Sobre a Extração de Fronteiras Culturais Considerando Hábitos Alimentares Observados na Web Social

Thiago H Silva
Universidade Tecnológica Federal do Paraná
Departamento Acadêmico de Informática
Curitiba, Brasil
thiagoh@utfpr.edu.br

Jussara M Almeida
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Belo Horizonte, Brasil
jussara@dcc.ufmg.br

Antonio A F Loureiro
Universidade Federal de Minas Gerais
Departamento de Ciência da Computação
Belo Horizonte, Brasil
loureiro@dcc.ufmg.br

ABSTRACT

New approaches to study urban social behavior use Foursquare check-ins to represent user's preferences. In this direction, recently, researchers have proposed a method for identifying cultural boundaries. Our study is based on that methodology, aiming to validate the results and to study some variations. We use a newer dataset to evaluate the results obtained previously. We found that the cultural separation results using our dataset agree with those presented previously. Furthermore, we evaluated the impact of the data observation window size in the results. Finally, we study two additional variations in the studied methodology. The cultural separation quality obtained using these variations is lower compared with the results obtained by the original approach. The results reinforce that, in fact, the methodology originally proposed might be useful to complement large-scale studies on cultural differences. Automatic identification of cultural differences is a valuable information that can enable the creation of new ubiquitous applications.

Author Keywords

Social Web; Foursquare; Culture; Food and Drink; Evaluation

ACM Classification Keywords

J.4 Computer Applications: Social and Behavioral Sciences; H.4 Information Systems Applications: Miscellaneous

INTRODUÇÃO

As formas tradicionais para estudar o comportamento social urbano, por exemplo questionários, podem ser um problema para a realização de estudos em larga escala. Recentes estudos, dentre eles [3, 4, 10, 11, 14, 16], revelaram uma nova forma de obtenção de dados considerando a Web Social, particularmente através de redes sociais baseadas em localização (LBSNs), que pode revolucionar o estudo do comportamento social urbano.

Especificamente em [16] os autores propuseram a utilização de dados públicos disponíveis a partir da LBSN Foursquare para mapear as preferências individuais de usuários. Isto é interessante porque um *check-in*[1] em uma LBSN expressa a preferência de um usuário por um determinado tipo de lugar. Além disso, LBSNs são acessíveis em quase todos os lugares e por qualquer pessoa, amenizando o problema de escalabilidade e permitindo que dados em diversas regiões do mundo sejam coletados.

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.

Sabemos que os hábitos alimentares e de bebidas são capazes de descrever fortes diferenças entre as pessoas [1]. Com base nisso, o objetivo de Silva et al. [16] foi propor uma nova metodologia para a identificação de fronteiras culturais e semelhanças entre sociedades, considerando hábitos alimentares e de bebida. Para isso, foram usados check-ins do Foursquare para representar as preferências do usuário em relação ao que se come e bebe localmente, por exemplo, em uma determinada cidade. Os autores estudaram como essas preferências mudam de acordo com a hora do dia e localizações geográficas. A partir disso criaram uma metodologia para a identificação de culturas semelhantes, que pode ser aplicada a regiões de tamanhos variados, como países, cidades ou até mesmo bairros.

Com isso, este presente trabalho visa avaliar a metodologia para o estudo de diferenças culturais proposta em [16].

[1] Ato de disponibilizar o local onde você se encontra para seus amigos.

WAIHCWS'16 was held as part of IHC'16, organized by the Brazilian Computing Society (SBC). October 04, 2016, São Paulo/SP, Brazil.

Nós executamos uma avaliação na metodologia estudada em vários aspectos. Usamos um dataset mais recente, ainda não utilizado, para avaliar os resultados obtidos anteriormente seguindo a mesma metodologia proposta. Verificamos que os resultados de separação cultural utilizando o nosso dataset concorda com os apresentados anteriormente. Além disso, avaliamos o impacto do tamanho da janela de observação dos dados nos resultados. Esta análise forneceu indicações de que a identificação das fronteiras culturais usando um tamanho de janela maior do que uma semana não se altera significativamente. Finalmente, avaliamos duas variações adicionais na metodologia estudada.

A correta identificação de fronteiras culturais é útil em muitas áreas e aplicações, incluindo aplicações ubíquas. Por exemplo, uma aplicação que pode utilizar a informação cultural é um sistema de recomendação de locais, o que é útil para os visitantes e moradores de uma cidade. Com base nessa informação, sistemas como o Foursquare e outros buscadores baseados em localização, como o proposto em [13], poderiam se beneficiar com a introdução de novos critérios e mecanismos em seus sistemas de recomendação considerando as diferenças culturais entre as áreas. Por exemplo, uma pessoa que gosta de uma área específica de Manhattan poderia receber uma recomendação de uma área similar ao visitar Londres.

O resto do trabalho está organizado da seguinte forma. A Seção 2 apresenta alguns dos trabalhos relacionados. A Seção 3 apresenta os datasets utilizados. A Seção 4 descreve a metodologia de agrupamento de regiões de acordo com a informação cultural. A Seção 5 avalia o impacto nos resultados para datasets cobrindo diferentes períodos. A Seção 6 propõe e avalia duas variações na metodologia original. Por fim, a Seção 7 apresenta as conclusões do trabalho.

TRABALHOS RELACIONADOS

A utilização de dados da Web Social para o estudo do comportamento social urbano é um tema recente de pesquisa. Essa fonte de dados é interessante, pois permite a realização de estudos em larga escala. Nessa direção, vários estudos concentraram em estudar as propriedades espaciais de dados compartilhados em redes sociais baseadas em localização, como o Foursquare [2, 9, 12]. No entanto, tais esforços visam principalmente a investigação de padrões de mobilidade do usuário ou propriedades de redes sociais e suas implicações. Salles et al. [11] estudaram o uso do Foursquare nas maiores cidades do Brasil, levando em consideração fatores socioeconômicos destas cidades. Cranshaw et al. [3] também consideraram dados do Foursquare para delimitar áreas da cidade em relação ao comportamento dos usuários da rede social estudada.

Além disso, estudos também mostraram como o uso de sistemas da Web social pode variar entre os países. Por exemplo, Hochman et al. [5] 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. [4] 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. [16] propuseram uma nova metodologia para a identificação de fronteiras culturais e semelhanças entre populações, considerando hábitos de comida e bebida. No entanto, os autores avaliaram essa metodologia considerando um dataset que abrange uma semana de dados. Apesar dos resultados serem promissores, uma melhor avaliação dessa metodologia ainda é necessária. O presente estudo baseia-se no trabalho [16] e visa avaliar a metodologia proposta de diversas formas. Um estudo preliminar realizado também por Silva et al. [15] forneceu mais indícios de que a metodologia apresentada em [16] é promissora. Este presente estudo complementa esses trabalhos trazendo uma análise mais robusta, provendo indícios mais fortes de que a metodologia apresentada em [16] é uma opção interessante para a extração de fronteiras culturais.

Estudos interculturais (isto é, o estudo das diferenças culturais) não constituem uma nova área de pesquisa. Na verdade, eles já vem sendo realizados por pesquisadores que trabalham nas ciências sociais, particularmente em antropologia cultural e psicologia [8]. No entanto, identificar de forma automática essas diferenças culturais é uma informação valiosa que pode habilitar novas aplicações ubíquas.

DESCRIÇÃO DOS DADOS

Analisamos um dataset do Foursquare, que é uma rede social baseada em localização bastante popular. Nesse sistema os usuários podem disponibilizar para seus amigos os seus locais visitados, os chamados check-ins. Os dados do Foursquare foram coletados através do Twitter[2], que é um serviço de *microblogging*, ou seja, ele permite que os seus usuários enviem e recebam atualizações pessoais de outros contatos em textos de até 140 caracteres, conhecidos como "*tweets*". Além de *tweets* de texto simples, os usuários também podem compartilhar localizações (ou check-ins) a partir de uma integração com o Foursquare. Neste caso, check-ins do Foursquare anunciados no Twitter passam a ficar disponíveis publicamente, o que por padrão não acontece quando o check-in é publicado unicamente no sistema do Foursquare. Nós consideramos dois datasets que representam dois períodos de tempo distintos: dataset 1 *(D1)* e dataset 2 *(D2)*. O dataset *(D1)* foi o mesmo utilizado em [16], com isso temos acesso aos mesmos dados usados na criação da metodologia que estamos estudando. Esse dataset é referente a uma semana de dados de maio de 2012. O dataset *(D2)* foi coletado por nós, não sendo utilizado previamente.

[2] http://www.twitter.com.

Em nossos datasets, cada check-in consiste da latitude e longitude, do identificador do usuário, da categoria do local, bem como do momento em que o check-in foi feito. Locais do Foursquare são agrupados em oito categorias: *Arts & Entertainment*; *College & University*; *Professional & Other Places*; *Residences*; *Great Outdoors*; *Shops & Services*; *Nightlife Spots*; e *Food*. Cada categoria, por sua vez, tem subcategorias. Por exemplo, *Rock Club* e *Concert Hall* são subcategorias de *Nightlife Spots*.

Como estamos interessados principalmente no que as pessoas comem ou bebem, nós agrupados manualmente as subcategorias de locais, disponíveis em nossos datasets, relacionadas com três classes: Bebida, Fast Food e Slow Food. Após essa separação, a classe Bebida resultou 21 subcategorias (por exemplo, *sake place*, *karaoke bar* e *pub*), ao passo que a classe Fast Food resultou em 27 subcategorias (por exemplo, *bakery*, *burger joint* e *wings joint*) e a classe Slow Food em 53 subcategorias, incluindo *Chinese restaurant*, *steakhouse* e *Greek restaurant*.

Table 1. Correlação (Spearman) do número de check-ins dados em cada subcategoria dos datasets D1, D2 e D3.

Classe Bebida	
Datasets usados	ρ (p-value)
D2, D3	0.99 (0)
D3, D1	0.94 (5.4e-07)
Classe Fast Food	
Datasets usados	ρ (p-value)
D2, D3	0.99 (0)
D3, D1	0.8 (1.2e-05)
Classe Slow Food	
Datasets usados	ρ (p-value)
D2, D3	0.99 (0)
D3, D1	0.96 (0)

D1 abrange uma única semana de Abril de 2012. O outro dataset, D2, abrange um período maior e mais recente: de 24 de abril de 2014 a 18 de junho de 2014. Ter acesso a dataset maior é particularmente interessante porque nos permite estudar a metodologia para capturar fronteiras culturais em diferentes janelas de observação. Nossa coleta de dados enfrentou alguns problemas, possivelmente, não capturando todos os dados compartilhados nos dias em que ocorreram problemas. Por esta razão, decidimos criar um novo dataset, dataset **D3**, que é um subconjunto de *D2* contendo apenas algumas semanas, sem dias com problemas de coleta, contendo as semanas 3, 7 e 8. Também é interessante ter esse dataset porque cobre parcialmente um evento mundial: Copa do mundo da FIFA de 2014 (semana 8).

A fim de estudar a semelhança dos nossos três datasets, nós correlacionamos o número de check-ins dados em cada uma das subcategorias de locais (para as classes Bebida, Fast Food e Slow Food), utilizando a correlação de Spearman. A Tabela 1 resume os resultados. Como podemos ver, os datasets D2 e D3 possuem alta correlação positiva. A correlação de D2 e D3 com D1 também é elevada. Apesar da sugestão de que D2 reflete corretamente o comportamento dos usuários, no resto do documento desconsideramos D2 na maioria das análises, a fim de evitar qualquer enviesamento nos resultados. Usamos D2 somente em uma análise específica relacionada com o tamanho das janelas de observação de dados, discutida na Seção 5, onde a incompletude deste dataset é uma característica interessante na avaliação.

Neste estudo consideramos as mesmas regiões, cidades e países estudadas em [16]. Ao todo analisamos 16 países em várias regiões do mundo (Argentina, Austrália, Brasil, Chile, Inglaterra, França, Indonésia, Japão, Coreia do Sul, Malásia, México, Rússia, Singapura, Espanha, Turquia e Estados Unidos), 27 cidades (Natal, Recife, Belo Horizonte, Rio de Janeiro, São Paulo, Manaus, Miami, Nova Iorque, Chicago, Dallas, Denver, Las Vegas, São Francisco, Paris, Londres, Istambul, Moscou, Banguecoque, Kuala Lumpur, Singapura, Jacarta, Bandung, Surabaia, Manila, Osaca e Tóquio), bem como regiões populares de Londres (8 regiões), Nova Iorque (8 regiões) e Tóquio (9 regiões). Para realizar essa separação de dados nós utilizamos as coordenadas geográficas do check-in e um sistema de informação geográfica.

METODOLOGIA PARA O AGRUPAMENTO DE ÁREAS

Para o agrupamento de regiões com hábitos alimentares e de bebida similares, utilizamos a mesma metodologia proposta em [16]. Primeiramente, cada área a é representada por um vetor de preferência composto de 808 características (*features*), ou seja, o número normalizado de check-ins em cada uma das 101 subcategorias consideradas em quatro períodos distintos do dia (madrugada, manhã, tarde e noite), durante a semana e nos fins de semana. Em seguida, aplicamos uma Análise de Componentes Principais (PCA) [7]. Finalmente, usamos o algoritmo k-means, uma técnica de agrupamento amplamente utilizada, para agrupar áreas no espaço definido pelos componentes principais encontrados.

Também seguindo a metodologia de [16], ao analisar países, definimos $k = 7$ (mesmo número de grupos, *clusters*, utilizados por Inglehart e Welzel [6], estudo que utilizou dados coletados de forma tradicional e agrupou países de acordo com aspectos culturais). Seguindo a mesma lógica, consideramos $k = 4$ para as cidades, uma vez que a ideia também é estudar cidades de 4 diferentes continentes e $k = 3$ para regiões dentro de uma cidade, porque consideramos 3 cidades nessa análise. Além desses valores de k computamos os grupos para $k = 2$ e $k = 10$ para todas as áreas consideradas, a fim de avaliar o resultado de agrupamento. Os parâmetros $k = 2$ e $k = 10$ são usados para estudar grupos "relaxados" e "compactos", respectivamente. Essa avaliação de grupos relaxados e compactos não foi feita por [16] e é interessante para entender a variabilidade dos grupos. Utilizamos a similaridade de cosseno para calcular a semelhança entre os locais.

Para ajudar na análise desses resultados, propomos neste trabalho um Índice de Similaridade de Grupo $c_{i,j}$ que

representa a semelhança entre um conjunto de grupos (*clusters*) (i) com outro conjunto de grupos (j). Este índice pode ser usado, por exemplo, para avaliar o quão boa é a correspondência entre os grupos obtidos utilizando o nosso novo dataset com o antigo. O Algoritmo 1 mostra os passos para calcular c. Este algoritmo analisa todos os pares de grupos que se deseja comparar. Para cada par ele calcula o número de elementos semelhantes (*hit*) entre os grupos, bem como o número de elementos diferentes (*miss*). O algoritmo usa esses valores para calcular um fator de desconto. O fator de desconto é usado para penalizar um agrupamento ruim, ou seja, grupos com valor baixo de "hit" e valor alto de "miss". O resultado de c tem um valor máximo de 1. Quanto mais perto de 1 mais semelhantes são os grupos comparados. O exemplo a seguir considera dois conjuntos hipotéticos de grupos, Clusters1 e Clusters2, para nos ajudar a entender o algoritmo. Clusters1: $(x, y, z), (a, b, c, d), (e, f)$. Clusters2: $(x, y, d), (a, b, c, z), (e, f)$. Resultado: $c_{1,2} = 0,68$. Explicação: $(2-1/2)+(3-1/3)+(2-0)$ (soma da interseção máxima com seu respectivo fator de desconto) dividido por 9 (número de elementos no total).

Algoritmo 1: Passos para calcular o índice de similaridade de grupo c.

```
listMaxs = []
paracada c1 em clusterSet1 fazer
    max = 0
    desconto = 0
    paracada c2 em clusterSet2 fazer
        hit = c1 ∩ c2
        se hit == 0 então
            continuar
        fim
        miss = tamanho(c2) − hit
        se miss ≠ 0 então
            desconto = miss/hit
        fim
        calc = hit−desconto
        se calc > max então
            max = calc
        fim
    fim
    listMaxs.append(max)
fim
c = soma(listMaxs)/numTotalElementosClusters
```

IMPACTO DO TAMANHO DA JANELA DE OBSERVAÇÃO

Ao reproduzir os resultados utilizando a metodologia mencionada acima, observamos que os resultados obtidos para o D1 e D3 são muito semelhantes. Com isso, uma pergunta natural é: qual é o impacto do tamanho da janela de observação nos resultados?

Lembre-se de que D1 tem uma semana completa, D3 tem três semanas completas, que estão contidas em D2 que tem oito semanas, mas algumas delas provavelmente não representam todos os dados que poderiam ser coletados. A fim de responder a questão colocada, investigamos o impacto nos resultados considerando cada semana de D2 individualmente. Esse tamanho da janela em particular foi escolhido para ir de acordo com o tamanho do dataset D1. A Figura 1 mostra o índice de similaridade de grupo para grupos obtidos utilizando cada semana individual do dataset D2 (1-8) com grupos obtidos utilizando D3, para os países (Figura 1a), cidades (Figura 1b) e regiões (Figura 1c). Os resultados referem-se a todos os valores de k considerados neste trabalho.

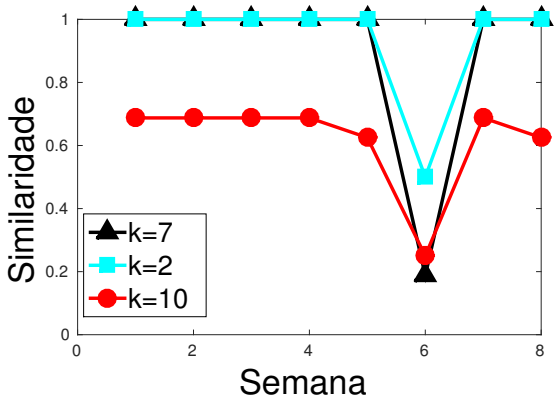
Encontramos $c = 1$, em todas as figuras, para a maioria dos grupos identificados para todos os valores de k, exceto $k = 10$. Outra ponto em comum em todos os resultados é o valor muito baixo de c para a semana 6. Isso é esperado porque esta semana em particular não possui quase nenhum dado (6 dias sem dados, o que representa uma janela de observação muito curta).

Considerando $k = 4$ no resultado de cidades (Figura 1b), temos dois casos de cidades que $c \neq 1$ (além da semana 6): utilizando a semana 1 e 7. Para ambos os casos o valor de c é $c = 0,7$ e os grupos são iguais aos grupos que foram encontrados utilizando D1. Isso sugere que as diferenças culturais observados usando D1 são representativas. Como esta é uma janela de observação pequena, as variações no comportamento das pessoas em qualquer situação atípica, por exemplo, más condições meteorológicas, são mais suscetíveis de serem capturadas. Este pode ser o caso de todas essas semanas mencionadas. Para as regiões considerando $k = 3$ o índice de similaridade de grupo é de $c = 1$ para todas as semanas, exceto a semana 6 (esperada) e na semana 7, mas que possui um valor de c muito alto: $c = 0,95$.

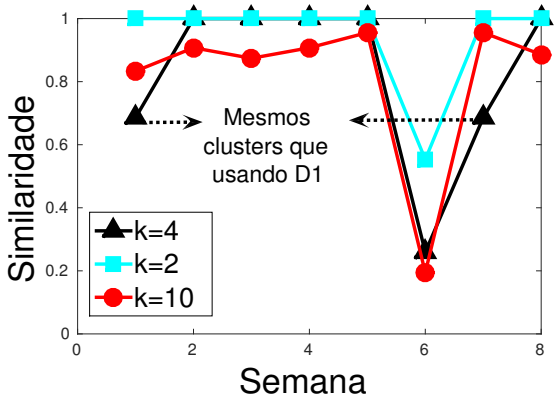
Analisando os resultados para $k = 10$ para países e cidades, observamos que a maioria dos grupos para todas as semanas são semelhantes entre si, explicando os valores similares de c, cerca de $0,7$ (para países) e $0,9$ (para cidades). Estes valores são consideravelmente altos, indicando que todos os grupos são semelhantes com os grupos encontrados usando D3.

Voltando nossa atenção para os grupos encontrados para regiões considerando $k = 10$, índices de similaridade de grupo baixos também foram observados, no entanto, com uma variação maior do que a observada para países e cidades. Este resultado pode ser explicado pelo fato de que as regiões, devido ao seu tamanho menor, tendem a ser mais suscetíveis à variação no comportamento das pessoas que vieram visitá-las, fato que talvez pode ser atenuado através de um dataset que abrange uma janela de tempo maior.

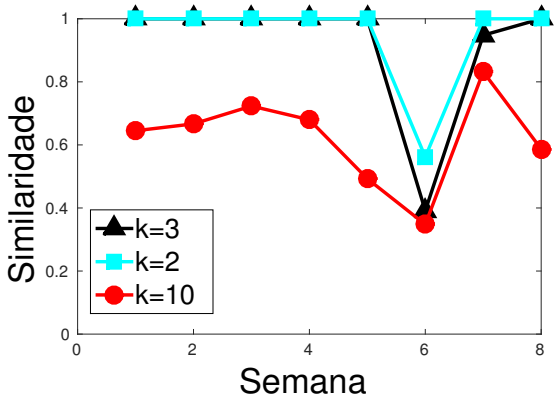
Todos estes resultados sugerem que os limites culturais identificados utilizando uma janela de observação maior do que uma semana não se alteraram significativamente, apesar da sugestão de que é possível obter resultados mais precisos quando considerando grupos mais compactos, ou seja, escolhendo um número grande de grupos para ser encontrado ($K = 10$ em nossos exemplos). Isso é especialmente válido em escala menor, tais como a nível de regiões. No entanto, usando um dataset que abrange consideravelmente menos do que uma semana,



(a) Países



(b) Cidades



(c) Regiões

Figure 1. Índice de similaridade de grupo dos grupos obtidos para cada semana individual de D2 (1 a 8) com os grupos para D3.

isto é, uma janela pequena de observação para capturar a rotina de usuários, tal como a semana 6, os resultados tendem a ser consideravelmente piores.

ANÁLISES ADICIONAIS

Nesta seção, o nosso objetivo é avaliar se a metodologia de agrupamento que estamos seguindo é satisfatória. Para isso, analisamos duas variações na abordagem original, o que poderia simplificar a abordagem original, melhorando o desempenho de processamento para um volume maior de dados.

Descrição das Análises

Nesta seção, nós ignoramos a dimensão tempo na nossa avaliação para propor duas análises adicionais (AA) para a identificação das fronteiras culturais.

- AA1: nesta análise o vetor de preferências dos usuários considera apenas os tipos de locais (subcategorias de lugares) apresentados em cada cidade. Não consideramos o número de check-ins realizados em cada local;

- AA2: nesta análise o vetor de preferências considera os tipos de locais, bem como a sua popularidade, isto é, que consideramos o número normalizado de check-ins realizados em cada uma das 101 subcategorias.

Com AA1 tentamos responder a pergunta: será que a existência de certos tipos de locais em uma área a são suficientes para explicar as diferenças culturais? AA2 nos ajuda a complementar a primeira questão, visando responder: a popularidade desses locais é útil/essencial nessa tarefa?

O resto da metodologia continua da mesma forma como apresentado na Seção 4. Em suma, agora representamos cada área a por um vetor de preferência como descrito em AA1 e AA2, desconsiderando a dimensão temporal. Em seguida, aplicamos a técnica PCA a esses vetores para obter os seus componentes principais. Finalmente, usamos o algoritmo k-means para agrupar áreas no espaço definido pelos componentes principais identificados. Nós realizamos essa análise para áreas que representam países, cidades e regiões. Para esta análise consideramos apenas o dataset D3.

Avaliando AA1

Ronald Inglehart e Christian Welzel propuseram um mapa cultural do mundo com base nos dados do World Values Surveys (WVS) 2005-2008 [6]. Além disso, ofereceram uma divisão do mundo em grupos, semelhante com o que fazemos neste trabalho. Primeiro estudamos os resultados obtidos para AA1. Os grupos encontrados para países considerando $k = 7$, $k = 2$ e $k = 10$ não vão de acordo com os dados do WVS e nem com o senso comum. Há sempre um grupo com o número máximo possível de acordo com o k. Em outras palavras, uma vez que temos 16 países, quando definimos $k = 7$ nós temos um grupo com 10 países e outros 6 grupos

Table 2. Índice de similaridade de grupo entre os grupos encontrados usando a metodologia original (considerando o dataset D3) e usando AA1. O índice é gerado para todos os tipos de áreas e k valores considerados.

Países	
Número de grupos (k)	$c_{aa1,D3}$
$k = 7$	0,18
$k = 2$	0,5
$k = 10$	0,31
Cidades	
Número de grupos (k)	$c_{aa1,D3}$
$k = 4$	0,39
$k = 2$	0,57
$k = 10$	0,32
Regiões	
Número de grupos (k)	$c_{aa1,D3}$
$k = 3$	0,39
$k = 2$	0,56
$k = 10$	0,34

contendo um país cada. Esses agrupamentos são praticamente selecionados aleatoriamente apenas para satisfazer o k escolhido, resultando em grupos muito diferentes dos observados considerando o WVS. De fato, o índice de similaridade de grupo encontrado considerando os grupos para $k = 7$ e os grupos encontrados para o WVS é: $c_{aa1,wvs} = 0,18$.

Como nós tendemos a ter muitos dados representando um país, o resultado insatisfatório para esta abordagem é esperado, pois é provável que encontremos todos os tipos de lugares (em nosso vetor de preferência) para todos os países. Por esta razão, as distâncias de cada vetor de preferências tendem a ser zero, tornando a qualidade do agrupamento muito baixa.

Nós também calculamos o índice de similaridade de grupo entre todos os resultados obtidos considerando AA1 com a metodologia original usando o dataset D3, gerando então $c_{aa1,D3}$, como mostra a Tabela 2. O índice é obtido para todos os tipos de áreas e valores de k considerados. Como podemos ver, os resultados para países, considerando todos os valores de k, obtidos com AA1 também são muito distintos daqueles obtidos com a metodologia original.

Avaliando AA2

Voltamos nossa atenção agora para os resultados obtidos para AA2. Estudando os resultados para países, observamos que eles vão consideravelmente mais de acordo com os encontrados por Ronald Inglehart e Christian Welzel usando dados do WVS do que aqueles obtidos utilizando AA1. No entanto, eles são menos precisos do que os resultados obtidos usando a metodologia original de [16]. Por exemplo, o grupo composto por Turquia e Austrália identificado utilizando a abordagem AA2 não é identificado utilizando a metodologia original de Silva e nem por Inglehart. Para estudar este caso a Tabela 3 mostra os valores de c entre os grupos encontrados usando a

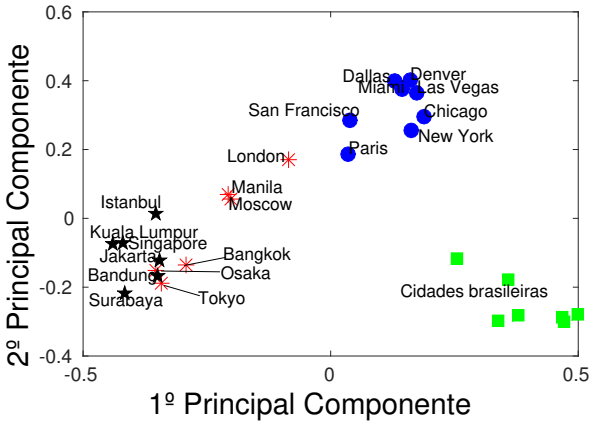


Figure 2. Resultados de agrupamento para cidades usando AA2 para $k = 4$ e considerando D3.

metodologia original e a abordagem AA2. Como podemos ver, os resultados para países não concordam consideravelmente com a metodologia original para $k = 7$ e $k = 10$.

Este não é o caso para cidades e regiões, casos em que os resultados com AA2 são mais semelhantes com aqueles obtidos utilizando a metodologia original. Apesar disso, a similaridade elevada não é sempre obtida. Além disso, utilizando a metodologia de [16] somos mais propensos a obter resultados que são esperados de acordo com o senso comum. Por exemplo, usando AA2 para $k = 4$, Figura 2, Londres foi agrupada com Bangkok, Tóquio, Manila, Moscou e Osaca, fato que não é observado usando a metodologia original. Além disso, a metodologia original tende a agrupar regiões dentro da mesma cidade melhor do que com AA2. Esta é outra indicação de que os resultados obtidos com a metodologia original separam melhor áreas distintas culturalmente.

Table 3. Índice de similaridade de grupo entre os grupos encontrados usando a metodologia original (considerando o dataset D3) e usando AA2. O índice é gerado para todos os tipos de áreas e k valores considerados.

Países	
Número de grupos (k)	$c_{aa2,D3}$
$k = 7$	0.4
$k = 2$	0.92
$k = 10$	0.59
Cidades	
Número de grupos (k)	$c_{aa2,D3}$
$k = 4$	0.95
$k = 2$	0.96
$k = 10$	1
Regiões	
Número de grupos (k)	$c_{aa2,D3}$
$k = 3$	1
$k = 2$	1
$k = 10$	0.88

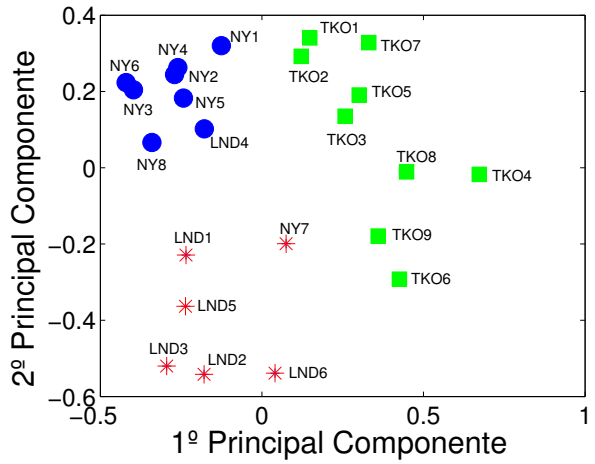


Figure 3. Resultados de agrupamento para regiões usando a metodologia original para o dataset D1 ($k = 3$) e considerando somente a classe Bebida.

Discussão

É importante ressaltar que a comparação realizada aqui com a metodologia original e as abordagens AA1 e AA2 foi em relação à identificação dos limites culturais. No entanto, temos que ter em mente que a redução da dimensão, na mesma direção usada nas abordagens AA1 e AA2, pode ser útil para obter outros tipos de informações sobre as áreas consideradas. A fim de deixar mais clara a utilidade de redução de dimensão, quando analisamos um subconjunto de características, por exemplo, hábitos de bebida durante os fins de semana em todas as regiões de Londres, Nova Iorque e Tóquio, resultado mostrado na Figura 3, nós descobrimos que algumas regiões de Londres e Nova Iorque são agrupadas. Isto é corroborado pelos resultados apresentados em [16]: para certas categorias, existem regiões de diferentes cidades que são muito semelhantes e, por isso, são agrupadas. Isto pode ser útil em um aplicativo, por exemplo, para sugerir áreas para consumir bebidas com os amigos.

CONCLUSÕES

Considerando datasets do Foursquare com diferentes volumes de dados e tamanhos de janela de observação, avaliamos uma metodologia para a identificação de fronteiras culturais em diferentes aspectos. Os resultados reforçam a sugestão de que a metodologia estudada, que usa de dados da Web social, particularmente sobre a preferência dos usuários por estabelecimentos alimentares, pode ser uma alternativa viável a métodos tradicionais para a extração de fronteiras culturais. Uma forma automática de identificação de fronteiras culturais pode habilitar a construção novas aplicações da Web social. Existem vários trabalhos futuros para este estudo, por exemplo, um estudo teórico do impacto da janela de observação de dados nos resultados, bem como a avaliação do impacto da qualidade dos dados utilizados nos resultados.

Agradecimentos

Este trabalho foi parcialmente financiado pelo projeto FAPEMIG-PRONEX-MASWeb, Modelos, Algoritmos e Sistemas para Web, processo número APQ-01400-14, bem como pelo Instituto Nacional de Ciência e Tecnologia para Web (INWEB), FAPEMIG, Fundação Araucária e CNPq.

REFERENCES

1. Carole, C. *Food And Culture: A Reader*, 2nd ed. Routledge, Dec. 1997.
2. Cho, E., Myers, S. A., and Leskovec, J. Friendship and mobility: user movement in location-based social networks. In *Proceedings of KDD'11*, ACM (San Diego, California, USA, 2011), 1082–1090.
3. Cranshaw, J., Schwartz, R., Hong, J. I., and Sadeh, N. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City. In *Proceedings of ICWSM'12* (Dublin, Ireland, 2012).
4. Garcia-Gavilanes, R., Quercia, D., and Jaimes, A. Cultural dimensions in twitter: Time, individualism and power. In *Proceedings of ICWSM'13* (Boston, USA, 2013).
5. Hochman, N., and Schwartz, R. Visualizing instagram: Tracing cultural visual rhythms. In *Proceedings of Workshop on Social Media Vis.*, AAAI (Dublin, Ireland, 2012), 6–9.
6. Inglehart, R., and Welzel, C. Changing Mass Priorities: The Link between Modernization and Democracy. *Perspectives on Politics 8*, 02 (2010), 551–567.
7. Jolliffe, I. T. *Principal Component Analysis*, second ed. Springer, 2002.
8. Murdock, G. *Social Structure*. Macmillan, 1949.
9. Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. An Empirical Study of Geographic User Activity Patterns in Foursquare. In *Proceedings of ICWSM'11* (Barcelona, Spain, 2011).
10. Noulas, A., Scellato, S., Mascolo, C., and Pontil, M. Exploiting Semantic Annotations for Clustering Geographic Areas and Users in Location-based Social Networks. In *Proceedings of ICWSM'11*, AAAI (Barcelona, Spain, 2011).
11. Sales, A., Alves, L., Araújo, M., Menezes, A., Morais, A., and Andrade, N. O uso de uma rede geossocial nas cidades brasileiras e sua relação com fatores socioeconômicos. In *Proceedings of IHC'13* (Manaus, Brasil, 2013), 142–147.
12. Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. Socio-spatial Properties of Online Location-based Social Networks. In *Proceedings of ICWSM'11* (Barcelona, Spain, 2011).
13. Shankar, P., Huang, Y.-W., Castro, P., Nath, B., and Iftode, L. Crowds replace experts: Building better location-based services using mobile social network interactions. In *Int. Conf. on Perv. Comp. and Comm. (Percom'12)* (Lugano, Switzerland, 2012), 20–29.
14. Silva, T., Vaz De Melo, P., Almeida, J., and Loureiro, A. Large-scale study of city dynamics and urban social behavior using participatory sensing. *Wireless Communications, IEEE 21*, 1 (Feb 2014), 42–51.
15. Silva, T. H., de Melo, P. O. S. V., Almeida, J. M., and Loureiro, A. A. F. Estudo de hábitos alimentares e de bebida usando mídia social. In *Proceedings of IHC'14*, Sociedade Brasileira de Computação (Porto Alegre, Brazil, Brazil, 2014), 337–340.
16. Silva, T. H., Vaz de Melo, P. O. S., Almeida, J. M., Musolesi, M., and Loureiro, A. A. F. You are What you Eat (and Drink): Identifying Cultural Boundaries by Analyzing Food & Drink Habits in Foursquare. In *Proceedings of ICWSM'14* (Ann Arbor, MI, USA, 2014).