Identifying exceptional cultural profiles in the world using Foursquare data

Paulo Henrique de S Coelho
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos 6627
Belo Horizonte, Minas Gerais, Brasil
paulohdscoelho@dcc.ufmg.br

Pedro O. S. Vaz de Melo
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos 6627
Belo Horizonte, Minas Gerais, Brasil
olmo@dcc.ufmg.br

Mario S Alvim
Universidade Federal de Minas Gerais
Av. Pres. Antônio Carlos 6627
Belo Horizonte, Minas Gerais, Brasil
msalvim@dcc.ufmg.br

ABSTRACT

The use of Location-Based-Networks allows diverse and unknown people around the world to daily record and share they activities, tastes and habits making them accesible both by other users and potential researchers. Analyzing the activity of the members in those networks, it's possible to identify shared behavior patterns that determine the formation of virtual groups of people who are culturally close even not knowing each other, and whose boundaries cross over the political and natural borders of countries in the real world. In this paper, we propose the development of a information model capable of explain how those virtual groups are formed and what is it's connexion with typical cultural habits that define certain countries and societies.

KEYWORDS

Data-Mining, Culture, Location-Based-Networks

1 INTRODUÇÃO

O conceito de Cultura pode ser entendido como o conjunto das práticas, gostos e hábitos de um determinado grupo de pessoas. O uso disseminado de Redes Sociais de Geolocalização (RSGL) na atualidade alavanca o acesso a estas informações (que chamaremos de culturais) em larga escala, sem a necessidade da aplicação de questionários. Lemmerich et. al define Comportamento Excepcional como *"padrões não usuais de comportamento entre subgrupos pertencentes a uma determinada amostra que diferem significativamente do comportamento geral"*[5]. É proposto neste trabalho o desenvolvimento de um modelo informacional capaz de medir comportamentos excepcionais presentes em dados de checkins do Foursquare.

Esse trabalho apresenta o resumo para o pôster apresentado no XXIII Simpósio Brasileiro de Sistemas Multimídia e Web e se organiza da seguinte maneira: a Seção 2 apresenta os trabalhos relacionados que orientaram o presente artigo; a Seção 3 apresenta a metodologia, com uma descrição da base de dados e seus desafios, as métricas de análise empregadas e as tecnologias utilizadas; a Seção 4 traz os resultados obtidos. A Seção 5 apresenta as conclusões e trabalhos futuros. Tabelas e Gráficos dos resultados podem ser consultados no pôster.

In: Sessao de Pôsteres do WebMedia'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.

2 TRABALHOS RELACIONADOS

Existem trabalhos importantes sobre modelos culturais e Redes Sociais Complexas que nos orientaram, como o de Kenneth Joseph e Kathleen Carley[4], que apresenta a validação de um modelo cultural que propõe que a influência das preferências culturais do indivíduo é relevante para sua rede de relacionamentos. Em outro trabalho importante, Axelrod[1] propõe um modelo de convergência cultural baseado em convergência local de indivíduos e interseção de features para criar uma polarização global de clusters culturais; conceito similar à presente pesquisa, no entanto, não focamos na interação entre usuários mas sim em como seu comportamento individual se aproxima do de outras pessoas e como eles se aproximam para criar um cluster cultural.

No estudo da cultura utilizando RSGL o trabalho desenvolvido por Mueller et. al[6] é muito importante pois utiliza uma base similar à desse trabalho mas enfocando no gênero dos usuários como fator determinante da criação dos perfis culturais. Ainda nessa mesma área, o artigo de Silva et. al 2014[7] realiza o trabalho de identificar a partir checkins em determinadas categorias de comidas e bebidas, quão culturalmente próximos são pessoas de diferentes Países e cidades, sem contudo, propor um modelo informacional para identificar tais padrões. O presente trabalho não enfoca na temporalidade, como a análise mais de Chorley et. al[2] que, utilizando uma RSGL específica para bebidas, analisa os hábitos alcoólicos de seus usuários. e sim na agregação, localização e correlação dos dados, similar ao realizado por Han et. al[3] que foca na mobilidade e dinâmica de grupos sociais para entender e explicar seus comportamentos e formação.

3 METODOLOGIA

3.1 Tecnologias utilizadas

Todos os *scripts* foram programados em *Python* os dados são armazenados e controlados pelo sistema de controle de versão *GitHub*. As métricas de análise e distância podem ser encontradas nas bibliotecas *Python* de Mineração de dados, particularmente **SciPy**, **Numpy** e **MatplotLib**, responsável pela plotagem dos gráficos apresentados.

3.2 Base de dados

Foi utilizada uma base de dados proveniente de checkins no Foursquare, obtida em uma única semana de Abril de 2012. Como checkins do Foursquare não são públicos, foi utilizada uma abordagem baseada em mineração de tuítes, na qual foram obtidos cerca de 4.7 milhões de tuítes contendo uma URL para o site do Foursquare

com informações sobre o checkin, particularmente a localização geográfica, a categoria e a data e hora da atividade.

As categorias do Foursquare à época da mineração se dividiam em 8 tipos: Arts & Entertainment; College & University; Professional & Other Places; Residences; Great Outdoors; Shops & Services; Nightlife Spots; e Food. Cada categoria possui subcategorias. Por exemplo: Rock Club and Concert Hall são subcategorias de Nightlife Spots[7]. A base de dados retornou um total de 19694387 checkins, distribuídos entre 1296453 usuários, 673 categorias e 115 países.

3.3 Métricas de análise empregadas

Kullback-Leibler Divergence (KLD) é a medida do quanto determinada distribuição de probabilidade diverge de outra segunda esperada. É geralmente chamada de O *Ganho de Informação* obtido se a distribuição P é usada ao invés de uma distribuição Q e também pode ser entendida como a Entropia relativa de P em relação a Q. A sua fórmula é dada por:

$$D_{KL}(P||Q) = \sum_i P(i) log \frac{P(i)}{Q(i)}$$

em que $P(i)$ são os elementos de P e $Q(i)$ os elementos de Q.

Os dados por País foram representados por vetores, cada célula correspondendo à soma total dos checkins da categoria. Tais vetores foram posteriormente modelados em (*Probability Mass Function*) contendo a probabilidade dos checkins ocorrerem em determinadas categorias. O vetor do País é denominado como P e a soma de todos os vetores como Global ou Q. Aplicando a métrica KLD sobre P e Q foram obtidos os resultados do grau de distância que cada P se encontra em relação a Q.

Para comparar perfis culturais excepcionais a mesma abordagem foi utilizada; no entanto, as distâncias foram calculadas entre as PMFs da distribuição dos checkins em uma dada categoria (e.g. bar) por país, ou seja, a dimensão utilizada passou a ser a Categoria, ao invés de País. O comportamento global considera a PMF da distribuição de todos os checkins por País; neste caso, quando a distância for muito grande entre as PMFs, então a categoria em particular é significativamente mais (ou menos) popular em determinados Países que o esperado.

Para analisar a correlação entre os dados, foi modelada uma *Matriz de Features* $M(mxn)$ onde m é o número de países e n o de categorias. Sobre M foi calculada a Co-Ocorrência entre categorias através da soma dos valores mínimos de checkins, ou seja, para cada valor K_i é calculado o produto cartesiano com as outras $K_i, \ldots, K_n$ categorias, sendo adicionado na célula correspondente à categoria K_i o mínimo do produto cartesiano.

A métrica *z-score* indica a quantidade de Desvios Padrão que determinado elemento se encontra em relação à média é calculada da seguinte forma: Dada a co-ocorrência de determinada categoria, encontra o quão excepcional ela é de acordo com seu valor de z-score em relação à média e é dado pela fórmula: $z = (X - \mu)/\sigma$ onde z é o z-score, X o valor do elemento, μ é a média da população e σ o desvio-padrão. Além do grau de surpresa sobre as Co-ocorrências, a aplicação dessa métrica nos ajuda encontrar em determinado perfil de categorias quais categorias podem ser consideradas excepcionais dado o seu valor de *z-score*.

4 RESULTADOS OBTIDOS

Ao aplicar o KLD sobre os dados de país foi obtido um mapa no qual percebe-se que países com maiores valores e, portanto, mais peculiares em relação ao pefil global são reconhecidamente peculiares: Iraque, Nepal, Costa do Marfim e Sudão por exemplo. O contrário acontece com os países menos peculiares, onde encontramos dois dos Países que sabemos possuem uma representatividade maior na amostra: Turquia e Estados Unidos. Entretanto, há a presença de Países com pouca representação, como Reino Unido, Alemanha e Holanda. Notamos o contraste social e econômico entre os dois grupos: o primeiro mais pobre em relação ao segundo; o que reflete questões culturais mas também econômicas, como o acesso à tecnologia.

Já na visão das categorias, os resultados apontam *Categorias Neutras* cujos checkins não nos dão uma informação muito relevante: são locais onde é esperado que as pessoas estarão, como Compras, Comida e Trabalho. Tais categorias retornam valores baixos de KLD sendo que sua relação com o perfil global é de dominância. Já as categorias com maior valor e, consequentemente mais excepcionais são tão específicas de um País que também não nos dizem muita coisa. Categorias como Huaiyan, Cretan Restaurants e Kafenia são locais que refletem uma prática cultural muito específica de determinado país. A categoria Huaiyang por exemplo ocorre apenas na China, com 4 checkins, enquanto Cretan Restaurants ocorre apenas na Grécia.

Dada a natureza massiva dos dados, a aplicação do z-score resultou uma matriz com dimensionalidade alta. Para contornar esse problema aplicamos uma técnica chamada *TSNE* para redução da dimensionalidade da matriz mantendo as proximidades originais. Realizando várias vezes a plotagem dos z-scores através do TSNE, embora a configuração dos pontos mude, a proximidade das categorias não, refletindo certas categorias que são agrupadas sempre próximas umas às outras, batizadas de categorias com afinidade. Os resultados mostram também a clara existência de clusterização em temática, como categorias escolares, de transporte ou por região.

5 CONCLUSÕES

Os resultados obtidos até o presente nos permitem perceber que os comportamentos culturais que definimos como excepcionais de países e categorias apontam para a e existência de determinadas características e sociedades peculiares que se relacionam, apesar da distância seja geográfica ou cultural.

Ainda não é possível validar completamente os dados ou determinar a métrica de análise mais adequada para os resultados; mas esperamos que ao dar prosseguimento à pesquisa seja possível desenvolver um modelo informacional de inferência automática de perfis culturais que nos permita entender melhor como se formam tais perfis e quais características presentes no comportamento dos usuários são fundamentais para a sua formação.

REFERÊNCIAS

[1] Robert Axelrod. 1997. The Dissemination of Culture. *Journal of Conflict Resolution* 41, 2 (1997), 203–226. https://doi.org/10.1177/0022002797041002001 arXiv:http://dx.doi.org/10.1177/0022002797041002001

[2] Martin Chorley, Luca Rossi, Gareth Tyson, and Matthew Williams. 2016. Pub crawling at scale:tapping untappd to explore social drinking. In *Proceedings of the 10th International Conference on Web and Social Media, ICWSM 2016*. AAAI, 62–71. http://publications.aston.ac.uk/27836/ -© 2015, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

[3] Jungkyu Han and Hayato Yamana. 2015. Why People Go to Unfamiliar Areas?: Analysis of Mobility Pattern Based on Users' Familiarity. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services (iiWAS '15)*. ACM, New York, NY, USA, Article 28, 10 pages. https://doi.org/10.1145/2837185.2837190

[4] Kenneth Joseph and Kathleen Carley. 2015. Culture, Networks, Twitter and foursquare: Testing a Model of Cultural Conversion with Social Media Data. (2015).

[5] Florian Lemmerich, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. 2016. Mining Subgroups with Exceptional Transition Behavior. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 965–974. https://doi.org/10.1145/2939672.2939752

[6] Willi Mueller, Thiago H. Silva, Jussara M. Almeida, and Antonio AF Loureiro. 2017. Gender matters! Analyzing global cultural gender preferences for venues using social sensing. *EPJ Data Science* 6, 1 (19 May 2017), 5. https://doi.org/10.1140/epjds/s13688-017-0101-0

[7] Thiago H Silva, Pedro OS Vaz de Melo, Jussara M Almeida, Mirco Musolesi, and Antonio AF Loureiro. 2014. You Are What You Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food and Drink Habits in Foursquare. In *Eighth International AAAI Conference on Weblogs and Social Media*.