Um brinde ao Untappd! Usando preferências por cervejas para o planejamento urbano e estudo de diferenças culturais

Saulo A de Brito, Ariane L Baldykowski, Sandro Miczevski, Thiago H Silva
Universidade Tecnológica Federal do Paraná
Departamento Acadêmico de Informática
Curitiba, PR, Brasil
saulobrito,ariane,sandrom@alunos.utfpr.edu.br;thiagoh@utfpr.edu.br

ABSTRACT

This study presents partial results of an ongoing investigation, in the context of the project CNPq-UrbComp, regarding the exploration of data from the location-based social network (LBSN) Untappd, which is an LBSN for sharing beer preferences. We explore Untappd data in two main directions. First, in the context of urban planning in Curitiba. Curitiba recently announced the creation of a Craft Beer Street, to promote local beers. By using this as a real case study we investigate a new approach that could help create this kind of attractions. Second, we show the potential to explore the preferences for beer to study urban social behavior, particularly related to the automatic identification of cultural aspects. Automatic identification of cultural differences is a valuable information that can enable new services.

KEYWORDS

Social media mining, Untappd, beer preferences, urban planning, cultural differences

1 INTRODUÇÃO

Este estudo apresenta resultados parciais de uma investigação em andamento dentro do projeto CNPq-UrbComp, que visa à caracterização, modelagem e utilização do comportamento social urbano e da dinâmica de cidades utilizando fontes de dados como as redes sociais baseadas em localização (LBSN). Nesse contexto, apresentamos uma exploração de dados da LBSN Untappd, que é uma LBSN para compartilhar preferências de cerveja. Foram estudados os dados do Untappd em duas direções principais.

A primeira direção foi no contexto do planejamento urbano. Para essa questão um aspecto fundamental é estudar dinâmicas urbanas em diferentes escalas espaciais, ponto que tem sido tradicionalmente desafiador, pois, geralmente, é um processo caro, uma vez que demanda a realização de entrevistas e questionários a um grande número de pessoas [14]. Neste trabalho, foi explorado o potencial e as possibilidades de usar dados de redes sociais, particularmente o Untappd, para medir o uso do espaço urbano na cidade.

Curitiba anunciou recentemente a criação de uma Rua da Cerveja, para promover cervejas artesanais locais. Ao usar isso como um estudo de caso real, foi investigada uma nova abordagem que poderia ajudar a criar esse tipo de atração. Estudar cerveja no contexto de Curitiba é interessante, pois um estudo anterior [15] mostra que em Curitiba as pessoas parecem ter maior interesse por cervejas artesanais do que em outras grandes cidades do Brasil.

Em seguida, foi apresentado o potencial de explorar as preferências de cerveja para estudar comportamentos sociais urbanos, particularmente relacionados à identificação automática de aspectos culturais. O estudo da influência de diferenças culturais no comportamento humano é um tema particularmente desafiador. Cultura é um conceito tão complexo e interessante que nenhuma definição simples pode capturá-lo. Entre os vários aspectos que definem a cultura de uma sociedade incluem suas artes, crenças religiosas e costumes. Neste trabalho investigamos se preferências por cervejas podem também ser uma boa característica descritiva de diferenças culturais. A identificação automática das diferenças culturais é uma informação valiosa que pode permitir novos serviços.

O trabalho está organizado da seguinte forma. A Seção 2 apresenta alguns dos trabalhos relacionados. A Seção 3 apresenta os datasets utilizados. A Seção 4 apresenta a abordagem utilizada que poderia auxiliar o planejamento de áreas temáticas de uma determinada cidade. A Seção 5 descreve a metodologia de agrupamento de regiões de acordo com a informação cultural. Por fim, a Seção 6 apresenta as conclusões do trabalho.

2 TRABALHOS RELACIONADOS

Nesta seção são apresentados alguns dos trabalhos relacionados. Estudos anteriores mostraram que as redes sociais, especialmente as redes sociais baseadas em localização, podem ser usadas para aprofundar a compreensão sobre o comportamento do usuário e a dinâmica da cidade [8, 11, 14, 16]. O consenso parece ser que estudar a cidade através dos dados das redes sociais pode ser uma alternativa para entender as atividades e interesses das pessoas com dados que podem ser obtidos mais facilmente [13].

Os dados das mídias sociais podem ser usados para descobrir áreas funcionais dentro das cidades [17]. Adicionalmente, a alta granularidade temporal e o grande volume de dados disponíveis oferecem oportunidades para estudar áreas urbanas em escala espacial com maior detalhamento do que métodos tradicionais [8].

Em [15] são demonstradas algumas das possibilidades em se analisar dados de redes sociais, especificamente do Untappd. O estudo em questão foi elaborado para auxiliar um empreendedor fictício a tomar suas decisões e planejar estratégias com foco em pequenas ou médias empresas que possuem pouco ou nenhum recurso para investir em pesquisas. De modo similar, com a ajuda do Untappd, diversas questões relacionadas aos hábitos de consumo cerveja foram analisadas por Chorley et al. [2].

In: XIV Workshop de Trabalhos de Iniciação Científica (WTIC 2017), Gramado, Brasil. Anais do XXIII Simpósio Brasileiro de Sistemas Multimídia e Web: Workshops e Pôsteres. Porto Alegre: Sociedade Brasileira de Computação, 2017.

ISBN 978-85-7669-380-2.

Na linha do estudo de diferenças culturais, estudos também mostraram como o uso de sistemas da Web social podem variar entre os países. Por exemplo, Hochman et al. [6] investigaram as preferências de cor em fotos compartilhadas através do Instagram, mostrando diferenças consideráveis nas preferências entre os países com culturas distintas. Garcia-Gavilanes et al. [5] estudaram variações de uso do Twitter entre os países, mostrando que as diferenças culturais não são apenas visíveis no mundo real, mas também observadas no Twitter. Nessa direção, Silva et al. [14] propuseram uma nova metodologia para a identificação de fronteiras culturais e semelhanças entre populações, considerando hábitos de comida e bebida.

3 DESCRIÇÃO DOS DADOS

Coletamos dados do Untappd compartilhados no Twitter. Conforme abordado por [15] a maioria dos tweets enviados através do Untappd segue um padrão. Um dos tweets coletados trazia como mensagem: "Medalha de ouro!!! - Drinking a Cacau Wee by @bodebrown at @riodejaneiro — https://t.co/NQd7fI6LYE #photo". Ao analisar este tweet é possível identificar que após a palavra "Drinking" é informado o tipo de cerveja sendo consumida pelo usuário, após a palavra "by" é informada a empresa fabricante da cerveja e, por fim, após a palavra "at" é informado o local onde o usuário está consumindo a cerveja.

Foi obtido a data e hora do compartilhamento, coordenadas geográficas, identificação do usuário, notas fornecidas para determinadas cervejas, entre outras informações. Isso possibilita adquirir informações diversas, como cervejas com maior consumo em determinadas cidades. Após a coleta dos dados foram realizadas limpezas para garantir que dados com alta qualidade fossem estudados. Entre os passos realizados podemos citar a remoção de check-ins com dados faltantes.

Para a realização deste estudo, foram coletados dados durante aproximadamente seis meses, totalizando 1,7 milhões de tweets. O período de coleta compreendeu o período de Novembro de 2016 a Abril de 2017. O processo de coleta ficou indisponível em alguns dias. Com isso, não consideramos esses dias nas análises quando uma questão temporal é importante. Dentre os tweets coletados, 702 mil tweets, ou 39,2% do total, eram geolocalizados e esses foram considerados.

4 PLANEJAMENTO URBANO COM DADOS DO UNTAPPD

Os resultados apresentados nesta seção foram publicados no artigo [12], onde o objetivo é apresentar o potencial uso dos dados considerados para a questão de planejamento urbano. Por limitação de espaço, mais detalhes podem ser obtidos em [12].

Foi considerado parcialmente o dataset mencionado anteriormente, contemplando um período sem nenhum problema de coleta: de 03 de novembro de 2016 a 05 de março de 2017 (conjunto de dados 2017), dataset denominado Dataset2017. Após a coleta de dados, realizamos um processo de filtragem. Consideramos apenas um check-in por usuário, em um estabelecimento específico. Além disso, consideramos apenas mensagens contendo informações geográficas sobre onde a cerveja estava sendo consumida e também mensagens contendo o tipo de cerveja que o usuário estava bebendo. Foi considerado também um dataset de 2013, que foi cedido pelos autores de [15] (06 de abril a 30 de abril de 2013, chamado de Dataset2013), que possui os mesmos tratamentos.

Para atingir nosso objetivo proposto mencionado anteriormente, foi preciso descobrir áreas populares na cidade com base no número de check-ins observados em nossos conjuntos de dados. Neste estudo, foi utilizado o algoritmo de agrupamento DBSCAN [3]. O DBSCAN é um algoritmo de agrupamento baseado em densidade que agrupa pontos próximos. Ele requer dois parâmetros: eps, usado para identificar pontos vizinhos; e o número mínimo de pontos necessários para formar uma região densa $minPts$. Devido às suas características, as particularidades do problema foram atendidas. Sendo assim, foi utilizado um pacote R disponível online[1] contendo a implementação desse algoritmo. Foi considerada a distância great-circle, que é a distância mais curta entre dois pontos na superfície de uma esfera, medida ao longo da superfície desta. Isto é calculado com a ajuda da fórmula de Haversine.

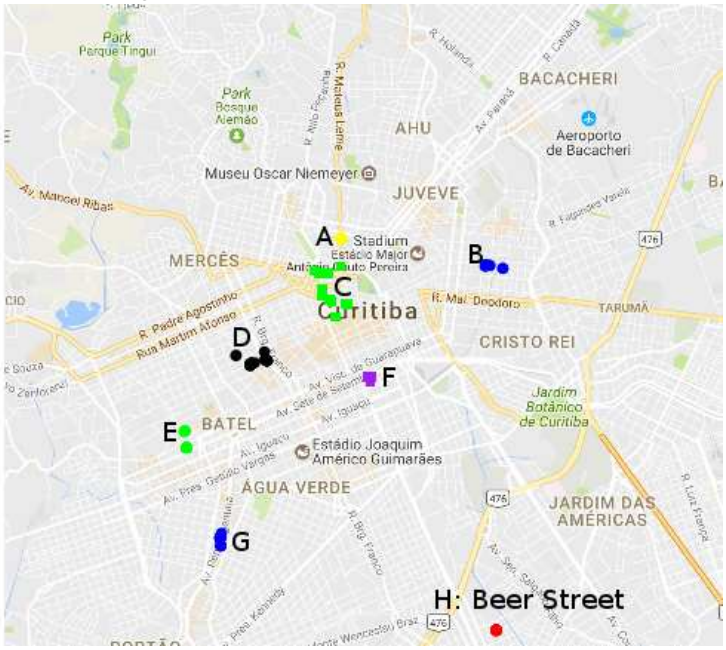
Foi definido o parâmetro eps como 250 metros e considerado $minPts = 10$ para todos os conjuntos de dados. No entanto, também são apresentados os resultados considerando $minPts = 5$ para o Dataset2013, uma vez que este conjunto de dados é menor. A figura 1 mostra os resultados do agrupamento. É possível observar na Figura 1a, que mostra os resultados de agrupamento para o Dataset2017, que o algoritmo encontrou oito clusters (todos eles são rotulados com uma letra). Observe que as cores e os símbolos são usados apenas para diferenciar os clusters. Entre os clusters, a área onde a Rua da Cerveja será construída foi identificada (veja o cluster H). Como a área da Rua da Cerveja foi encontrada pelas análises, esta é a primeira evidência de como os dados das mídias sociais poderiam ser utilizadas no planejamento.

Se os tomadores de decisão de Curitiba estivessem usando a estratégia discutida aqui para ajudar a tomar melhores decisões de planejamento urbano, neste momento eles teriam essas oito áreas candidatas para criar a Rua da Cerveja. A primeira observação é que todos esses clusters são áreas importantes em relação ao consumo de cervejas artesanais, e isso surgiu a partir dos dados que expressam implicitamente a preferência dos usuários. Conforme foi verificado, existem outras áreas para consumo de cerveja artesanal, no entanto, as áreas identificadas emergiram como as mais populares. Isto significa que a cidade pode concentrar esforços em áreas mais estratégicas. Em seguida, no artigo [12], foram apresentados mais detalhes de como tomadores de decisão poderiam atuar no caso específico, e esses passos poderiam ser utilizados para outras situações. Um ponto importante a ressaltar aqui é o potencial da abordagem. As pessoas já estão usando os espaços e, potencialmente, criando pontos estratégicos para diversas características. Com dados de redes sociais essas áreas talvez possam ser identificadas e exploradas.

O mesmo processo de agrupamento foi realizado para Dataset2013 (Figura 1b). Note-se que, para $eps = 250$ metros e $minPts = 10$, o único cluster encontrado foi a área onde a Rua da Cerveja será criada. Como podemos ver, a área tem sido popular há muito tempo e ainda popular hoje. Este pode ser outro critério a ser considerado pelo tomador de decisão. É esperado encontrar menos clusters nesta configuração do Dataset2013, devido à restrição de $minPts = 10$ (o

[1] http://github.com/mhahsler/dbscan.

Um brinde ao Untappd! Usando preferências por cervejas para o planejamento urbano e estudo de diferenças culturais WebMedia'17: Workshops e Pôsteres, WTIC, Gramado, Brasil



(a) 2016-2017 - minPts =10, eps=250 metros

(b) 2013 - minPts = 10, eps=250 metros

(c) 2013 - minPts = 5, eps=250 metros

Figura 1: Clusters encontrados em Curitiba usando dados do Untappd. Imagem de [12].

Dataset2013 é menor). Na prática, o *minPts* pode ter que ser adaptado para diferentes conjuntos de dados ou metas, o tomador de decisão deve ter esse controle. Por esse motivo, também foi considerado $minPts = 5$ para Dataset2013 (Figura 1c). Desta forma também outros clusters, *D*, *E* e *F* da Figura 1a, também foram encontrados.

5 IDENTIFICAÇÃO DE DIFERENÇAS CULTURAIS

Em outra etapa do trabalho foi estudado o potencial para explorar as preferências de cerveja para identificar diferenças culturais. Para isso foram selecionadas cidades populares em diversas partes do mundo (entre parênteses é mostrado o número de check-ins obtido): Cidade do México (1.004), Chicago (6.493), Los Angeles (2.210), Nova York (8.080), Portland (6.451), São Francisco (3.062), Belo Horizonte (436), Curitiba (465), Rio de Janeiro (1.070), São Paulo, 2.555, Tokyo (2.764), Berlim (771), Bruxelas (1.160), Barcelona

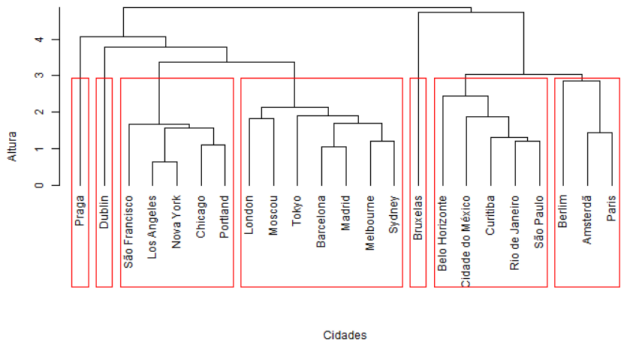


Figura 2: Dendograma para o agrupamento realizado com as cidades consideradas.

(1.877), Madrid (770), Paris (383), Amsterdã (1.556), Dublin (837), Londres (6.658), Praga (476), Moscou (857), Melbourne (1.808) e Sydney (2.597).

Nesta análise, as preferências dos usuários são calculadas de acordo com as notas informadas no check-in. Com isso, identificou-se a necessidade de eliminar as notas zeradas. Isto se deve à impossibilidade de identificar quando um usuário realmente atribuiu uma nota zero ou esqueceu-se de avaliar uma cerveja fornecendo uma nota a ela. Este fato pode vir a gerar falsos positivos, impactando de forma negativa os resultados e análises da pesquisa. O dataset não possui classes de cervejas provenientes do Untappd e sim somente o nome da cerveja. Outro passo realizado no dataset foi a criação de uma classificação de acordo com o material disponibilizado pela Brewers Association [1], que agrupa as cervejas por características étnicas.

Além disso, para o agrupamento de regiões com preferência por cervejas similares foram realizados os seguintes passos. Primeiramente, cada cidade c é representada por um vetor de preferência composto de nove classes de cervejas (*features*). Em seguida calcula-se a distância entre cada uma das cidades com base nesse vetor de preferências utilizando a distância canberra [7, 9]. Finalmente, realizamos um agrupamento hierárquico com um critério de agrupamento por *complete linkage* [4]. O resultado é apresentado na forma de um dendograma [10], que pode ser observado na Figura 2.

Como é possível observar com base nesta figura, as cidades dos Estados Unidos ficaram agrupadas no mesmo cluster. Isso também é válido para as cidades brasileiras. Note ainda que outras cidades do mesmo país estão, no geral, muito próximas uma das outras. Isso demonstra o potencial de exploração dessas características para o estudo de diferenças culturais.

Julgamos que as os critérios utilizados nesta avaliação são interessantes para o problema estudado. No entanto, outros critérios poderiam ser utilizados. Essa avaliação ainda será realizada na pesquisa em andamento. Além disso, é ainda trabalho futuro da pesquisa a avaliação de outros métodos de agrupamento, mas os resultados atuais já são bastante promissores.

6 CONCLUSÕES

Neste trabalho em andamento, foram explorados dados da LBSN Untappd em duas direções principais. Primeiramente, no contexto do planejamento urbano em Curitiba investigando o caso da criação da Rua da Cerveja anunciada pela prefeitura, onde apresentou-se que uma estratégia com dados de redes sociais poderia auxiliar na tomada de decisão na criação desse tipo de atração. Além disso, foi possível identificar o potencial de explorar as preferências por tipos de cervejas para estudar comportamentos sociais urbanos, particularmente relacionados à identificação automática de aspectos culturais. Os resultados sugerem que a preferência dos usuários por tipos de cervejas é uma característica cultural importante que pode ser explorada, por exemplo, na construção de novas aplicações da Web social.

AGRADECIMENTOS

Este trabalho foi parcialmente financiado pelo projeto CNPq UrbComp, processo número 403260/2016-7, bem como pela Fundação Araucária e CNPq.

REFERÊNCIAS

[1] 2017. *Brewers Association 2017 beer style guidelines*. Citado na página 41.
[2] Martin Chorley, Luca Rossi, Gareth Tyson, and Matthew Williams. 2016. Pub crawling at scale: tapping Untappd to explore social drinking. In *Proc. of ICWSM'16*.
[3] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Proc. of KDD'96*, Vol. 96. 226–231.
[4] Edward B Fowlkes and Colin L Mallows. 1983. A method for comparing two hierarchical clusterings. *Journal of the American statistical association* 78, 383 (1983), 553–569.
[5] Ruth Garcia-Gavilanes, Daniele Quercia, and Alejandro Jaimes. 2013. Cultural Dimensions in Twitter: Time, Individualism and Power. In *Proceedings of ICWSM'13*. Boston, USA.
[6] Nadav Hochman and Raz Schwartz. 2012. Visualizing Instagram: Tracing Cultural Visual Rhythms. In *Proceedings of Workshop on Social Media Vis.* AAAI, Dublin, Ireland, 6–9.
[7] Giuseppe Jurman, Samantha Riccadonna, Roberto Visintainer, and Cesare Furlanello. 2009. Canberra distance on ranked lists. In *Proceedings of Advances in Ranking NIPS 09 Workshop*. 22–27.
[8] Felix Kling and Alexei Pozdnoukhov. 2012. When a city tells a story: urban topic analysis. In *Proc. of the 20th international conference on advances in geographic information systems*. ACM, 482–485.
[9] Godfrey N Lance and William T Williams. 1967. Mixed-Data Classificatory Programs I - Agglomerative Systems. *Australian Computer Journal* 1, 1 (1967), 15–20.
[10] Oded Maimon and Lior Rokach. 2005. *Data mining and knowledge discovery handbook*. Vol. 2. Springer.
[11] Daniel Preoţiuc-Pietro and Trevor Cohn. 2013. Mining user behaviours: a study of check-in patterns in location based social networks. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 306–315.
[12] Ville Santala, Sandro Miczevski, Saulo A de Brito, Ariane Lao, Tatiana Gadda, Nadia Kozievitch, and Thiago H Silva. 2017. Making Sense of the City: Exploring the Use of Social Media Data for Urban Planning and Place Branding. In *Proc. of Workshop de Computação Urbana (CoUrb)*. Belém, Brasil.
[13] Raz Schwartz, Mor Naaman, and Ziad Matni. 2013. Making sense of cities using social media: Requirements for hyper-local data aggregation tools. In *Proc. of ICWSM'13*. 15–22.
[14] Thiago Silva, Pedro Vaz de Melo, Jussara Almeida, Mirco Musolesi, and Antonio Loureiro. 2014. You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food e Drink Habits in Foursquare. In *Proc. of ICWSM*. Ann Arbor, USA.
[15] T. H. Silva and A. R. Graeml. 2016. Exploring Collected Intelligence from Untappd to Support the Location Decision for New SMEs. In *Proc. of Brazilian Symposium on Multimedia and the Web*.
[16] Thiago H. Silva, Pedro O. S. Vaz de Melo, Jussara M. Almeida, Juliana Salles, and Antonio A. F. Loureiro. 2014. Revealing the City That We Cannot See. *ACM Trans. Internet Technol.* 14, 4, Article 26 (Dec. 2014), 23 pages. https://doi.org/10.1145/2677208
[17] Shoko Wakamiya, Ryong Lee, and Kazutoshi Sumiya. 2011. Crowd-based urban characterization: extracting crowd behavioral patterns in urban areas from twitter. In *Proc. of 3rd ACM SIGSPATIAL international workshop on location-based social networks*. ACM, 77–84.