

Comer, Comentar e Compartilhar: Análise de uma Rede de Ingredientes e Receitas

Willyan Michel Ferreira
Universidade Federal de Ouro
Preto
Ouro Preto, Minas Gerais
willyanmichel@gmail.com

Ana Paula Couto da Silva
Universidade Federal de
Minas Gerais
Belo Horizonte, Minas Gerais
ana.coutosilva@dcc.ufmg.br

Fabrcio Benevenuto
Universidade Federal de
Minas Gerais
Belo Horizonte, Minas Gerais
fabricio@dcc.ufmg.br

Luiz H. C. Merschmann
Universidade Federal de Ouro
Preto
Ouro Preto, Minas Gerais
luizhenrique@iceb.ufop.br

RESUMO

A troca de receitas é um hábito de muitas pessoas. Um meio online e colaborativo de compartilhar esse tipo de informação é através de websites especializados que permitem que usuários postem novas receitas, comentem e/ou avaliem receitas existentes. Apesar de extremamente populares, pouco se sabe sobre esses sistemas e os padrões de interações que eles permitem. Visando preencher essa lacuna, esse trabalho apresenta uma extensa caracterização do site Tudo Gostoso, um importante site brasileiro de compartilhamento de receitas. Para isso, foram coletadas todas as receitas existentes no site juntamente com informações associadas aos comentários e avaliações dos usuários. Além de explorar as interações existentes entre os usuários do site, este trabalho analisa uma rede formada por ingredientes que co-ocorrem em receitas e investiga a viabilidade de se extrair possíveis alterações nas receitas a partir de comentários. As análises apresentadas revelam padrões de uso de ingredientes importantes da culinária brasileira e podem ser úteis para inspirar a construção de diversas novas aplicações nesse contexto.

ABSTRACT

Exchanging recipes is an habit of many people. An online and collaborative way of sharing this type of information is through specialized websites that allow users to post new recipes, comment and/or rate existing recipes. Although extremely popular, little is known about these systems and about the patterns of interactions that they promote. Aiming at filling this gap, this paper presents an extensive characterization of the site Tudo Gostoso, an important Brazilian site for sharing recipes. To do that, we crawled all recipes on the site along with information associated to com-

ments and evaluations. In addition to explore the interactions between users of the website, our work analyze a network of co-occurring ingredients in recipes and investigates the feasibility of extracting existing changes on recipes from commentaries. The analysis of this graph reveals patterns of ingredients to use basic ingredients of Brazilian cuisine.

Categories and Subject Descriptors

J.4 [Computer Applications]: Social and Behavioral Sciences

General Terms

Human Factors, Measurement

Keywords

Recipes, Ingredient networks

1. INTRODUÇÃO

Desde seu início, a Web tem sido um dos principais meios para o compartilhamento colaborativo de receitas culinárias. Sites de receitas colaborativos fazem sucesso na Web devido à necessidade que muitas pessoas têm de compartilhar suas receitas, que variam desde antigas receitas de famílias passadas por várias gerações até recentes invenções. Uma vez compartilhadas, receitas são avaliadas por outros usuários, que podem dar notas ou mesmo postar comentários sobre as mesmas.

Em um passado recente, receitas eram armazenadas somente em livros ou cadernos pessoais, construídos ao longo de anos. Sistemas colaborativos para compartilhamento de receitas trazem inúmeras diferenças tais como: 1) Sites de receitas são gratuitos e podem ser acessados livremente na Web; 2) Usuários dão notas às receitas, o que pode indicar as receitas mais promissoras; 3) Usuários comentam receitas, sugerindo modificações e aprimoramentos. Tais mecanismos fizeram com que sites de receitas se tornassem grandes repositórios de receitas com uma quantidade enorme de acessos. Como exemplo, enquanto em livros de culinária básica encontramos apenas uma receita de pão de queijo, no site Tudo Gostoso, objeto de estudo deste trabalho, encontramos mais de

100 receitas de pão de queijo que vão muito além do pão de queijo tradicional mineiro, tais como: pão de queijo recheado com carne (ex. calabresa, lombo defumado, etc.), pão de queijo de liquidificador, de forma, de sanduicheira, ou mesmo o pão de queijo da Dona Alilás.

A quantidade de informação disponível nesse tipo de sistema colaborativo nos permite não só aprender a cozinhar, mas também estudar questões relacionadas às preferências dos usuários ou os padrões típicos de receitas brasileiras. Como exemplo, através desses sistemas, podemos identificar combinações importantes de ingredientes. Além disso, comentários postados em receitas possuem informações que podem ser exploradas não só para inferir a qualidade das receitas, mas também para sugerir modificações e aprimoramentos. Apesar de oferecerem tantas possibilidades interessantes para investigação, pouco se sabe sobre esses sistemas, suas receitas e os padrões das interações sociais neles existentes.

Visando preencher essa lacuna, este trabalho apresenta uma ampla caracterização do site Tudo Gostoso, um importante portal brasileiro de receitas culinárias. Para isso, nós coletamos todas as receitas existentes no site juntamente com informações associadas aos comentários, avaliações recebidas, bem como os ingredientes existentes nas receitas. Além de explorar as interações existentes entre os usuários do site, nosso trabalho analisa uma rede formada por ingredientes que co-ocorrem em receitas. A análise desse grafo de ingredientes revela padrões de uso de ingredientes importantes da culinária brasileira. Finalmente, nós investigamos a viabilidade de se extrair possíveis alterações nas receitas a partir de comentários. Nós acreditamos que nossas análises e descobertas possam ser úteis para a construção de novas aplicações nesse contexto, além de identificar padrões estruturais do uso de ingredientes em receitas que estão intrinsecamente ligados a aspectos culturais da culinária brasileira.

O restante do artigo está organizado da seguinte maneira. A seção a seguir discute brevemente os trabalhos relacionados. Na Seção 3, descrevemos a base de dados coletada, bem como nossa estratégia de coleta e tratamento dos dados. A Seção 4 descreve aspectos do comportamento dos usuários que interagem no sistema. Na Seção 5, analisamos as receitas postadas no site e apresentamos as principais relações observadas entre os ingredientes a partir de um grafo gerado visando facilitar a visualização e entendimento da rede. A Seção 6 investiga a viabilidade de se extrair possíveis alterações nas receitas a partir dos comentários encontrados no site. Por fim, na Seção 7, apresentamos conclusões e direções de trabalhos futuros.

2. TRABALHOS RELACIONADOS

Em um artigo recente, Ahn, Ahnert, Bagrow e Barabási [1] criaram uma rede que conecta ingredientes de acordo com os sabores de cada alimento. Esses foram extraídos a partir dos elementos químicos predominantes em cada ingrediente. A partir de receitas extraídas de diferentes tipos de cozinha (ex. asiática, indiana, italiana) e da análise desse grafo, eles mostram diferentes padrões culturais no uso dos alimentos. Mais recentemente, Teng, Lin e Adamic [6] propõem observar a co-ocorrência de ingredientes e investigam os elementos

que impactam nas avaliações recebidas pelas receitas. Como exemplo, eles mostram que receitas que envolvem processos mais elaborados, como defumar ou fermentar, tendem a receber notas melhores. Baseado nessas e outras características, os autores utilizam uma abordagem supervisionada de classificação para prever as avaliações de receitas. Nosso trabalho aborda uma rede diferente da rede de sabores e não explora o problema da predição de avaliações, sendo, portanto, complementar a esses esforços.

Trabalhos envolvendo receitas têm sido alvo de diversos estudos que visam a recomendação de receitas baseando-se em avaliações prévias de seus usuários, seja por meio de voto [2] ou mesmo respondendo questionários [5]. Existem trabalhos que se baseiam em recomendar receitas onde o ingrediente central seja o mesmo da receita acessada pelo usuário. Por exemplo, se a receita acessada foi “batata recheada”, onde o ingrediente principal é a batata, seriam recomendadas receitas que possuem batata como ingrediente principal [8]. Outra forma de recomendação utilizada foi observar o histórico de navegação dos usuários, buscando determinar relações entre receitas a partir de como usuário navega pelo site [7]. Uma abordagem semelhante foi proposta por Geleijnse e seus colaboradores [3]. Eles desenvolveram um sistema de recomendação baseado nos ingredientes de receitas consultadas pelo usuário no passado de forma a explorar esse histórico de receitas para recomendar outras receitas que o usuário possa gostar. Diferentemente dos esforços acima, nosso trabalho não aborda questões relacionadas a recomendação, mas sim uma caracterização de padrões de interação e uso de ingredientes nesses sistemas. Acreditamos que resultados apresentados em nosso estudo possam formar a base de sistemas de recomendação voltados para o contexto específico das receitas brasileiras.

3. BASE DE DADOS

Esta seção descreve brevemente o site de receitas utilizado, nossa estratégia de coleta e os aspectos do tratamento e extração de ingredientes das receitas coletadas.

3.1 Site de Receitas

O site *Tudo Gostoso*¹ é um portal de compartilhamento de receitas que foi criado em 2005. Ele permite aos usuários acessar, postar, avaliar, fazer comentários e responder a comentários das receitas. Nesse site, as receitas são organizadas em diferentes categorias. A Tabela 1 mostra as diferentes categorias, bem como o total de receitas em cada uma delas. A base de dados coletada é composta por todas as receitas públicas, resultando na coleta de todos os ingredientes que as compõem, bem como os comentários, respostas e avaliações dos usuários do site.

3.2 Metodologia de Coleta

Para obtenção da base de dados de receitas, foi desenvolvido um *crawler* na linguagem Python². O processo de coleta se dividiu em duas etapas. Para cada uma das categorias listadas na Tabela 1, coletamos todas as receitas públicas existentes na categoria. Em seguida, para cada receita, coletamos as informações sobre as avaliações e comentários recebidos. Sendo assim, nossos dados resultam de

¹tudogostoso.uol.com.br/

²www.python.org/

Categoria	Número de Receitas
<i>Bolos e Tortas</i>	26.948
<i>Carnes</i>	6.839
<i>Aves</i>	6.216
<i>Peixes e Frutos do Mar</i>	5.236
<i>Saladas, Molhos e Acompanhamentos</i>	19.315
<i>Sopas</i>	2.796
<i>Massas</i>	16.777
<i>Bebidas</i>	4.557
<i>Doces e Sobremesas</i>	32.217
<i>Lanches</i>	17.667
<i>Prato Único</i>	7.057
<i>Total</i>	145.625

Tabela 1: Categorias e número total de receitas.

uma coleta **exaustiva** do site, que corresponde a um total de **145.626** receitas que receberam **383.189** comentários. A quantidade de receitas coletadas é bastante significativa, visto que outros trabalhos da literatura realizaram seus experimentos com um número muito menor de receitas. Por exemplo, Teng, Lin e Adamic [6] analisaram 46.337 receitas e Ahn, Ahnert, Bagrow e Barabási [1], 56.498. No total, foram identificados **236.189** usuários diferentes que postaram comentários e/ou criaram receitas. A partir desses dados foram extraídas informações como o nome da receita, nome do usuário que criou a receita, número de avaliações recebidas e avaliação média recebida, número de comentários e de usuários que postaram os comentários, os ingredientes de cada receita, etc.

A seguir, discutimos as estratégias utilizadas para extrair os ingredientes presentes nas receitas.

3.3 Extração de Ingredientes

Para gerar a base de ingredientes, os arquivos coletados contendo informações sobre as receitas passaram por um processo de identificação de ingredientes. Nesse processo foram retiradas as quantidades e medidas que acompanham o ingrediente. Por exemplo, do texto “1 colher de chá de sal”, extraímos apenas “sal”. A identificação de ingredientes se deu a partir da análise manual de milhares de receitas. Dessa forma, vários ingredientes foram acrescentados manualmente em nossa lista de ingredientes.

Outro grande problema num país continental e multicultural como o Brasil é a diferenciação regional para os nomes de alguns ingredientes. Um exemplo bem conhecido é o caso da “mandioca”, “macacheira” e “aipim”, que apesar de serem nomes diferentes, correspondem a um mesmo ingrediente. Para tratarmos esse problema, três voluntários percorreram manualmente todos os ingredientes identificados realizando uma padronização de nomes desses ingredientes. Nesse processo, alguns ingredientes foram agrupados e passaram a ter um único nome. Por exemplo, todos os ingredientes citados na base inicial como “canela em pau”, “canela de pau”, “canela em casca”, “canela em rama” e “pau de canela” passaram a ser denominados “canela em pau”. Além disso, houve uma padronização para que todos os nomes de ingredientes ficassem escritos na língua portuguesa (por exemplo, “whisky” foi reescrito como “uísque”) e também uma identificação e correção de nomes de ingredientes que apresentavam erros de grafia. Ao todo foram identificados 1.079.835 ocorrências de ingredientes nas receitas, sendo um total de **1.397** ingredientes diferentes.

4. CARACTERIZAÇÃO DOS USUÁRIOS

Nesta seção apresentamos uma caracterização dos usuários (participantes) do site. O objetivo principal é o entendimento dos diferentes tipos de usuários e como eles utilizam e interagem no site. Por exemplo, tal entendimento pode ser útil para identificar usuários que contribuem com muitas receitas e comentários interessantes e que podem receber algum tipo de privilégio ou premiação do próprio site. Além disso, após a identificação dos usuários mais influentes, empresas que produzem os ingredientes de diferentes receitas podem incentivar esses usuários a postarem receitas que utilizam seus produtos, como uma forma de propaganda e difusão dos mesmos.

Os dados coletados representam 236.189 usuários, com um usuário especial chamado *Tudo Gostoso*. Esse usuário, como esperado, é de gerência do próprio site e é o maior criador de receitas, contribuindo com aproximadamente 2,2% de todas as receitas. Do total de usuários, aproximadamente 21% (49.861) criaram pelo menos uma receita. A maior parte dos usuários coletados comentam receitas ou simplesmente consultam o site.

A dinâmica do site é composta por duas atividades principais dos usuários: criação das receitas e postagem de comentários. Nos comentários os usuários sugerem acréscimo, exclusão ou substituição de ingredientes; indicam o melhor clima (frio, calor) para degustar a receita, além de sugerir as festividades nas quais as receitas se enquadram melhor (carnaval, festa junina, natal). Obviamente, esses comentários dependem da experiência e preferências de um usuário em particular. No entanto, o comportamento médio dos usuários pode auxiliar na melhoria da qualidade do site. Por fim, comentários também podem conter dúvidas, elogios e críticas.

O total de 384.189 comentários feitos pelos usuários estão organizados da seguinte forma. Aproximadamente 48,4% dos usuários postaram ao menos 1 comentário, em qualquer uma das categorias listadas na Tabela 1. O usuário mais ativo, em termos de comentários, foi responsável por 533 comentários, ou seja 0,018%. A categoria que recebeu maior número de usuários distintos que colaboram com um comentário foi a de Bolos e Tortas, com 16,7% dos usuários. Interessante é que a categoria com maior atividade dos usuários considerando criação e comentários também é a categoria de Bolos e Tortas, com 21,7% de usuários ativos nesta categoria.

A cada comentário gerado, uma resposta pode ser dada. Por exemplo, o caso mais natural para comentário com resposta associada é o caso de dúvidas. No entanto, poucos usuários respondem a comentários (somente 8,12% do total). O usuário que responde a mais perguntas é o *Tudo Gostoso*, com 30,3% do total de respostas (5.802 respostas). Esse valor elevado é esperado dado que este usuário é o administrador do site, e, de uma certa forma, deve interagir mais com os demais usuários do site. Novamente, a categoria de Bolos e Tortas se destaca, com o maior número de respostas (4.890).

Uma questão interessante que foi investigada é se existe alguma correlação entre usuários que mais criam receitas e aqueles que participam mais ativamente com comentários.

Categoria	Homens (%)	Mulheres (%)	Não Identificado (%)
<i>Bolos e Tortas</i>	3.000 (5,9)	47.768(93,3)	403 (0,8)
<i>Carnes</i>	1.281 (9,3)	12.370 (89,9)	115 (0,8)
<i>Aves</i>	818 (7,3)	10.226 (91,5)	93 (0,8)
<i>Peixes e Frutos do Mar</i>	954 (8,9)	9.693 (90,4)	72 (0,7)
<i>Saladas, Molhos e Acom.</i>	1.922 (7,3)	24.022 (91,5)	313 (1,2)
<i>Sopas</i>	381 (7,0)	5.004 (92,4)	32 (0,6)
<i>Massas</i>	2.184 (7,2)	28.038 (92,0)	241 (0,8)
<i>Bebidas</i>	746 (10,5)	6.311 (88,5)	78 (1,1)
<i>Doces e Sobremesas</i>	2.953 (6,3)	43.535 (92,8)	400 (0,9)
<i>Lanches</i>	2.161 (6,4)	31.296 (92,9)	245 (0,7)
<i>Prato Único</i>	1.014 (6,7)	13.944 (92,5)	114 (0,8)
<i>Total</i>	16.446 (7,0)	217.878 (92,2)	1.865 (0,8)

Tabela 2: Percentagem de comentários feito por homens e mulheres.

Para tal, foi calculado o coeficiente de correlação de Pearson. O valor para essa métrica foi igual a 0,87, o que significa que usuários que participam ativamente na criação de receitas são também aqueles com o maior número de comentários postados.

Investigamos também se usuários simplesmente postam receitas ou se posteriormente realizam comentários e/ou respondem as dúvidas relacionadas com suas receitas. Observouse que 3,86% do total de usuários criadores de receitas (6.499 usuários) comentaram suas próprias receitas, sendo responsáveis por um de total 7.366 dos comentários. Em relação às respostas, 12,21% do total de usuários que criaram receitas (1.033 usuários) responderam algum comentário feito em suas receitas.

Finalizando, a Tabela 2 mostra como os comentários estão divididos entre homens e mulheres, além de comentários onde o gênero não foi identificado. Como podemos observar, comentários são realizados predominantemente por mulheres. Além disso, se o número de comentários for utilizado como uma métrica de preferência, mulheres concentram sugestões em categorias relacionadas a doces (Bolos, Tortas e Sobremesas).

5. CARACTERIZAÇÃO DAS RECEITAS E INGREDIENTES

Nesta seção é realizada a análise das informações disponibilizadas no site considerando duas granularidades: receitas e ingredientes.

5.1 Receitas

O site *Tudo Gostoso* é composto por 145.625 receitas com 1.397 ingredientes distintos.

A categoria com o maior número de receitas é a de Doces e Sobremesas, com 22,1% das receitas (32.217 receitas); e a categoria com o menor número de receitas é a de Sopas, com 1,92% do total de receitas, ou seja, 2.796. Provavelmente a categoria de Sopas tem o menor número de receitas pelo fato de morarmos em um país tropical, onde os invernos geralmente são amenos.

Receitas com o maior número de ingredientes distintos, 645, fazem parte da categoria de Saladas, Molhos e Acompanhamentos. O menor número de ingredientes distintos é encontrado na categoria de Sopas, com o total de 409 ingredientes

distintos. Adicionalmente, um subconjunto com o total de 74 ingredientes ocorre somente na categoria de Bebidas, conforme mostrado na Tabela 3.

Categoria	Ingredientes Distintos
<i>Bolos e Tortas</i>	603
<i>Carnes</i>	537
<i>Aves</i>	425
<i>Peixes e Frutos do Mar</i>	489
<i>Saladas, Molhos e Acompanhamentos</i>	645
<i>Sopas</i>	408
<i>Massas</i>	596
<i>Bebidas</i>	442
<i>Doces e Sobremesas</i>	570
<i>Lanches</i>	583
<i>Prato Único</i>	579

Tabela 3: Ingredientes distintos por categoria.

A Figura 1 mostra a distribuição de probabilidades do total de ingredientes presentes em cada uma das receitas do site. A distribuição segue uma forma semelhante a de uma distribuição normal com número médio de ingredientes igual a 7,4. A receita com maior número de ingredientes é a de um cozido, com o total de 33 ingredientes³.

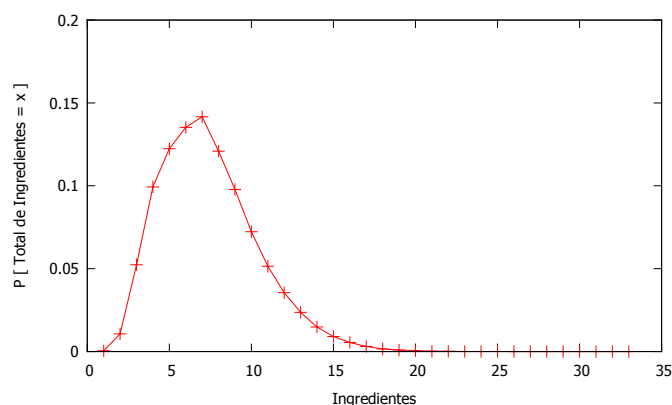


Figura 1: Total de ingredientes por receita.

O site permite aos usuários avaliarem as receitas, atribuindo conceitos entre 0 e 5. Somente 31% das receitas receberam esta avaliação, sendo o conceito médio das avaliações aproximadamente igual a 4,4. A categoria com mais avaliações

³<http://tudogostoso.uol.com.br/receita/77761-cozido.html>

é a de Bolos e Tortas, com 78.725 avaliações (24,56%). A categoria com menor número de avaliações é a de Sopas, com apenas 5.056 avaliações (1,58%). A distribuição do total de avaliações para cada uma das receitas segue a lei de potência, onde poucas receitas recebem um número maior de comentários.

5.2 Ingredientes

Para a análise dos ingredientes utilizados nas receitas, uma rede será utilizada. Esta rede será representada por um grafo $\mathcal{G}(\mathcal{V}, \mathcal{E})$, com \mathcal{V} o conjunto de vértices e \mathcal{E} o conjunto de arestas.

Seja \mathcal{G}_i o grafo não direcionado que modela a rede de ingredientes. Nesta rede, os vértices representam os ingredientes encontrados nas receitas analisadas. Sem perda de generalidade, uma aresta $v \rightarrow u$ ocorre se, e somente se, os ingredientes v e u são utilizados em uma receita r_k . Assim, todos os ingredientes utilizados em uma receita r_k , estarão ligados entre si. Este modelo, pode revelar, por exemplo, quais são os ingredientes mais típicos ou aqueles que combinam mais com outros ingredientes. Além disso, se a rede de ingredientes estiver atrelada a classificação de comunidades que representam regiões diferentes do Brasil, é possível verificar os ingredientes preferenciais para cada região.

Propriedades

Para caracterização e análise das redes de ingredientes, diferentes métricas de teoria de grafos são utilizadas, especialmente métricas topológicas. Métricas topológicas são definidas como medidas baseadas em atributos estruturais de um grafo. Estas métricas podem considerar cada vértice ou a visão global do grafo [4].

(1) **Grau:** O grau de entrada de um vértice v , $k_e(v)$, é o total de arestas que incidem no vértice. Da mesma maneira, o grau de saída de um vértice v , $k_s(v)$, é o total de arestas que tem origem no vértice. Graus de entrada e saída estão relacionados com grafos direcionados. O grau de um vértice v é a soma de $k_e(v)$ e $k_s(v)$. O grau médio, $\langle k \rangle$, do grafo \mathcal{G} é dado por:

$$\langle k \rangle (\mathcal{G}) = \sum_{v \in \mathcal{V}} k(v) / |\mathcal{V}|.$$

Para um grafo não-direcionado, $k_e(v) = k_s(v)$.

A distribuição de grau define a probabilidade de um vértice v ter o valor de grau igual a y , para $y \in \{0, 1, \dots, |\mathcal{V}| - 1\}$.

(2) **Betweenness:** O *Betweenness* de um vértice v é a fração dos caminhos mínimos, calculados usando *breadth-first search*, que ligam qualquer par de vértices e que passam pelo vértice v . Em outras palavras, seja $\sigma_{u,j}$ o total de caminhos mínimos entre u e j , e $\sigma_{u,j}(v)$ o número total de caminhos que passam por v . A métrica de *Betweenness* é definida como:

$$\text{Betweenness}(v) = \sum_{\forall v \neq u \neq j \in \mathcal{V}} \frac{\sigma_{u,j}(v)}{\sigma_{u,j}}.$$

(3) **Diâmetro:** Seja $l(u, v)$ o menor caminho entre u e v . O diâmetro, \mathbf{d} , é definido como o maior caminho entre qualquer

dois vértices do grafo:

$$\mathbf{d}(\mathcal{G}) = \max_{v, u \in \mathcal{V}} l(u, v).$$

Esta propriedade reflete o grau de dispersão dos vértices do grafo \mathcal{G} .

(4) **Coefficiente de Clusterização:** O coeficiente de clusterização (*clustering coefficient*) de um vértice v , ou seja, $cc(v)$ é definido como sendo a divisão da quantidade de arestas existentes entre os vizinhos de v e o número máximo de arestas possíveis entre os vizinhos do vértice v .

O cálculo do coeficiente de clusterização de uma rede ou grafo é calculado a partir da média dos cc de todos os nós.

(5) **PageRank:** O PageRank é um algoritmo iterativo que visa definir um peso numérico para cada vértice do grafo com o propósito de estimar sua importância dentro do mesmo. A equação que calcula o PageRank (PR) de um vértice i , $PR(i)$, pode ser definida da seguinte forma:

$$PRs(i) = (1 - d) + d \sum_{v \in S(i)} \frac{PR(v)}{N_v}.$$

sendo $S(i)$ o conjunto de vértices que apontam para i , N_v , denomina o número de arestas que saem do vértice v , e d é o *damping factor* que varia entre 0 e 1.

Análise

Esta seção foca na análise da rede dos ingredientes que compõem as receitas encontradas no site *Tudo Gostoso*. Analisar a rede de ingredientes pode responder perguntas interessantes, como quais os ingredientes que são mais comuns nas receitas, quais combinam melhor entre si, se existe alguma relação entre ingredientes e questões culturais, entre outras.

A rede de ingredientes, representada pelo grafo \mathcal{G}_i , é formada por 1.397 vértices e 66.487 arestas. O diâmetro da rede, \mathbf{d} , é igual a 3, com o menor caminho médio de aproximadamente igual a 2,1. Estes valores indicam que a rede é extremamente densa, sugerindo que, apesar do número alto de receitas no site, os ingredientes não variam muito de receita para receita.

Em relação a métrica de grau, a rede possui grau médio de $\langle k \rangle = 95,54$. Mapeando em afinidade, pode-se dizer que cada ingrediente, em média, provavelmente harmoniza com outros ≈ 96 ingredientes. Podemos ainda notar que 80% dos ingredientes se ligam, a no máximo, outros 150 ingredientes.

O coeficiente de clusterização da rede de ingredientes é igual a 0,867, podendo significar alta densidade da rede. Além disso, seguindo a definição de clusterização, este valor implica a existência de grupos de três ingredientes que combinam entre si. Este fato é reforçado com o valor do coeficiente de clusterização obtido quando uma rede aleatória, com o mesmo grau médio, é gerada. Para este caso, o valor é igual a 0,086, ou seja, uma ordem de grandeza menor. Assim, a formação da rede de ingredientes segue um processo mais elaborado, provavelmente ligado a combinação química entre os ingredientes, bem como aspectos culturais [1].

Métricas de centralidade, como grau, *betweenness* e *page-rank*, são amplamente utilizadas para detectar os vértices de maior importância na rede analisada. A Tabela 4 apresenta o *ranking* com os *top-10* ingredientes encontrados nas receitas. Conforme esperado, os ingredientes com os maiores valores de centralidade são os ingredientes básicos, como sal, leite, água, ovo, que estão presentes tanto em receitas salgadas como em receitas doces.

Ranking	Grau	Betweenness	PageRank
1	Sal (1069)	Sal (82.749)	Sal (0,010)
2	Leite (956)	Leite (72.075)	Leite (0,009)
3	Cebola (872)	Açúcar (50.047)	Açúcar (0,007)
4	Água (868)	Água (40.895)	Cebola (0,007)
5	Açúcar (844)	Cebola (38.197)	Água (0,007)
6	Ovo (844)	Ovo (30.997)	Alho (0,007)
7	Alho (832)	Alho (30.927)	Ovo (0,007)
8	Far. Trigo (788)	Azeite (23.965)	Far. Trigo (0,006)
9	Óleo (778)	Far. Trigo (23.033)	Azeite (0,006)
10	Azeite (766)	Tomate (22.826)	Tomate (0,006)

Tabela 4: Ranking de ingredientes, com os respectivos valores para cada métrica considerada.

Todas as medidas mostradas até o momento fornecem características dos nós, individualmente, ou da rede como um todo. Uma análise complementar a ser feita está relacionada ao peso das arestas entre os ingredientes. Por exemplo, é interessante saber quais são os pares de ingredientes que mais ocorrem na rede de ingredientes. A Tabela 5 mostra os *top-10* pares de ingredientes que mais ocorrem nas receitas analisadas. Como era de se esperar, pares que contêm açúcar, ovo, sal e água são os que possuem maior ocorrência, dado que receitas de todas as categorias possuem estes ingredientes.

Ranking	Aresta	Peso
1	Ovo ↔ Leite	31.150
2	Açúcar ↔ Leite	29.287
3	Açúcar ↔ Ovo	28.980
4	Sal ↔ Leite	25.725
5	Far. Trigo ↔ Ovo	24.043
6	Sal ↔ Cebola	23.594
7	Leite Condensado ↔ Leite	22.986
8	Fermento ↔ Ovo	22.870
9	Far. Trigo ↔ Leite	22.845
10	Sal ↔ Ovo	22.561

Tabela 5: Ranking das arestas.

A Figura 2 mostra o grafo da rede de ingredientes, onde as arestas entre dois ingredientes estão ponderadas no número de vezes que ambos os ingredientes ocorrem nas receitas. Como é de se esperar, os ingredientes com maiores valores de centralidade (Tabela 4) aparecem com maior destaque no grafo.

6. PERSONALIZANDO RECEITAS

Existem em nossa base de dados receitas muito similares que se diferenciam apenas por um ingrediente ou outro. Essas pequenas diferenças podem estar relacionadas com as preferências de quem as executa. Desse modo, substituições de ingredientes são realizadas na receita original acreditando-se que o sabor da mesma não será alterado. Além das substituições, inclusões e exclusões de ingredientes também acontecem visando deixar a receita com sabor mais agradável.

Essas preferências pessoais levam os usuários a postarem as modificações como uma nova receita ou a utilizarem o formu-

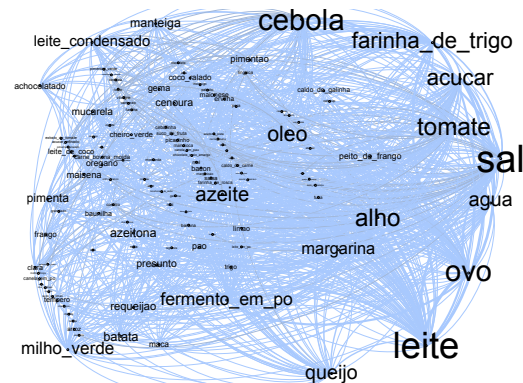


Figura 2: Representação gráfica da rede de ingredientes.

lário de comentários para compartilhar suas sugestões de alterações.

Buscando observar quais tipos de sugestões de modificações os usuários postaram, analisamos os comentários encontrados nas 319.129 receitas do site Tudo Gostoso. Para isso, buscamos nos comentários termos que caracterizam as sugestões de adição, remoção ou substituição de ingredientes.

Para definição desses termos, analisamos uma amostra formada por 1.000 comentários que foram extraídos de 100 receitas escolhidas aleatoriamente na base de dados. Todas essas 100 receitas escolhidas na base de dados possuíam 10 ou mais comentários. Portanto, para formarmos uma amostra contendo 1.000 comentários, coletamos os 10 primeiros de cada uma das 100 receitas selecionadas.

A análise dessa amostra foi realizada de forma manual, ou seja, os comentários foram lidos buscando-se observar quais tipos de comentários são postados em um sistema como o estudado neste trabalho. A partir dessa análise, observamos 11 tipos de comentários diferentes. Vale ressaltar que, em algumas situações, numa única postagem o usuário utiliza mais de um tipo de comentário. A Tabela 6 apresenta a frequência desses tipos de comentários na amostra avaliada.

Tipo de Comentário	Quantidade
Elogio	721
Crítica	92
Sug. Adição	90
Sug. Remoção	29
Sug. Substituição	24
Dúvida	122
Resposta	71
Dica de modo de preparo	42
Modificação de medidas	57
Sug. de data para execução	23
Ainda vai preparar	38

Tabela 6: Tipos de comentários encontrados na amostra.

A partir da análise desses 1.000 comentários extraídos da base de dados, listas de termos-chave que sugerem modificações nas receitas foram elaboradas. A partir dessas listas de termos-chave pudemos automatizar o processo de identificação dos casos de adição, remoção e substituição de ingredientes em todas as receitas da base.

Nos casos de adição e remoção, a seguinte estratégia foi implementada para a identificação dos ingredientes nas frases dos comentários. Optamos por coletar o primeiro ingrediente mencionado após algum termo-chave e, se logo após desse ingrediente for encontrado o conectivo “e”, coletamos também o ingrediente citado após esse conectivo. Desse modo, na frase “acrescentei creme de leite, a receita ficou maravilhosa adoro receitas que utilizam leite condensado!”, considerando que “acrescentei” é um termo-chave, somente o ingrediente creme de leite será corretamente coletado. Já na frase “acrescentei milho verde e cenoura”, essa estratégia nos permite coletar tanto o ingrediente milho verde como a cenoura. As seguintes expressões regulares foram elaboradas a partir da lista de termos-chave para a coleta dos ingredientes relacionados nas sugestões de adição e remoção: “acrescentei|acrescento|acrescente|acrescetar|acrescentamos|acrescentei|adicionei X” e “nao coloquei|nao acrescentei|nem colocar|nao uso|nao coloco|nao pode colocar|n coloquei|tirar|nao coloquei|nao usei|n coloquei X”.

Já no caso de substituições, duas listas de termos-chave foram criadas. A primeira lista foi criada para contemplar as situações em que precisamos coletar os dois ingredientes mencionados logo após um termo-chave. Por exemplo, na frase “troquei leite por creme de leite” precisamos coletar os ingredientes leite e creme de leite, os quais aparecem após o termo-chave “troquei”. A segunda lista contém os termos-chave que possibilitam a coleta dos ingredientes que aparecem em situações como a que ocorre na seguinte frase: “coloquei leite no lugar do creme de leite”. Nesse caso, os ingredientes que precisam ser coletados aparecem intermediados pelo termo-chave “no lugar de”. As seguintes expressões regulares foram elaboradas a partir das listas de termos-chave para a coleta dos ingredientes relacionados nas sugestões de substituição: “no lugar|substitui|em vez|troquei|substituida X por Y” e “Y ao inves| de X”.

A partir das listas de termos-chave construídas, todos os comentários da base de dados de receitas foram processados para obtenção da caracterização das modificações sugeridas nas receitas pelos usuários. No entanto, um pré-processamento no texto dos comentários foi realizado para a separação dos trechos de texto (frases) que foram processados utilizando-se as expressões regulares descritas anteriormente. As pontuações encontradas no texto (vírgula, ponto final, ponto de exclamação, ponto de interrogação e o ponto e vírgula) foram utilizadas para dividir o texto de cada comentário em um conjunto de frases. Após o processamento dessas frases foram encontradas 22.800 sugestões de adição envolvendo 461 ingredientes distintos e 4.885 sugestões de remoção contemplando 311 ingredientes diferentes. A Tabela 7 mostra os 10 ingredientes que apareceram com maior frequência nas sugestões de cada uma das categorias de adição e remoção. Nessa tabela, os valores entre parênteses correspondem às frequências de ocorrência dos ingredientes.

Além das adições e remoções, o processamento dos textos de comentários nos permitiu identificar 5.511 substituições, das quais 2.319 são distintas e que utilizaram 632 ingredientes diferentes. A Tabela 8 apresenta os 10 pares de ingredientes que apareceram com maior frequência nas sugestões de substituição. Nessa tabela, os valores entre parênteses correspondem às frequências de ocorrência dos pares de in-

Ranking	Sugestão de Adição	Sugestão de Remoção
1	Leite (1.733)	Leite (396)
2	Creme de Leite (922)	Creme de Leite (216)
3	Queijo (884)	Açúcar (213)
4	Coco (781)	Água (153)
5	Molho de Tomate (549)	Queijo (137)
6	Açúcar (518)	Sal (129)
7	Far. Trigo (505)	Coco (123)
8	Milho Verde (480)	Cobertura de Sorvete (113)
9	Chocolate (453)	Ovo (111)
10	Cebola (429)	Chocolate (89)

Tabela 7: 10 ingredientes mais frequentes nas sugestões de adição e remoção.

gredientes.

Ranking	Sugestão de Substituição
1	Água/Leite (142)
2	Leite/Leite de Coco (69)
3	Óleo/Azeite (66)
4	Leite/Água (65)
5	Açúcar/Leite Condensado (55)
6	Margarina/Manteiga(52)
7	Óleo/Margarina (46)
8	Açúcar/Adoçante (37)
9	Margarina/Óleo (37)
10	Leite/Creme de Leite (31)

Tabela 8: 10 ingredientes mais frequentes nas sugestões de substituição.

A Figura 4 mostra em que percentual de quantidade de sugestões de adição e remoção aparece cada um dos 20 ingredientes mais frequentes nas receitas da base de dados. Por exemplo, podemos observar nesse gráfico que o ingrediente batata está presente em 1,7% das sugestões de adição e 0,6% das sugestões de remoção. Considerando-se esses 20 ingredientes mais frequentes nas receitas, observa-se que, proporcionalmente, todos eles (exceto a água) aparecem com mais frequência nas sugestões de adição do que nas sugestões de remoção. Isso indica que, se seguirmos as sugestões dos usuários, esses ingredientes tendem a aumentar a quantidade de ingredientes utilizados nas receitas.

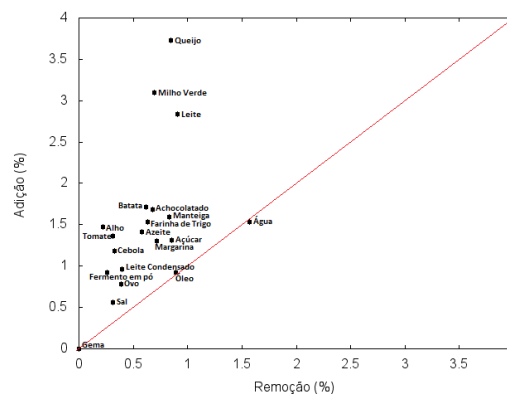


Figura 3: Participação de ingredientes nas sugestões de adição e remoção.

A partir das sugestões de substituições encontradas nos comentários foi gerado o grafo apresentado na Figura 4. Nesse grafo, cada nó representa um ingrediente e cada aresta indica a possibilidade de substituição entre os ingredientes repre-

sentados pelos nós unidos por aquela aresta. Além disso, o tamanho de cada nó está associado com o seu grau, ou seja, os nós maiores indicam que aquele ingrediente está vinculado a uma maior quantidade de outros ingredientes nas sugestões de substituição. Desse modo, é possível observarmos, por exemplo, que o creme de leite é o ingrediente mais flexível com relação às substituições, dado que ele aparece vinculado a um grande número de outros ingredientes.

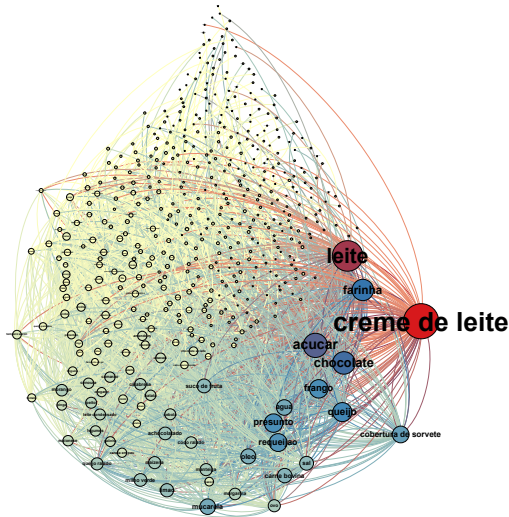


Figura 4: Rede de substituições.

Essa caracterização dos comentários aliada às análises das sugestões de adição, remoção e substituição de ingredientes são etapas fundamentais para a proposição de sistemas de recomendação que tenham como meta tornar as receitas mais saudáveis e/ou adequadas ao gosto de um determinado usuário.

7. CONCLUSÕES

Neste trabalho foi apresentada uma extensa caracterização de um importante site brasileiro de compartilhamento de receitas: o Tudo Gostoso. A caracterização foi realizada em três níveis: usuários, receitas e ingredientes. Em relação aos usuários, a análise revela que um número elevado de usuários compartilham suas opiniões (48,4%). Estes comentários englobam desde sugestões para a época do ano mais adequada para fazer a receita, sugestões que tornam, por exemplo, as receitas mais saudáveis e dúvidas referentes ao preparo das receitas. Adicionalmente, a maioria dos comentários realizados são feitos por mulheres.

Em relação as receitas, a categoria com maior número de receitas é a de Doces e Sobremesas, o que pode revelar uma tendência de gosto dos usuários do site. Interessante é que a categoria de Sopas possui somente 1,92% das receitas compartilhadas, fato este que pode estar ligado ao fato de sermos um país tropical, com invernos amenos. Receitas possuem, em média, 7,4 ingredientes distintos.

A modelagem da rede de ingredientes baseadas nas receitas coletadas fornece indícios de padrões de uso de ingredientes encontrados na culinária brasileira. As receitas compartilhadas no site são compostas por 1.397 ingredientes distin-

tos, sendo que estes ingredientes formam uma rede extremamente densa, sugerindo que, apesar do número grande de receitas no site, os ingredientes não variam muito de receita para receita. Além disso, cada ingrediente, em média, provavelmente harmoniza com outros ≈ 96 ingredientes.

Além da caracterização, foi proposto um processo automático para identificação dos casos de adição, remoção e substituição de ingredientes em receitas. A análise realizada, juntamente com o processo automatizado proposto, são essenciais para o desenvolvimento de sistemas de recomendação que poderão ser utilizados para tornar receitas mais saudáveis, ou mais próximas ao gosto de um determinado usuário.

Como trabalhos futuros, além do desenvolvimento de sistemas de recomendação para personalização de receitas, pretende-se formalizar e analisar a rede de sabores presentes nos ingredientes utilizados nas culinária brasileira, estendendo os resultados apresentados em [1].

Agradecimentos

Este trabalho é parcialmente financiado pelas Agências de Fomento CNPq, CAPES e FAPEMIG.

8. REFERENCES

- [1] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási. Flavor network and the principles of food pairing. *Nature, Scientific Reports*, (196), 2011.
- [2] J. Freyne and S. Berkovsky. Recommending food: reasoning on recipes and ingredients. In *Proceedings of the 18th international conference on User Modeling, Adaptation, and Personalization*, UMAP'10, pages 381–386, 2010.
- [3] G. Geleijnse, P. Nachtigall, P. van Kaam, and L. Wijgengangs. A personalized recipe advice system to promote healthful choices. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 437–438, 2011.
- [4] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.
- [5] M. Svensson, K. Höök, and R. Cöster. Designing and evaluating kalas: A social navigation system for food recipes. *ACM Trans. Comput.-Hum. Interact.*, pages 374–400, 2005.
- [6] C.-Y. Teng, Y.-R. Lin, and L. A. Adamic. Recipe recommendation using ingredient networks. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 298–307, 2012.
- [7] M. Ueda, M. Takahata, and S. Nakajima. User's food preference extraction for cooking recipe recommendation. In *SPIM*, pages 98–105, 2011.
- [8] Q. Zhang, R. Hu, B. M. Namee, and S. J. Delany. Back to the future: Knowledge light case base cookery. In *ECCBR Workshops*, pages 239–248, 2008.