Estudo de Preferências por Receitas do AllRecipes pelo Mundo

Juliana Viscenheski
Univ. Tecnológica Federal do Paraná
Curitiba, PR, Brasil
jviscenheski@alunos.utfpr.edu.br

Artur Ziviani (in memoriam)
LNCC
Petrópolis, RJ, Brasil
ziviani@lncc.br

Thiago H Silva
Univ. Tecnológica Federal do Paraná
Curitiba, PR, Brasil
thiagoh@utfpr.edu.br

ABSTRACT

A considerable part of the culture and behavior of societies are derived from the habits and preferences built up over time. One representative characteristic that a group can present is the preference for certain food groups, thus, building the gastronomic identity of each region around the world. With the increasingly broad connections established by social networks, it is now more feasible to analyze such preferences on a large scale. This study examines recipes from Allrecipes.com network in three continents: America, Europe, and Asia. Based on the evaluations made by the users, a score was developed, allowing the separation of the recipes in two broad groups: well evaluated and poorly evaluated. All the ingredients of these recipes were extracted and used to assemble a network whose links were made via pointwise mutual information. This measure of association, used in pairs of ingredients, allowed us to find the main ingredients common to the countries. Our study may help to better understand the success, or otherwise, of a recipe, in a specific locality, based on its main ingredients. Thus, one of the main utilities envisioned for this work is to establish better recommendations for recipes.

KEYWORDS

Food, Recipes, Complex Networks, Allrecipes

1 INTRODUÇÃO

A preferência por certos tipos de alimentos em detrimento de outros é uma característica muito peculiar de cada pessoa. Entretanto, pode ser válido tentar relacionar os gostos particulares com o gosto do senso comum através de aspectos de origem histórica, tradicional ou até mesmo gênero. Um exemplo desta generalização está nas chamadas receitas típicas de cada região. Tais ingredientes foram selecionados como típicos por constituirem uma receita que além de ser criada na região, foi aceita pelo paladar da maioria de seus moradores. Com a evolução dos meios de comunicação, os simples diários de receitas de família se tornaram redes complexas que conectam pessoas pelos seus gostos em comum e as permitem compartilhar não só receitas, mas também experiências e opiniões. Um exemplo disso são as redes sociais pensadas exclusivamente nas pessoas que enxergam na cozinha um universo de descobertas, tais como Allrecipes, Pip! (criada em Florianópolis, Brasil) e All Chefs. Todas essas plataformas contribuem para o crescimento do interesse pela culinária, assim como podem revelar comportamentos alimentares de determinadas regiões baseado nas ações de seus usuários.

O objetivo deste trabalho é estudar os fatores que fazem uma receita obter sucesso em uma determinada região considerando seus ingredientes e a interação dos usuários. Para isso, foram coletadas receitas da rede Allrecipes do Brasil, França, Alemanha, Itália, Índia e Estados Unidos. Com os dados obtidos, o presente trabalho conseguiu construir um rede entre os ingredientes por meio das preferências dos usuários dos países estudados. Tais conclusões podem evoluir na direção de recomendação de combinações de ingredientes que podem ser melhor aceitos em uma região específica, por exemplo. Além disso, futuramente pode-se levar em conta outras características dos ingredientes (valores nutricionais, por exemplo), para dar maior acuracidade às recomendações.

A organização deste está constituída da seguinte forma: A Seção 2 apresenta alguns dos principais trabalhos relacionados. A Seção 3 explica detalhes da fonte dos dados utilizada e a importância dela no compartilhamento de experiências do ramo culinário entre os usuários. Além disso, nesta seção ainda são abordados os parâmetros recolhidos para cada receita. A Seção 4 explica a abordagem proposta para identificar receitas que são populares de acordo com a opinião dos usuários. Além disso, nessa mesma seção se evidencia o número de receitas recolhidas e devidadamente tratadas de cada localidade. Os resultados obtidos até o presente momento são expostos e discutidos na Seção 5. A Seção 6 discute possíveis ameaças aos resultados. Por fim, a Seção 7 apresenta as conclusões obtidas, assim como uma descrição das ideias que futuramente serão implementadas a este trabalho.

2 TRABALHOS RELACIONADOS

A tentativa de analisar as preferências alimentares das pessoas por meio de seus comportamentos virtuais não é nova no ramo da computação. Wagner e Aiello [1] buscaram diferenças entre as imagens de comidas (associadas a tags) publicadas por homens e mulheres em uma mesma rede social e, com isso, conseguiram especificar quais as preferências de cada gênero no que diz respeito a consumo de alimentos. Da mesma maneira, Salvador et al. [2] buscaram compreender se o tempo de preparo de uma receita influencia na preferência das pessoas, encontrando uma correlação muito evidente.

O trabalho de Ahn et al. [3] identificou os sabores que são compartilhados por determinados ingredientes. Sabendo os locais onde aqueles ingredientes são mais comuns, conseguiram identificar similaridades entre diversas regiões do mundo.

O trabalho de BLABLA [4] utilizou dados de receitas postados antes e durante a quarentena imposta pela COVID-19 a fim de identificar padrões de consumo que pudessem ter sido alterados em razão do isolamento forçado. Além disso, ao utilizar inteligência artifical para otimizar o tempo de processamento e também a qualidade

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

da extração de informação, os autores expõe uma oportunidade de melhoria para o presente trabalho.

Na linha de pesquisa tendo o Allrecipes como base de dados, Teng et al. [5] estudaram a probabilidade dos ingredientes aparecerem sempre juntos nas receitas, estabelecendo uma interdependência entre eles. Além disso, com um processo parecido com o utilizado no presente trabalho, conseguiram encontrar os ingredientes mais comuns em cada região, assim como especificar qual o tipo das mudanças sugeridas pelos usuários a cada receita.

A diferença do presente trabalho está, primeiramente, na diversificação da base de dados. Isso se deve porque foram estudados os Allrecipes locais de cada país e não as categorias internas de classificação por região existente no Allrecipes.com. Ademais, as relações estabelecidas neste trabalho tem potencial para relacionar aspectos dos mais diversos (tempo de preparo, tempo de cozimento, avaliação dada pelos usuários que a realizaram, número de estrelas, ingredientes) de uma receita e estabelecer sua probabilidade de sucesso em cada um das regiões estudadas. Da mesma forma, o caminho inverso pode ser estabelecido, isto é, ao desejar obter sucesso com uma receita em determinada localidade, este trabalho poderá indicar os seletos ingredientes que são comuns às receitas muito bem avaliadas naquela região.

3 A REDE SOCIAL ALL RECIPES

Criada em 1997, por estudantes universitários da cidade de Seattle, a rede Allrecipes foi escolhida como fonte de dados para esse trabalho. A ideia para a plataforma surgiu quando os estudantes perceberam que automatizar a busca por receitas poderia ser de grande valia. A rede social Allrecipes tem por objetivo o compartilhamento de receitas entre seus usuários, assim como suas experiências e opiniões acerca de cada uma delas. As receitas são avaliadas por número de estrelas, quantas pessoas já a fizeram e quantas pessoas a avaliaram. Além disso, informações como tempo de preparo, tempo de cozimento e ingredientes estão disponíveis na plataforma e foram armazenados durante o processo de coleta de dados. Na figura 1 está um exemplo do formato de uma receita e seus atributos coletados (destacados em azul).

4 METODOLOGIA

Inicialmente foram escolhidos os países cujas receitas seriam coletadas. A fim de obter resultados dispostos em várias localidades, Brasil, França, Alemanha, Itália, Índia e Estados Unidos foram escolhidos. Com um processo de *webscraping* os dados foram coletados, traduzidos para o inglês com o auxílio da Google Translate API e armazenados em um banco de dados. Número de estrelas, número de avaliações, informações relacionadas ao preparo das receitas, além da quantidade de pessoas que já as fizeram foram atributos selecionados para coleta.

Foram coletadas receitas das seguintes categorias: prato principal, salada, sobremesa, café da manhã e aperitivos totalizando quase 39 mil receitas. Após o processo de coleta e tratamento dos dados, 37.489 receitas das 6 localidades distintas foram selecionadas, sendo elas Brasil, França, Alemanha, Itália, Índia e Estados Unidos. A tabela 1 mostra o número de receitas para cada local.

Em outro momento, foi necessário categorizar as receitas e identificar um padrão de avaliação que pudesse qualificar uma receita

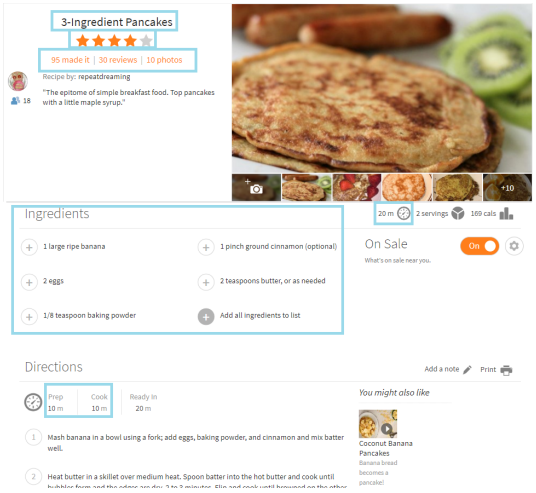

Figure 1: Estrutura da receita

Brasil	França	Alemanha	Itália	Índia	EUA
7795	5569	6985	4006	997	12167

Table 1: Distribuição das receitas por país

perante a todas as outras. Apesar da grande quantidade de atributos de classificação, categorizar uma receita apenas pelas suas avaliações, deixando de fora o número de pessoas que a realizaram, por exemplo, poderia trazer incertezas a tal categorização. A fim de evitar injustiças e atribuir às receitas uma nota final que melhor as qualificassem perante o gosto os usuários, foi elaborada uma fórmula que une todos os atributos de avaliação:

$$Score = log(A+1) + log(P+1) + (S+1)^2$$

sendo:
A : número de avaliações da receita
P : número de pessoas que fizeram a receita
S : número de estrelas que a receita possui

O conhecimento das receitas bem e mal avaliadas permitiu a extração dos ingredientes das mesmas e, por fim, a construção de uma rede $G(V, E)$, onde os vértices $v_i \in V$ representam os ingredientes únicos do dataset. O critério utilizado para criar uma aresta $e(x, y)$, conectando dois ingredientes v_x e v_y, foi o *pointwise mutual information* (PMI), baseado em pares de ingredientes (x, y):

$$PMI(x,y) = \frac{p(x,y)}{(p(x)p(y))},$$

na qual

$$p(x,y) = \frac{\#dereceitascontendoxey}{\#totaldereceitas}$$

$$p(x) = \frac{\#dereceitascontendox}{\#totaldereceitas}$$

$$p(y) = \frac{\#dereceitascontendoy}{\#totaldereceitas}.$$

5 RESULTADOS

Com a atribuição do *score* a cada um das receitas, tornou-se possível analisar o sucesso de uma receita perante as outras. A Figura 2 mostra a distribuição dos *scores* em uma função de distribuição acumulada (CDF - Cumulative Density Function), para todos os países considerados neste trabalho.

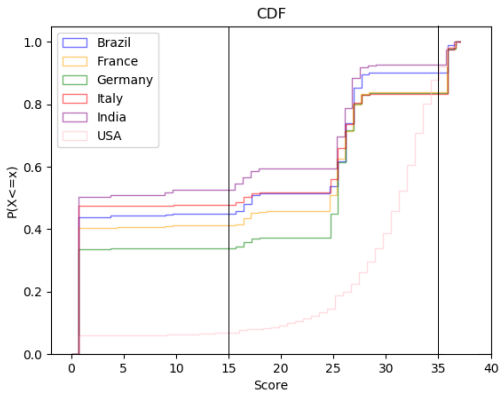

Figure 2: Função de distribuição acumulada do *score* de cada país

Com o auxílio da Figura 2, percebe-se que a curva característica do Brasil, França, Alemanha, Itália e Índia são muito semelhantes, além disso, todas compartilham uma distribuição parecida para o *score*. Isso sugere que o padrão de comportamento na avaliação de receitas nesses países é similar. Existe um grupo distinto de receitas que as pessoas não gostam, bem como existe outro grupo de receitas que as pessoas gostam muito. O limite de *score* que separam esses grupos são bastante parecidos.

É possível observar uma ligeira mudança nas curvas das CDFs por volta do *score* 15 e uma mudança brusca por volta do *score* 25 em todos os países (exceto os EUA), indicando que são receitas mais bem avaliadas. Existe ainda um conjunto de receitas com um *score* muito alto, acima de 37, e para esses países essas receitas representam cerca de 5%. A curva característica dos Estados Unidos, no entanto, não se aproxima de nenhuma das outras curvas. De maneira singular, as receitas dos Estados unidos costumam ter, de maneira geral, avaliações muito positivas.

A fim de analisar a causa desta discrepância significativa entre a CDF dos EUA em comparação com a CDF dos outros países, foi elaborado o gráfico boxplot dos EUA e do Brasil em função do número de estrelas atribuído às receitas (figura 3).O gráfico boxplot evidencia aspectos da distribuição dos dados por meio de quartis, tais como: centro da distribuição (linha da mediana), amplitude dos dados (tamanho do quadrado), simetria ou assimetria do conjunto (hastes) e a presença de outliers (pontos acima ou abaixo das hastes).

Em uma primeira análise, percebe-se que a mediana do número de estrelas dos Estados Unidos está muito acima da mediana das receitas do Brasil (e, por comparação com a CDF, dos outros países estudados). Isso reafirma que a função de densidade acumulada mostrou: as receitas estadunidenses são muito bem avaliadas de maneira geral. Outra indicação consiste na presença de pontos

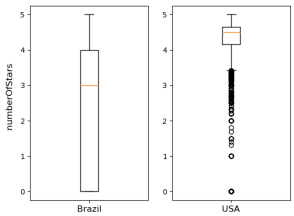

Figure 3: Boxplot do *Score* do Brasil e Estados Unidos

abaixo do limite inferior, mostrando que receitas mal avaliadas são dados ímpares quando comparados a todo o conjunto.

Outra maneira de visualizar as diferenças entre as receitas é por meio de nuvem de palavras. As receitas foram divididas com base no valor das notas de corte do *score* (muito evidente no gráfico das CDFs) e, para cada grupo, foi gerada uma nuvem de palavras com os ingredientes mais comuns. As nuvem de palavrass das receitas mal avaliadas (*score* baixo, isto é, um score de até 15.00), receitas razoáveis, tendo um *score* entre 15.00 e 35.00 e as receitas de sucesso, com *score* maior que 35.00 estão dispostas nas Figuras 4, 5, 7.

Figure 4: Ingredientes das receitas com *score* baixo. Brasil (esquerda), Estados Unidos (centro) e Índia (direita).

Figure 5: Ingredientes das receitas com *score* médio. Brasil (esquerda), Estados Unidos (centro) e Índia (direita).

De maneira geral, as receitas mal avaliadas na Índia e Brasil possuem o sal como ingrediente em maior destaque. No entanto é importante notar que existem diferenças significativas entre os três países. Ovo (do inglês *egg*), por exemplo, só aparece em grande evidência nas receitas dos Estados Unidos.

Na categoria das receitas com score médio, açúcar e sal apareceram em destaque tanto no Brasil como na Índia, sugerindo que receitas com avaliações intermediárias pertencem a uma coleção mista de doces e salgados nessas duas regiões. Outra análise nos permite dizer que pimenta parece ser um ingrediente não muito bem quisto pelos brasileiros. Olhando para as receitas muito bem avaliadas, Índia e Brasil compartilham o açúcar e leite como ingredientes mais recorrentes.

O grafo da figura 6 evidencia os ingredientes compartilhados pelos países. Além disso, é possível notar as singularidades de cada

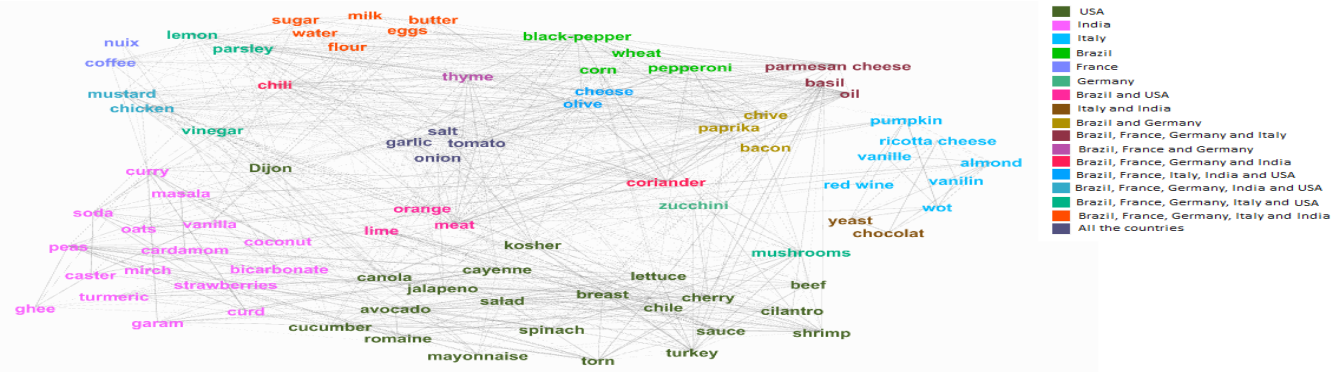

Figure 6: Rede de ingredientes pertencentes a receitas bem avaliadas

Figure 7: Ingredientes das receitas com *score* alto. Brasil (esquerda), Estados Unidos (centro) e Índia (direita).

região, isto é, ingredientes que só são extraídos de receitas bem avaliadas de determinado país.

É possível perceber a diversidade de combinações de países nas quais o Brasil participa, indicando uma forte influência externa na gastronomia local. Possivelmente, tal influência remete ainda aos tempos de colonização, no qual diversas culturas foram introduzidas no território brasileiro por meio da imigração. Este comportamento é diferente do apresentado pela Índia, por exemplo, cujos ingredientes são, em sua grande maioria, de uso único da região e quando são compartilhados, o fazem de uma forma muito sutil. Da mesma forma, os EUA se apresentam como um país cujo compartilhamento de ingredientes com outras localidades é pequeno, mas o número de ingredientes de uso quase exclusivo é bastante evidente. Essa característica pode revelar que a gastronomia desses países é regida por diferentes gostos e peculiaridades, quando comparados aos outros países estudados.

6 POSSÍVEIS AMEAÇAS PARA OS RESULTADOS

Ao utilizar a rede social Allrecipes, o usuário tem o poder de fazer qualquer comentário acerca das receitas disponíveis. Após analisar padrões presentes na relação entre os diversos atributos das receitas, conjecturou-se para esse trabalho que o número de avaliações seria, majoritariamente, algo positivo para a receita em questão. Isto é, as avaliações seriam oriundas de usuários que realizaram a receita e obtiveram um retorno positivo da mesma. Entretanto, as avaliações podem conter reclamações, críticas e sugestão de mudanças. Devido ao fato do *score* não ser baseado somente no número de avaliações, mas sim em um compilado de vários atributos da receita, esta parcela de avaliações negativas foi desconsiderada neste trabalho.

Outro ponto que merece atenção é quanto à representatividade dos dados, isso porque os comportamentos e preferências observados por este trabalho podem não ser refletir em sua totalidade as preferências dos habitantes de cada país, mas somente daqueles que utilizam a plataforma estudada.

7 CONCLUSÃO E TRABALHOS FUTUROS

O presente trabalho conseguiu atingir o objetivo principal ao encontrar, na rede de ingredientes, quais são os responsáveis pela grande popularidade de algumas receitas em determinadas regiões.Foi possível visualizar uma considerável semelhança entre as distribuições das preferências alimentares de alguns países estudados. O grafo dos ingredientes possibilita a visualização da quantidade destes que são compartilhados pelos países e poderia ser um ponto de partida para futuros trabalhos.

AGRADECIMENTOS

Este trabalho foi apoiado por CNPq (processo 403260/2016-7), CAPES, FAPERJ e FAPESP.

REFERENCES

[1] C. Wagner and L. M. Aiello, "Men eat on mars, women on venus?: An empirical study of food-images," in *Proceedings of the ACM Web Science Conference*, (Oxford, UK), p. 63, ACM, 2015.

[2] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, "Learning cross-modal embeddings for cooking recipes and food images," *Training*, vol. 720, no. 619-508, p. 2, 2017.

[3] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási, "Flavor network and the principles of food pairing," *Scientific Reports*, vol. 1, pp. 196 EP –, Dec 2011. Article.

[4] T. Eftimov, G. Popovski, M. Petković, B. K. Seljak, and D. Kocev, "Covid-19 pandemic changes the food consumption patterns," *Trends in Food Science & Technology*, vol. 104, pp. 268–272, 2020.

[5] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic, "Recipe recommendation using ingredient networks," in *Proceedings of the 4th Annual ACM Web Science Conference*, WebSci '12, (New York, NY, USA), pp. 298–307, ACM, 2012.