Analyzing A Touristic Event Popularity Using Social Networks

Pedro Pongelupe Lopes
pedro.pongelupe@sga.pucminas.br
Departamento de Engenharia de Software
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – MG – Brazil

Humberto T. Marques-Neto
humberto@pucminas.br
Departamento de Ciêcia da Computação
Pontifícia Universidade Católica de Minas Gerais
Belo Horizonte – MG – Brazil

ABSTRACT

Context-aware recommendation systems use contextual data to recommend a fully personalized suggestion to their users, for instance, using the massive workloads produced by the usage of mobile apps. In this paper, we collected and analyzed a dataset from social media related to the Belo Horizonte's 2020 Carnival to understand how this event attracts tourists (or Belo Horizonte's non-resident), analyzing their interactions with a large recommendation system. We built a point-of-view of an event mixing its features in two social networks: Twitter and Google Review. Our results show that combining traits lead to better information about the given context using social networks. It helps both the tourists choosing where to travel and the local establishments to provide better services.

KEYWORDS

Social Media, Tourists, Complex Networks, Mobility, Cultural Behavior

1 INTRODUÇÃO

A quantidade massiva de dados que é produzida diariamente por diversas aplicações de forma distribuída estabelecem um cenário interessante e desafiador. As aplicações sensíveis ao contexto (*context-aware*, em inglês) protagonizam este fenômeno, especialmente como aplicativos de *smartphones*. A partir da análise de dados gerados por esse tipo de aplicação é possível compreender comportamentos, preferências e tendências dos usuários e, então, gerar recomendações personalizadas aos mesmos [4]. Esse modelo é categorizado como sistema de recomendação sensível ao contexto (CARS, do inglês *Context-Aware Recommendation System*). Nesses sistemas, a natureza das recomendações é atrelada ao contexto. Exemplo de contexto é o âmbito de mobilidade humana, de forma a realizar predições de localização de pessoas [2].

Na área de turismo esse tipo de aplicação é bem importante. As recomendações a serem feitas podem ser oriundas de diferentes análises como a hora do dia, a disponibilidade do turista em explorar determinada área, seus interesses, seu perfil e suas crenças [3, 6, 10]. Assim, o problema que este trabalho trata é **a análise de popularidade de um evento turístico utilizando dados de redes sociais on-line**.

A diversificação de análises de usuários em sistemas voltados para o setor de turismo é um assunto relevante, pois é uma área importante da economia e de profundo impacto no planejamento urbano [10]. Esses modelos, que são complexos, possuem variáveis intrínsecas e extrínsecas ao turista, como suas crenças [6, 10], e com o trânsito local, respectivamente.

Tendo em vista o panorama contextualizado, o objetivo principal deste trabalho é analisar a popularidade de um evento de turismo utilizando dados de redes sociais on-line, mais especificamente, o Twitter e o Google Review. Para alcançar esse objetivo, são propostos os seguintes objetivos específicos:

(1) Categorizar perfis de usuários como residente ou turista (ou não-residente);
(2) Estabelecer uma medida de popularidade para um evento de turismo utilizando dados de redes sociais on-line.

O Carnaval de Belo Horizonte de 2020 foi o evento alvo escolhido para compreender a popularidade de um evento de turismo sob ótica dessas redes sociais: o Twitter e o Google Review. Foi estabelecido grau de popularidade para cada bloco no evento dentro de cada rede social isoladamente e em conjunto, apontando que uma análise mais sólida é construída utilizando mais de uma rede. Além disso, analisou-se a proporção entre turistas (ou não-residentes) e residentes utilizando dados do Google Review.

Este trabalho está divido em sete seções. Na Seção 2 é apresentada a fundamentação teórica, a qual aprofunda nos conceitos necessários para o entendimento da pesquisa. A Seção 3 aborda os trabalhos relacionados. Em sequência, a Seção 4 apresenta os materiais e métodos. Na Seção 5 é apresentados os resultados. A Seção 6 trata as discussões de implicações e limitações. Finalmente, a Seção 7 apresenta as conclusões e os trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção são apresentados os conceitos chaves para o entendimento deste trabalho. O primeiro é mobilidade humana, em seguida, *CARS*, situando o tipo da aplicação e seus contextos de aplicação. Modelos de mobilidade humana são explicados, considerando as diferentes fontes de dados: homogêneos e heterogêneos.

2.1 Mobilidade Humana

A análise da mobilidade humana é baseada em modelos que buscam prever informações de deslocamento das pessoas a partir de uma análise de dados. Essas informações são relevantes, pois possibilitam planejamento e tomada de decisões mais acertadas baseando-se na vida das pessoas que habitam uma determinada região [5, 8, 9].

In: I Concurso de Trabalhos de Iniciação Científica (CTIC 2021), Minas Gerais, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2021.

ISSN 2596-1683

2.2 Sistemas de Recomendações Sensíveis ao Contexto

Sistemas de recomendações podem ser categorizados pelo grau de conhecimento contextual de um determinado objeto de estudo. Numa abordagem *top-down* de observação contextual são dispostos como: totalmente observável, parcialmente observável e não observável [1]. No totalmente observável os valores e estruturas no momento da recomendação são caracterizados como vitais, enquanto no parcialmente observável utiliza-se de valores, mesmo parcialmente, na recomendação. Por fim, o não observável não se embasa em características explícitas do contexto ao fazer uma recomendação.

3 TRABALHOS RELACIONADOS

Nesta seção são tratados os trabalhos que abordam a aplicação de modelos de previsão para mobilidade humana em aplicações voltadas ao turismo. Esses trabalhos aprofundam em tópicos significativos o objetivo do trabalho de analisar o comportamento de turistas com o propósito de criar soluções tecnológicas para problemas reais.

Carvalho et al. (2018) classificam que existem três fases claras de uma viagem, a preparação (*pre-trip*), a viagem em si (*trip*) e o momento após (*after trip*) e de modo a buscar uma melhor experiência ao turista é necessário integrar essas fases continuamente, preferencialmente em tempo real [3]. Tal integração deve extrapolar as tecnologias de sistema de posicionamento global (GPS do inglês, *Global Positioning System*) e compor um contexto a cerca dos hábitos do turista. Partindo desse contexto, Carvalho et al. (2016) propuseram um *framework* de integração agregando dados sobre os turistas e os comércios que podem os servir gerando um perfil único e acessível ao turista possibilitando um refinamento das recomendações. Este trabalho propõe uma abordagem sistemática na fase definida como *trip*, extrapolando o uso tecnologias de GPS juntamente a análise de redes sociais.

Silveira et al. (2015) propõem um modelo de mobilidade humana que atua em insumos de dados de diversas fontes, dados heterogêneos. *MobDatU*, o modelo proposto, utiliza-se de insumos de *tweets* georreferenciados e chamadas telefônicas [9]. O funcionamento é dado pelo enquadramento da área estudada em *grid*, subdividindo-a em regiões que são associadas com uma medida de popularidade por visitas, assim como o *SMOOTH*. Além disso, como o *Leap Graph*, é criado um grafo simulando a movimentação dos usuários, diferenciando do modelo de referência que mapeia as relações por indivíduo e não grupo. Neste estudo, são analisados os impactos de modelos de mobilidade humana baseado em fontes heterogêneas, como o *MobDatU*. Focalizando no contexto de turismo e na rede mercadológica que atende esse público. Este trabalho se posiciona entre os tratados nesta seção por tratar o assunto da análise da mobilidade humana direcionada ao turismo. Além disso, a utilização de análises sob um elemento de um contexto para estabelecer relações sobre o objeto analisado.

4 MATERIAIS E MÉTODOS

O estudo apresentado neste trabalho é uma busca descritiva com características quantitativas. Esta seção aborda as etapas necessárias para a realização dos objetivos propostos deste estudo.

4.1 Procedimentos

Propôs-se coletar informações de redes sociais sobre um evento regional que mobilizou tanto moradores locais e turistas, o Carnaval de Belo Horizonte de 2020. As redes escolhidas foram o Twitter e o Google Review, pois, ambas permitem, através de informações públicas realizar um mapeamento da localização do usuário. Além disso, possibilitam o estudo de outras variáveis como o estudo de sentimentos e possível desenho de um perfil de turista deste usuário.

4.2 Coleta de dados

A coleta de dados deste trabalho foi executada em 3 grandes fases, uma etapa para recuperar dados sobre o Carnaval de Belo Horizonte de 2020 e depois uma fase para cada rede pesquisada. Todos os *scripts* produzidos foram desenvolvidos em Python 3.6 e armazenado em dois bancos, um relacional com funções espaciais, Postgis, e um *NoSQL*, o MongoDB.

4.2.1 Dados Comuns. A primeira etapa foi a extração das informações oficiais do Carnaval de 2020 em Belo Horizonte, disponibilizados na página da prefeitura do município. Foram 382 blocos iniciando oito de fevereiro e terminando em primeiro de março, sendo que 38 foram listados em dois dias diferentes de evento, portanto, contabilizando 344 blocos distintos. De cada bloco foram coletadas as seguintes informações: nome, uma breve descrição, data, estilos musicais, perfil de público alvo e a rota do bloco. Foi desenvolvido uma ferramenta de *web scraping* para coleta e armazenamento no banco de dados espacial.

Em seguida, houve uma tratativa quanto às rotas de blocos coletadas. As rotas são compostas por endereços provindos pelo sistema da prefeitura. Todavia, houve a necessidade de converter-los em coordenadas geográficas, longitude e latitude. Outro *script* buscou os endereços do banco de dados e os atualizava em sequência, utilizando a *API* de *Geocode* do Google. Foram convertidos 2004 endereços diferentes.

4.2.2 Google Review. Utilizando as coordenadas convertidas, foi possível recuperar os estabelecimentos listados pelo *Google Places* em até 150 metros de distância de algum dos 2004 pontos de rota de blocos de carnaval coletados. Consumindo o serviço de *search places nearby*, foram encontrados 12.866 estabelecimentos de 88 tipos dos 96 listados no serviço. Limitando as localizações por tipo focados ao contexto de turismo, restaram apenas sete tipos. Entre estes, os tipos locais selecionados que oferecem serviços de alimentação ou de bebidas alcoólicas, são: bar, café, lojas de bebidas e restaurantes. Além disso, dois serviços de apoio foram cotados, farmácias e lojas de conveniência. Finalmente, o último tipo foi de atrações turísticas, resultando um conjunto de 1.209 estabelecimentos.

Em sequência, foi desenvolvido um *script* para recuperar as *reviews* dos estabelecimentos filtrados. A busca levou em conta as datas dos *reviews* para recuperar dados apenas do período do Carnaval. Uma limitação do Google Review é que a data da *review* é em tempo relativo, portanto foi estabelecido um período relativo que abrangesse o período do Carnaval, durante a coleta era de 3 ou 4 meses atrás. Os dados das *reviews* são o texto, a *URL* do perfil quem a comentou, a quantidade de estrelas nesta *review* e a data relativa.

Finalmente, a medida que as *reviews* eram coletadas foram buscadas informações sobre os autores das mesmas na plataforma do

Google Review. Para cada usuário foi recuperado todas as localizações das últimas *reviews* feitas por esse usuário. Primeiramente, os *reviews* contendo fotos e depois apenas aqueles com texto.

4.2.3 Twitter. Utilizando os dados coletados sobre os blocos, foi coletado os dados durante os oito últimos dias de evento. A coleta foi realizada em lotes e a cada interação poderia resultar em até 200 *tweets*, utilizando a biblioteca Tweepy. Os critérios de busca utilizados foram a data da busca e o texto do *tweet*. O primeiro critério foi estabelecido para buscar apenas *tweets* de blocos que já passaram ou estava acontecendo naquele momento. O segundo critério foi sobre o conteúdo que deveria conter o nome do bloco, utilizando um *script*. [7]

Finalmente, foi estabelecido um critério para evitar duplicidade e buscas fora do período. A utilização de *id* do último *tweet* feito sobre um bloco na consulta da *API* que garante apenas resultados posteriores aquele. Esta coleta foi realizada utilizando um serviço que a cada minuto buscava pela *API* armazenando no banco de dados *NoSQL*. A *API* utilizada tinha uma limitação quanto ao tamanho do texto do *tweet*, assim marcando os grandes como truncados. Portanto, intercalado ao serviço por minuto, havia um serviço que a cada hora recuperava o texto completo de até 1.000 *tweets* buscando pelo *id*.

4.3 Métodos de Avaliação

Esse trabalho é de caráter quantitativo e descritivo. Portanto, as análises fundamentadas em recomendações sensíveis ao contexto em aplicações de turismo propostas foram avaliadas sob óptica do Carnaval de Belo Horizonte em 2020, assim descrevendo a popularidade nesse evento.

A popularidade foi analisada quanto quantidade de menções nas redes, diretamente ou através dos lugares que estão prestando serviço ao bloco. A popularidade foi calculada com a seguinte fórmula:

$$P = (((Qgr/Tgr) * \alpha gr) + ((Qtt/Ttt) * \alpha tt))/N \quad (1)$$

A Fórmula 1 é uma média ponderada para a popularidade de cada rede. Sendo a popularidade de cada rede calculada com um Q, e um T e um fator de ajuste de popularidade, α. Q representa a quantidade de interações, *tweets* ou *reviews*. O T representa a maior quantidade de interações de um bloco. O α é um fator de ajuste de cada rede, para o Google Review é definido pela média de estrelas daquele bloco dividido pelo máximo possível de estrelas, 5. Para o Twitter, o αtt é definido como uma constante, 1, por limitações discutidas na seção de discussões. A popularidade máxima é 1.

O método postulado para Categorizar um participante do Carnaval de Belo Horizonte de 2020 foi analisar as variáveis definidas no contexto com análise de fatores das redes sociais. O método utilizado não possui exatidão, portanto foram definidos um α, como é descrito na seção de resultados. Apenas a base do Google Review foi submetida nessa análise, devido a limitações exploradas quanto a base do Twitter, abordados na seção de discussões.

A forma utilizada caracteriza um usuário do Google Review como turista ou residente analisando *point-of-interest* [10] utilizando a localização das últimas postagens na plataforma. Definido um α, o perfil é considerado turista caso o número de postagem com a localização diferente de **Belo Horizonte - MG** seja superior ao α, caracterizando pessoas não residentes de Belo Horizonte.

5 RESULTADOS

Esta seção apresenta os resultados obtidos pela aplicação dos métodos de avaliação propostos para compreender as métricas definidas. A seção está em 2 partes. A primeira, os resultados que categorizam as variáveis presentes no Carnaval de Belo Horizonte de 2020. Na segunda estão os resultados da classificação entre turista e residente.

5.1 Resultados de análise de contexto

Primeiramente, foi possível extrair variáveis que caracterizaram o Carnaval de Belo Horizonte de 2020 utilizando o Twitter e o Google Review. Com um total de 266.494 interações dessas redes, sendo 239.951 *tweets* e 26.534 *reviews*.

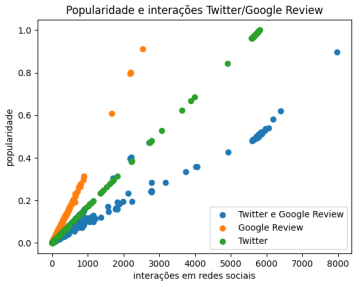

Figure 1: Popularidade dos blocos de nas redes sociais

Para as três análises existem três patamares de popularidade, estabelecendo níveis de qualidade de blocos, observando a Figura 1. No Google Review e no Twitter o primeiro patamar se apresenta por volta de $P = 0,35$, juntos, é por volta de $P = 0,3$. O segundo patamar, no Twitter é até $P = 0,55$, no Google Review e no conjunto é de até $P = 0,6$. O último patamar, se estabelece logo acima do segundo até $P = 1$.

A Figura 1 retrata a popularidade individual de cada rede e também em conjunto. O Google Review apresenta menor quantidade total de interações, *reviews*. A alta densidade na parte inicial da distribuição indica que não é uniforme a distribuição de *review* por estabelecimento, em sua maioria até 1.000. Além disso, a angulação é dita pelo αgr, apontando uma boa classificação média de estrelas nos estabelecimentos, condizendo com a média mensurada de 4,38 (87,6%).

Os blocos sob a ótica do Twitter, representados na Figura 1 possuem uma distribuição mais uniforme da interações, *tweets*, e maior quantidade total. A angulação da curva representa o αtt, que por limitações neste trabalho, é 1. Esse valor gera uma reta que representa o melhor cenário, consequentemente, os blocos possuem a maior classificação individual possível, em paralelo ao Google Review, equivale a 5 estrelas.

Ao analisar as duas redes juntas pela Figura 1, nota-se uma distribuição mais pulverizada das interações, soma de *tweets* e *reviews*. Pulverização que implica em nivelamentos mais criteriosos quanto a popularidade.

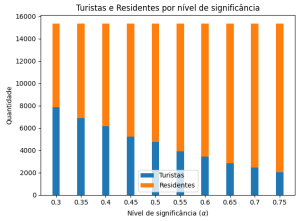

Figure 2: Quantidade de *Reviews* por usuários

Blocos avaliados sob duas redes compõem novas perspectivas sobre o mesmo evento, refinando a informação. O Bloco Duro que é o mais popular de todos. Bloco que é o mais popular apenas quando se combina as redes, individualmente o Bloco da Bicicletinha e o Daquele Jeito foram os mais populares, do Google Review e do Twitter, respectivamente.

5.2 Resultados categorização de perfis

A categorização de perfis foi realizada para aprofundar em análises sobre a mobilidade humana, através da utilização de redes sociais. O Google Review foi utilizado para categorizar turistas e residentes durante o Carnaval de Belo Horizonte de 2020.

A base de usuários que fizeram as *reviews* no Google Review é composta por 17.222 usuários que possuam pelo menos um *review* registrado com a visibilidade pública. O α, variou entre 30% até 75%, pois garantiu análise de 15.361 (89,2%) dos usuários. Foi reduzido a quantidade de usuários para possibilitar o aumento do α. No cenário de $\alpha = 0,75$ de um usuário com menos 4 *reviews* implicaria em classificação automática como residente.

Aplicando o algoritmo descrito ao usuários com a partir de quatro *reviews* é evidente a relação inversamente proporcional entre do número de turistas e o α, conforme apresenta a Figura 2. Além disso, a quantidade de turistas no α máximo é de 2.027 (13,19%) que aponta percentual não desprezível da base analisada.

O decaimento da quantidade de residentes sob α, na Figura 2, revela que há um terceiro grupo, residentes que simpatizam com turismo. Apesar de serem classificados como residentes, os integrantes desse grupo possuem significativa quantidade de postagens fora de Belo Horizonte.

6 DISCUSSÕES DE IMPLICAÇÕES E LIMITAÇÕES

Neste estudo houve algumas limitações durante o processo de coleta de dados. Não foi possível extrapolar as análises do *Twitter*, como o αtt e o classificador de turista e residente. Na coleta de dados no Google Review houve limitações quanto a data de pesquisa e a quantidade de *reviews* feitas por um usuário. A primeira, é dado pela medida ser em tempo relativo, de modo forçar inferência do fato. A segunda, é a impossibilidade de recuperar todo o histórico de *reviews* limitando a análise para os *reviews*, mais recentes, em torno de 150.

7 CONCLUSÕES E TRABALHOS FUTUROS

Estudo propõe fomentar as análises de elementos de contexto em aplicações de recomendação sensíveis ao contexto direcionado ao turismo. Portanto, foram aplicas os conceitos de mobilidade humana sob o Carnaval de Belo Horizonte de 2020 utilizando dados de redes sociais, Twitter e Google Review. Os resultados foram obtidos por análises das redes individualmente e em conjunto, apontando características a popularidade do evento. A análise em conjunto de redes sociais em detrimento delas separadas apontam um cenário mais real do contexto e a classificação de turistas e residentes pode ser realizado utilizando redes sociais.

Para trabalhos futuros, propõe-se uma análise do conteúdo dos *tweets* e das *reviews*. Propor novas variáveis extraídas do contexto e somando com as propostas neste trabalho.

REFERENCES

[1] Gediminas Adomavicius, Bamshad Mobasher, Francesco Ricci, and Alex Tuzhilin. 2011. What a Recommender System Knows About Contextual Factors. *Springer* (2011), 67–80. https://link.springer.com/chapter/10.1007/978-0-387-85820-3{_}7

[2] Roni Bar-David and Mark Last. 2014. Context-Aware Location Prediction. *The Fifth International Workshop on Mining Ubiquitous and Social Environments* (2014), 51–66.

[3] António Carvalho, Elisabete Paulo Morais, and Carlos R. Cunha. 2018. Location based mobile services & Context-aware: An approach to the tourism sector. *Proceedings of the 32nd International Business Information Management Association Conference, IBIMA 2018 - Vision 2020: Sustainable Economic Development and Application of Innovation Management from Regional expansion to Global Growth* (2018), 6828–6836.

[4] Arati R Deshpande. 2016. Context based Recommendation Methods : A Brief Review. *International Journal of Computer Applications* (2016), 14–19.

[5] Wei Dong, Nick Duffield, Zihui Ge, Seungjoon Lee, and Jeffrey Pang. 2013. Modeling cellular user mobility using a leap graph. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 7799 LNCS (2013), 53–62. https://doi.org/10.1007/978-3-642-36516-4-6

[6] Ana P G Ferreira, Thiago H Silva, and Antonio A F Loureiro. 2020. Uncovering Spatiotemporal and Semantic Aspects of Tourists Mobility Using Social Sensing. arXiv:2005.09033 [cs.SI]

[7] Rômulo Meloca, Gustavo Pinto, Leonardo Baiser, Marco Mattos, Ivanilton Polato, Igor Scaliante Wiese, and Daniel M German. 2018. Understanding the Usage, Impact, and Adoption of Non-OSI Approved Licenses. In *Proceedings of the 15th International Conference on Mining Software Repositories* (Gothenburg, Sweden) *(MSR '18)*. Association for Computing Machinery, New York, NY, USA, 270–280. https://doi.org/10.1145/3196398.3196427

[8] Aarti Munjal, Tracy Camp, and William C. Navidi. 2011. SMOOTH: A simple way to model human mobility. *MSWiM'11 - Proceedings of the 14th ACM International Conference on Modeling, Analysis, and Simulation of Wireless and Mobile Systems* (2011), 351–360. https://doi.org/10.1145/2068897.2068957

[9] Lucas Maia Silveira, Jussara M. Almeida, Humberto Marques-Neto, and Artur Ziviani. 2015. MobDatU: A New Model for Human Mobility Prediction Based on Heterogeneous Data. *Proceedings - 33rd Brazilian Symposium on Computer Networks and Distributed Systems, SBRC 2015* (2015), 217–227. https://doi.org/10.1109/SBRC.2015.34

[10] David A. M. Veiga, Gabriel B. Frizzo, and Thiago H. Silva. 2019. Cross-Cultural Study of Tourists Mobility Using Social Media. In *Proceedings of the 25th Brazillian Symposium on Multimedia and the Web* (Rio de Janeiro, Brazil) *(WebMedia '19)*. Association for Computing Machinery, New York, NY, USA, 313–316. https://doi.org/10.1145/3323503.3360620