambientes fechados.

Para resolver esse problema, este trabalho propõe uma solução não-supervisionada que utiliza dados georeferenciados não rotulados. Para isso, o algoritmo proposto recebe como entrada dados históricos de latitude, longitude, célula e nível de sinal, faz agrupamentos de registros próximos em uma mesma célula, e cria um modelo de classificação do tipo de ambiente com base no nível de sinal de cada agrupamento. A partir desse modelo, é possível classificar se um novo registro refere-se a um ambiente aberto ou fechado. A premissa do trabalho é que registros próximos geograficamente e utilizando a mesma célula possuem padrões de distribuição de nível de sinal diferentes para ambientes abertos e fechados. O algoritmo então separa essas distribuições em dois grupos, um para ambientes abertos (com valores de nível de sinal maiores) e outro para ambientes fechados (com valores de nível de sinal menores).

O restante deste trabalho está organizado da seguinte forma. A Seção 2 descreve os principais estudos relacionados encontrados na literatura. Os detalhes da solução e uma análise dos resultados são apresentados nas Seções 3 e 4, respectivamente. Por fim, o trabalho é concluído na Seção 5.

2. Trabalhos Relacionados

Alguns trabalhos utilizam os chamados *fingerprints*, que são medidas de nível de sinal coletadas de diferentes localizações em ambientes fechados. A proposta do trabalho [Lakmali e Dias 2008] utiliza uma base de dados previamente coletada com informações sobre a localização conhecida e o nível de sinal de diferentes antenas, para então estimar a localização de um objeto com base em seu nível de sinal. O trabalho descrito em [Gallagher et al. 2010] utiliza WiFi para localização em ambientes fechados, e GPS para ambientes abertos, em um campus universitário. Já o estudo publicado em [Kuo et al. 2010] propõe uma solução de localização em ambientes fechados que utiliza sensores e a tecnologia ZigBee, fazendo uma separação dos ambientes fechados em zonas.

Outros trabalhos não necessitam de dados detalhados de nível de sinal, mas utilizam a localização de pontos de monitoração fixos. O trabalho [Mizuno et al. 2007] utiliza GPS para ambientes abertos e nível de sinal de *Bluetooth* para fechados. O estudo publicado em [Luo et al. 2011] compara algoritmos de localização utilizando identificação por rádio frequência, que avalia o nível de sinal recebido. Essas soluções requerem a instalação de leitores fixos (i.e.,*beacons*) em locais estratégicos para o bom funcionamento.

Alguns trabalhos são mais elaborados e utilizam diferentes técnicas em conjunto para a localização. Os autores de [Pereira et al. 2011] descrevem uma solução flexível que faz uso de técnicas como registros de localização e nível de sinal, pontos de acesso com localizações conhecidas, e células de operadoras com localizações conhecidas. O trabalho descrito em [Reyero e Delisle 2008] propõe um modelo que visa identificar, a cada momento, qual a melhor alternativa para localizar um dispositivo móvel, com base em sinais de GPS e de pontos de acesso disponíveis na proximidade. Por outro lado, os autores de [Kohtake et al. 2011] utilizam o próprio *chipset* do GPS para a localização de objetos móveis tanto em ambientes abertos quanto fechados.

Outros trabalhos focam em detectar apenas em qual tipo de ambiente o usuário se encontra: fechado ou aberto. No trabalho [Gallagher et al. 2011], para o caso de mudança de ambiente aberto para fechado, é identificada uma redução brusca no sinal do GPS. Caso contrário, são utilizados pontos de transição entre um ambiente e outro (i.e., portas), e um temporizador que contabiliza o período de tempo sem nenhum sinal de localização interna em ambientes abertos. Já o estudo de [Li et al. 2015] utiliza sensores diversos, como de luz, magnetômetro, e sinal de torres celulares para fazer essa detecção de transição entre um ambiente e outro. Por fim, o trabalho apresentado em [Radu et al. 2014] descreve uma solução de aprendizado semi-supervisionado, que utiliza valores de intensidade da luz, hora do dia e nível de sinal como parâmetros do modelo.

Apesar dos bons resultados, essas últimas soluções apresentam alguns pontos fracos. Primeiro, requerem a coleta de dados de sensores menos usuais (e.g., luminosidade), consumindo recursos do dispositivo móvel. Além disso, a análise com base no nível de sinal não utiliza a informação de georeferenciamento, considerando somente a variação do nível para detectar uma troca de ambiente. Por fim, alguns trabalhos requerem que um conjunto de dados rotulados seja fornecido para treinamento, o que inviabiliza a sua utilização em escala. A solução proposta neste artigo também visa identificar se um usuário móvel está em um ambiente aberto ou fechado. Porém, faz uso apenas do nível de sinal e da geo-localização do dispositivo, criando um modelo não-supervisionado que não necessita de dados rotulados para funcionar.

3. Solução Não-supervisionada

Seja c_i=<lat,lng,$celula$,$sinal$> os dados contextuais do acesso i de um usuário móvel em uma localização definida pela sua latitude e longitude, as informações de acesso à rede celular definidas pelo identificador da célula de acesso, e a qualidade do sinal definido pela força do sinal. O objetivo é inferir se o usuário está em um ambiente fechado (I, do inglês *Indoor*) ou aberto (O, do inglês *Outdoor*).

A solução proposta neste artigo visa utilizar dados históricos não-rotulados (i.e., não requer a informação de qual ambiente se encontra para o treinamento de um modelo supervisionado) para resolver o problema. Em linhas gerais, a solução funciona da seguinte maneira:

- Para criar o modelo de aprendizado, é utilizado um histórico de dados R contendo vários registros r_i=<lat_i,lng_i,$celula_i$,$sinal_i$> de acesso com as informações de latitude, longitude, célula, e nível de sinal;
- Para reduzir a abrangência de um conjunto de acessos próximos, a precisão da latitude e longitude é reduzida para 3 casas decimais, fazendo com que os registros com mesmas localizações e células estejam a aproximadamente 100 metros de distância no máximo;
- Foram criados grupos G_k em que $r_i \in G_k$ se $r_i[lat, lng, celula]$=$r_j[lat, lng, celula]$ $\forall r_j \in G_k$. Para que o treinamento seja estatisticamente confiável, foram considerados somente os grupos G_k em que $|G_k|$>=30;
- Para cada grupo G_k formado por valores únicos de latitude, longitude e célula, é aplicado um algoritmo de agrupamento que separa os níveis de sinal em duas categorias distintas. A premissa básica desse passo é que, em uma mesma localidade (dentro de um raio de 100 metros alcançado pela redução da precisão da latitude

e longitude) e com a mesma célula, teremos duas categorias distintas de nível de sinal, sendo que o conjunto com menores valores tendem a indicar acessos em ambientes fechados, e o conjunto com valores maiores em ambientes abertos.

Para classificar um acesso, os dados de latitude, longitude e célula são utilizados para recuperar as duas categorias de agrupamento criadas na etapa de treinamento. Então, verifica-se, dentre as duas categorias (i.e., de ambientes abertos ou fechados) qual se aproxima mais do valor do nível de sinal do acesso de entrada. Nesse ponto, duas abordagens foram avaliadas:

1. *Predição Original*: utiliza a predição original do algoritmo de agrupamento, em que a categoria com centro mais próximo ao valor do nível de sinal de entrada é alocada a ele.
2. *Predição com Tratamento de Fronteira*: utiliza a predição original do algoritmo de agrupamento, mas considera como *Desconhecido* caso o registro esteja equidistante, com uma margem de erro, dos dois centros das categorias.

A solução *Predição Original* funciona muito bem quando as categorias são bem separadas, não ocorrendo nenhum registro de fronteira (i.e., valor do nível de sinal está praticamente equidistante dos dois centros). Porém, quando um valor que se aproxima das duas categorias é encontrado, a predição irá associa-lo ao mais próximo, mesmo que seja por uma diferença muito pequena. Esta estratégia cria um problema de precisão pois, apesar de possuir uma revocação de 100%, pode resultar em uma classificação errônea de registros de fronteira.

A solução *Predição com Tratamento de Fronteira*, por outro lado, visa resolver esse problema, atribuindo a esses registros a categoria de *Desconhecido*. Com isso, o objetivo é aumentar a precisão (i.e., os registros classificados serão mais corretos), a um preço de se reduzir a revocação (i.e., menos registros serão classificados). Essa solução apoia-se na premissa de que uma classificação incorreta tem um impacto mais negativo que uma não-classificação, pois ações podem ser tomadas de forma indevida. Em outras palavras, erros do tipo falso-positivo e falso-negativo são mais críticos do que simplesmente não classificar um acesso.

4. Resultados

Para avaliar a proposta, foi utilizado um conjunto de dados real de milhares de usuários de todo o Brasil, contendo 8.878.370 registros. Desses registros, foram encontrados 3.118.305 combinações diferentes de <latitude, longitude, célula>, considerando que a latitude e longitude tiveram suas precisões reduzidas para três casas decimais. Para essas combinações, descartamos aquelas que possuem menos de trinta registros, para que os resultados sejam estatisticamente válidos de acordo com o Teorema Central do Limite, resultando em 25.497 grupos para análise. Para cada uma dessas combinações restantes, foi utilizado o algoritmo *K-Means* para separar os níveis de sinal em duas categorias distintas. Após essa etapa, o modelo de aprendizado está criado.

Para avaliar a qualidade da solução, foram separados e rotulados manualmente 177 acessos aleatórios que não fizeram parte do conjunto original de treinamento. Para a *Predição Original*, o mesmo algoritmo *K-Means* foi utilizado para classificar os registros desconhecidos. Para a *Predição com Tratamento de Fronteira*, o algoritmo do *K-Means* foi adaptado para tratar os acessos de fronteira como *Desconhecidos*. Para isso, foram

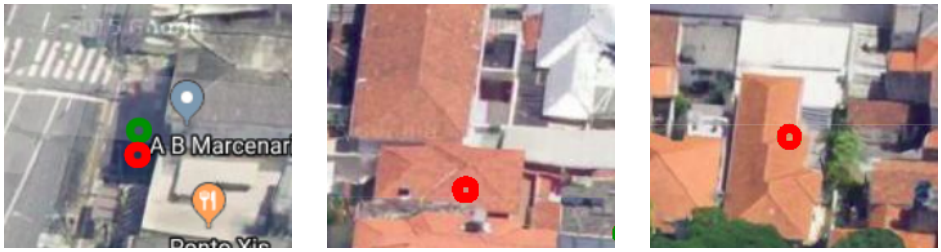

Figura 1. Em vermelho estão exemplos de registros classificados incorretamente pela Predição Original, mas que foram inferidos como *Desconhecidos* pela *Predição com Tratamento de Fronteira*

considerados de fronteira se a diferença da distância do nível de sinal para o centro das duas categorias (i.e., ambiente fechado e aberto) for menor que 20%. Foi preciso que a margem de erro fosse ajustada a este valor por haver locais com muitos prédios e árvores, o que poderia interferir na precisão da análise.

A Figura 1 ilustra alguns dos resultados avaliados. Nela são representados em verde pontos em que a *Predição Original* acertou na classificação, e em vermelho pontos em que a mesma solução errou na classificação. Nestes mesmos casos, a *Predição com Tratamento de Fronteira* classificou como *Desconhecidos* os registros que foram classificados de forma errônea pela *Predição Original*.

A partir dos resultados encontrados, foi possível construir uma matriz de confusão, que pode ser vista na Tabela 1. Observou-se então que, apesar de a quantidade de registros classificados pela *Predição com Tratamento de Fronteira* ter diminuído, isso somente fez com que se reduzisse o número de erros do tipo falso-positivo e falso-negativo, não interferindo muito nas classificações corretas da *Predição Original*.

Tabela 1. Matriz de Confusão da Classificação das Soluções

	Predição Original		***Predição com Tratamento de Fronteira***	
	Fechado	***Aberto***	***Fechado***	***Aberto***
Fechado	44	29	44	21
Aberto	36	68	28	65

Com base nos valores apresentados na Tabela 1, foi possível calcular a porcentagem de precisão e revocação de cada uma das soluções. Para a *Predição Original* observou-se um total de 100% de revocação e 63,27% de precisão. Já para a *Predição com Tratamento de Fronteira* foi obtido um total de 89,26% de revocação e 68,98% de precisão. Estes últimos dados confirmam a hipótese de aumento da precisão em virtude da não-classificação de alguns registros de fronteira.

5. Conclusão

Este trabalho apresentou uma solução não-supervisionada para o problema *Indoor-Outdoor Detection*, que identifica o tipo de ambiente, se aberto ou fechado, que um usuário móvel se encontra durante um acesso a algum serviço. A solução requer as informações de latitude, longitude, célula e nível de sinal, e com base em um modelo não-supervisionado

treinado, infere o tipo de ambiente. Os resultados preliminares são promissores, sendo que foi alcançada uma boa precisão. Além disso, a solução é simples e de fácil implementação. Como trabalhos futuros, pretende-se avaliar a solução com um conjunto maior de dados. Também é importante tratar casos em que não seja possível separar os valores de nível de sinal em dois grupos, por serem muito similares. Por fim, a aplicação de outras técnicas de agrupamento devem ser avaliadas.

Referências

Gallagher, T., Li, B., Dempster, A. G., e Rizos, C. (2011). Power efficient indoor/outdoor positioning handover. In *Proceedings of the 2nd International Conference on Indoor Positioning and Indoor Navigation (IPIN11)*.

Gallagher, T. J., Li, B., Dempster, A. G., e Rizos, C. (2010). A sector-based campus-wide indoor positioning system. In *Indoor Positioning and Indoor Navigation (IPIN), 2010 International Conference on*, pages 1–8. IEEE.

Kohtake, N., Morimoto, S., Kogure, S., e Manandhar, D. (2011). Indoor and outdoor seamless positioning using indoor messaging system and gps. In *Proceedings of the International Conference on Indoor Positioning and Indoor Navigation (IPIN2011), Guimarães, Portugal*, pages 21–23.

Kuo, W.-H., Chen, Y.-S., Jen, G.-T., e Lu, T.-W. (2010). An intelligent positioning approach: Rssi-based indoor and outdoor localization scheme in zigbee networks. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, volume 6, pages 2754–2759. IEEE.

Lakmali, B. S. e Dias, D. (2008). Database correlation for gsm location in outdoor & indoor environments. In *Information and Automation for Sustainability, 2008. ICIAFS 2008. 4th International Conference on*, pages 42–47. IEEE.

Li, M., Zhou, P., Zheng, Y., Li, Z., e Shen, G. (2015). Iodetector: A generic service for indoor/outdoor detection. *ACM Transactions on Sensor Networks (TOSN)*, 11(2):28.

Luo, X., O'Brien, W. J., e Julien, C. L. (2011). Comparative evaluation of received signal-strength index (rssi) based indoor localization techniques for construction jobsites. *Advanced Engineering Informatics*, 25(2):355–363.

Mizuno, H., Sasaki, K., e Hosaka, H. (2007). Indoor-outdoor positioning and lifelog experiment with mobile phones. In *Proceedings of the 2007 workshop on Multimodal interfaces in semantic interaction*, pages 55–57. ACM.

Pereira, C., Guenda, L., e Carvalho, N. B. (2011). A smart-phone indoor/outdoor localization system. In *International conference on indoor positioning and indoor navigation (IPIN)*, pages 21–23.

Radu, V., Katsikouli, P., Sarkar, R., e Marina, M. K. (2014). A semi-supervised learning approach for robust indoor-outdoor detection with smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, pages 280–294. ACM.

Reyero, L. e Delisle, G. Y. (2008). A pervasive indoor-outdoor positioning system. *JNW*, 3(8):70–83.