Análise e Caracterização de Receitas Gastronômicas na Web

Alternative Title: Analysis and Characterization of Gastronomic Recipes on the Web

Edwaldo S. Rodrigues
Universidade Federal de Ouro Preto
Departamento de Computação
Ouro Preto - MG - Brasil
edwaldoroadsf1@yahoo.com.br

Álvaro R. Pereira Jr.
Universidade Federal de Ouro Preto
Departamento de Computação
Ouro Preto - MG - Brasil
alvaro@iceb.ufop.br

RESUMO

A internet nos dias atuais tem desempenhado um importante papel em toda a sociedade, facilitando a realização de serviços e tendo diversos fins. Um dos serviços que surgiram a partir da internet foram os sistemas colaborativos, onde diversos usuários criam o conteúdo dos sistemas através de experiências pessoais. Um dos tipos de sistemas colaborativos existentes atualmente é o de compartilhamento de receitas gastronômicas. A área de Recuperação da Informação na *Web* tem crescido o interesse no que diz respeito a recuperar as informações contidas nesse ambiente e estudá-las de forma a descobrir novo conhecimento, como a descoberta de receitas saudáveis, o que acontece por meio do uso de técnicas de mineração de dados textuais. Nesse escopo, o presente trabalho tem o objetivo de estudar características de dados de receitas gastronômicas presentes em *sites* de compartilhamento de receitas, levando em consideração características dos ingredientes, comentários, número de usuários, categorias, entre outras informações associadas às receitas. Objetiva-se ainda identificar os verbos associados às instruções de preparo que representam ações utilizadas no preparo da receita, como "assar", "fritar" e outros. Dessa forma, foi possível analisar algumas relações entre as ações verbais e a maneira como se dá o processo de preparo de uma receita.

Palavras-Chave

Mineração de Dados, Receitas Gastronômicas, Mídia Social

ABSTRACT

Internet has played nowadays an important role in society, being the means for services of diverse purposes to be delivered. One of the services that have gained attention on internet is the collaborative systems, in which multiple users create content based on their own personal experiences. An emerging class of collaborative systems is currently the gastronomical recipes sharing services. The area of Web Information Retrieval has grown interest in retrieving information in the recipes environment in order to discover new knowledge, such as the discovery of healthy recipes, which happens by employing textual data mining techniques. In this scope, this work analyzes features of gastronomic recipes present in specialized sites, taking into account characteristics of ingredients, comments, number of users, categories, among other information related to the recipes in the target sites.

SBSI 2016, May 17th-20th, 2016, Florianópolis, Santa Catarina, Brazil

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data Mining

General Terms

Experimentation

Keywords

Data Mining, Gastronomic Recipes, Social Media

1. INTRODUÇÃO

A facilidade de acesso à *Web* tem crescido a cada dia. Atividades são realizadas instantaneamente e pode-se verificar uma demanda enorme de uso para diversificados fins. Desde os primórdios da *Web*, percebe-se que uma série de aplicativos tem sido desenvolvidos no intuito de facilitar a realização de algumas tarefas. Uma aplicação que tem crescido recentemente na *Web* são os sistemas colaborativos que, segundo [6], são sistemas onde as informações presentes são dispostas por vários usuários, havendo comunicação, coordenação e cooperação entre eles, mantendo, assim, a troca de informações e gerando conteúdo.

O trabalho desenvolvido em [7] apresenta conceitos, exemplos e questionamentos a respeito dos sistemas colaborativos. Um dos questionamentos tem por objetivo verificar se o futuro da sociedade estará nas multidões, uma vez que cada vez mais se utilizam de informações coletivas, das multidões, com o intuito de criar produtos e soluções. Apresenta, ainda, conceitos sobre temas relacionados como: *crowdsourcing*, inteligência coletiva, interação multidão-computador,

entre outros. Observa-se que atualmente diversas aplicações utilizam-se do conhecimento coletivo como fonte de dados em seu desenvolvimento. Um exemplo é o Flickr[1], que conforme salientado em [8], utiliza-se de informações compartilhadas por usuários em marcações de fotografias.

Uma classe de serviços colaborativos que vem crescendo na *Web* é representada por meio de *sites* de compartilhamento de receitas gastronômicas. Usuários de diversas localidades acessam os *sites* e compartilham suas experiências gastronômicas, criando um espaço de conhecimento e gerando conteúdo diariamente, a partir da inserção de diversificadas receitas, cada uma com suas particularidades e seus toques pessoais, além do conhecimento proveniente dos comentários realizados pelos usuários cadastrados nestes *sites*. Ressalta-se que a maioria dos *sites* de compartilhamento de receitas podem ser considerados sistemas colaborativos, uma vez que, as receitas ali apresentadas, são incluídas por diversos usuários.

Percebe-se, portanto, a necessidade de estudar os dados que estão sendo gerados e, por meio desses, encontrar maneiras de aplicar o conhecimento encontrado, melhorando a prestação de serviços e angariando cada vez mais usuários.

Informações aprofundadas sobre as receitas se fazem importantes para que se possa entender algumas características relacionadas à gastronomia. Como exemplo, pode-se destacar: quais os principais ingredientes utilizados nas receitas, quais as principais receitas, quais as categorias de receitas mais presentes, principais formas de preparo de uma receita tendo como base as instruções de preparo, enfim, características que permitam que informações "escondidas" possam ser descobertas.

Com o objetivo de encontrar novos conhecimentos sobre as receitas, este trabalho utiliza das informações presentes nas receitas, como as instruções de preparo e a lista de ingredientes, em busca de identificar características que auxiliem na identificação de receitas que possam apresentar determinadas características de interesse do usuário, como exemplo, encontrar receitas que possam ser consideradas mais saudáveis, levando-se em consideração apenas suas formas de preparo. Dessa forma, mediante as informações textualmente disponibilizadas nas instruções de preparo e na lista de ingredientes, este trabalho apresenta uma análise do comportamento dos dados associados às receitas. Por exemplo, uma análise efetuada consiste na identificação da principal característica da forma de preparo da receita, que pode ser assada, cozida, crua, refogada, frita, entre outras; ou até mesmo uma combinação desses processos.

Nesse contexto, o trabalho desenvolvido inicialmente efetua a coleta de receitas gastronômicas extraídas de diversificadas fontes de dados (*sites* de compartilhamento de receitas gastronômicas). Posteriormente, analisam-se as informações presentes na lista de ingredientes e instruções de preparo, buscando novas informações sobre as receitas. Por fim, uma caracterização da base de dados é efetuada, de maneira a permitir conhecer alguns detalhes das fontes de dados trabalhadas, elucidando algumas características importantes e oferecendo suporte para que novos estudos possam ser realizados.

O restante do artigo está organizado da seguinte maneira. Na Seção 2 são discutidos os trabalhos relacionados. A Seção 3 explicita os processos de coleta e processamento dos dados. A Seção 4 apresenta os resultados das análises realizadas. Por fim, na Seção 5 são apresentadas as conclusões do trabalho, bem como sugestões de trabalhos futuros.

[1]https://www.flickr.com

2. TRABALHOS RELACIONADOS

Existem diversos trabalhos que envolvem receitas gastronômicas associadas a, principalmente, análise e caracterização da base de dados e a criação de sistemas de recomendação de receitas gastronômicas. A seguir serão apresentados alguns trabalhos que abordam as áreas de estudo trabalhadas.

Os autores do trabalho [2] realizam a caracterização e análise de uma rede de ingredientes e receitas, a partir da realização da coleta de um *site* de receitas brasileiro. Após a realização da coleta, foram feitas análises verificando a co-ocorrência de ingredientes, a viabilidade de exclusão de determinado ingrediente e estudos afins. Foi trabalhado o conceito de redes de ingredientes de forma a possibilitar a identificação de ingredientes similares que possivelmente poderiam ser usados em detrimento de um outro de acordo com a similaridade de suas características.

No trabalho [1], os autores criaram uma rede de receitas, onde os ingredientes conectam-se de acordo com seus componentes químicos. Os estudos foram realizados em dois repositórios: um do Ocidente e um do Oriente, onde visualizaram que em média as receitas possuem 8 ingredientes. Verificaram ainda que as bases de dados do Leste Asiático e o Sul da Europa possuem receitas onde os ingredientes não compartilham componentes químicos.

No trabalho [12] são realizadas coletas de receitas de um *site* de receitas e, posteriormente, usando classificadores *Support Vector Machines* (SVMs), tentam descobrir qual a avaliação de uma determinada receita. Os atributos utilizados nesse procedimento foram: ingredientes, instruções de preparo e comentários. O estudo apresentou que a maioria das receitas foram avaliadas com as pontuações 3 e 4, com acurácia de 62% quando analisado somente os comentários.

Os autores do trabalho [11] e [10] propõem sistemas de recomendação de receitas, levando-se em consideração as preferências dos usuários em relação aos ingredientes. Os autores desenvolveram dois trabalhos que se diferem no que tange a pontuação que é dada ao ingrediente quanto à preferência do usuário. No primeiro trabalho [11], verificou-se somente se o usuário gostava ou não de determinado ingrediente, sem levar em consideração a ordem de preferência dos ingredientes. Já o trabalho [10] trata a questão do gosto do usuário, elaborando um *ranking* dos ingredientes preferidos. Em seguida, o usuário entra com informações sobre seu histórico alimentar dos últimos dias no sistema de recomendação, assim o sistema recomenda receitas que levam em consideração as preferências do usuário, além de evitar propor pratos que foram as refeições realizadas recentemente.

O presente trabalho assemelha-se com os trabalhos relacionados, principalmente no que tange à análise e caracterização dos dados, diferindo no ponto em que se realiza o estudo de receitas por meio de ações verbais como aquelas que indicam o modo de preparo (assar, fritar, entre outras). Este trabalho apresenta uma análise mais completa de dados coletados de cinco fontes de dados com características diversas, diferentemente dos demais trabalhos, que na maioria das vezes executa suas análises sobre uma única fonte.

3. COLETA E PROCESSAMENTO DOS DADOS

Nesta seção apresenta-se as fontes de dados utilizadas neste trabalho e o processo de coleta das receitas (Seção 3.1), além dos procedimentos realizados no pré-processamento dos dados (Seção 3.2), e por fim, é apresentado o processo de identificação de ações verbais (Seção 3.3).

3.1 Fontes de dados e processo de coleta

Este trabalho foi realizado sobre receitas coletadas de cinco fontes de dados. A escolha por essas fontes de dados deu-se de acordo com a quantidade de receitas, a importância da fonte no cenário nacional por meio da exibição de programas televisivos, e a consideração de uma das preocupações atuais das pessoas que é a alimentação saudável. Ressalta-se que apesar deste trabalho ser baseado no conhecimento coletivo, nem todas as fontes de dados utilizadas são de caráter coletivo, visto que suas receitas não são disponibilizadas por usuários da internet. A seguir são apresentadas as fontes de dados escolhidas:

- Tudo Gostoso[2]: é um *site* brasileiro criado em 2005 que apresenta receitas compartilhadas por usuários;
- Receitas.com[3]: receitas disponibilizadas por meio do *site* Globo.com, contendo receitas gastronômicas que são apresentadas em programas da emissora Globo, como Mais Você e Estrelas. O *site* também permite o envio de receitas por usuários cadastrados que compartilham seus conhecimentos culinários;
- Edu Guedes[4]: mais um *site* onde se verifica apelo televisivo, apresentando receitas de um programa de culinária da rede Record. As receitas disponibilizadas são apresentadas unicamente pelo *chef* Edu Guedes;
- Cybercook[5]: *site* onde são apresentadas receitas culinárias e que também é de propriedade da rede Record, no entanto, não representa um programa televisivo. As receitas são disponibilizadas por usuários;
- Dieta e Receitas[6]: *site* que apresenta receitas culinárias referentes a uma dieta chamada Dukan[7]. Todas as receitas presentes são especiais para dietas.

Vale ressaltar que a coleta foi realizada entre 30/08 e 07/11 de 2014, e que para esse processo utilizou-se o *Crawler4j*[8]. Todas as receitas de cada um dos cinco *sites* foram coletadas. Para o processo de coleta, foi necessário analisar a estrutura das páginas a serem coletadas, visando obter as *tags* referentes aos campos que se desejava coletar. Feito isso, foi realizada a adaptação do coletor com as *tags* necessárias e implementada uma aplicação para receber os dados coletados e salvá-los em formato *Extensible Markup Language* (XML), uma vez que esse formato oferece facilidades no manuseio dos dados. Foi desenvolvida uma aplicação diferente para efetuar a coleta de cada uma das fontes de dados.

[2]http://www.tudogostoso.com.br
[3]http://www.gshow.globo.com/receitas
[4]http://receitas.eduguedes.com.br
[5]http://www.cybercook.com.br
[6]http://www.dietaereceitas.com.br
[7]http://www.dietadukan.com.br
[8]https://code.google.com/p/crawler4j/

3.2 Pré-processamento dos dados

As receitas coletadas possuem similaridades em relação à sua estrutura. Em todas as fontes de dados, as receitas apresentam título, autor, tempo de preparo, rendimento, lista de sentenças contendo cada ingrediente e suas unidade de medida e quantidade, e instruções de preparo.

As demais informações presentes nas receitas divergem entre as fontes de dados. A maioria das fontes apresentam informações como o número de votos, avaliação, data de postagem, descrição, tipo de cozinha, categoria, informações referentes à interação dos usuários em relação à receita em redes sociais e comentários inseridos por usuários.

Há ainda campos que são apresentados somente em uma fonte de dados, como observa-se no *site* Receitas.com, onde há o número de pessoas que favoritaram determinada receita. O *site* Edu Guedes também apresenta algumas especificidades, como o número de "gostei" e de "visualizações" do vídeo que segue a receita. Já o *site* Dieta e Receitas apresenta campos como tempo de cozimento, tempo de espera, quantidade de calorias, grau de dificuldade e fase da dieta.

Após a coleta das receitas, foi efetuado o pré-processamento dessas, onde primeiramente foram removidos os acentos e caracteres especiais e a caixa de texto foi convertida para caixa baixa. Esse procedimento foi realizado para todos os dados coletados.

Os ingredientes com sua quantidade e sua unidade de medida são encontrados em uma mesma sentença. Ressalta-se que neste trabalho se denomina como sentenças as expressões onde se verificam a presença dos ingredientes com sua quantidade e unidade de medida, quando as expressões existirem, conforme observa-se no exemplo: "1 copo de leite". Visando tratá-los de maneira separada, fazia-se necessário criar uma heurística que conseguisse desmembrar tais informações.

A princípio, várias receitas foram estudadas de forma a encontrar um padrão que se repetia em grande maioria. Nessa fase, conseguiu-se encontrar alguns padrões que acometia em um número considerável de receitas.

Os principais padrões identificados, relacionavam-se ao uso da preposição "de". Diversas sentenças são apresentadas sem a utilização da preposição "de", como visualiza-se em: "1 cebola picada". Nesse caso, observa-se que não há a presença da unidade de medida, apresentando inicialmente a quantidade seguida pelo ingrediente. Observa-se ainda diversos casos em que há a presença da preposição "de" uma única vez, como ocorre em: "2 copos de leite". Visualiza-se nesse caso que primeiramente, apresenta-se a quantidade; em seguida, após o espaço, a unidade de medida; e por fim, após a preposição, é apresentado o ingrediente. Outro padrão que ocorre com muita frequência é visto em: "1 colher de chá de margarina", onde se verifica a ocorrência da preposição "de" por duas vezes, sendo que a primeira ocorre de forma a unir duas palavras que compõem uma unidade de medida, e que após a segunda ocorrência é apresentado o ingrediente.

Além dos principais padrões citados anteriormente, vários outros padrões foram identificados. O processo de desenvolvimento da heurística foi incremental, uma vez que ao encontrar um novo padrão esse era inserido no código. Uma das principais dificuldades encontradas foi em tratar as várias maneiras de escrita das sentenças. Isso pode ser visto nos exemplos: "2 1/2 xícaras de café", 2 colheres 1/2 de açúcar" ou ainda "2 e 1/2 copos de água". Nos exemplos

citados, verifica-se que a quantidade é a mesma, entretanto, encontra-se escrita de maneiras diferentes, ampliando os padrões encontrados, e elucidando a diversidade encontrada na maneira de se escrever as sentenças. Não faz parte do escopo do presente trabalho apresentar a heurística com todos os detalhes, no entanto, a heurística completa é apresentada no documento[9] "Pseudocódigo da heurística desenvolvida".

3.3 Processo de identificação de ações verbais

Uma das maneiras de encontrar características no processo de preparar uma receita é por meio das ações verbais que são encontradas nas instruções de preparo. Nesse escopo, observou-se a necessidade de identificar os verbos, associando-os às instruções de preparo. Dessa forma, primeiramente, buscou-se na *Web* uma lista de verbos no infinitivo e com suas conjugações. Foi encontrada uma lista de verbos no Wikicionário[10], contendo 77.647 conjugações verbais. Em seguida os verbos foram armazenados na base de dados de receitas.

De posse das instruções de preparo de cada uma das receitas e tendo a lista de verbos, deu-se o processo de identificação de verbos nas instruções de preparo. Para isso, para cada termo de uma instrução de preparo verifica-se no arquivo de verbos se o mesmo está presente. Quando encontrado um verbo, este é armazenado na base de dados de receitas e estabelece-se a relação entre o verbo com a instrução de preparo de uma determinada receita. Desta forma, no final deste processo, obtém-se a lista de verbos encontrados e associados com as instruções de preparo de cada receita.

4. RESULTADOS

Nesta seção é apresentada uma análise das características da base de dados (Seção 4.1), além de analisar as ações verbais identificadas (Seção 4.2).

4.1 Caracterização da base de dados

Esta seção apresenta uma análise sobre os dados coletados de receitas, visando identificar algumas características. A Figura 1 apresenta a composição da base de dados de acordo com a porcentagem de receitas de cada uma das fontes de dados utilizadas. Observa-se na Figura 1 que aproximadamente 60% das receitas que compõem a base de dados são extraídas da fonte Tudo Gostoso. Observa-se ainda que as fontes Dieta e Receitas e Edu Guedes representam cada, menos de 1% das receitas. Entretanto, ao iniciar a coleta não se sabia o número de receitas que cada uma das fontes apresentavam. Ressalta-se ainda que essas fontes foram identificadas como importantes mediante suas características, como a exposição dada às receitas em um programa televisivo, como acontece com a fonte Edu Guedes, e devido à característica da dieta que se observa em Dieta e Receitas. Vale salientar que o número total de receitas coletadas foi de 238.049.

O número total de sentenças que contém ingrediente, quantidade e unidade de medida encontrados nas receitas coletadas foi de 1.971.405, possuindo cada receita em média 8,28 sentenças. Verificou-se ainda que o número de linhas correspondentes a instruções de preparo foi de 1.621.258, tendo em média 6,81 linhas de instruções por receita.

[9]http://migre.me/ozaPp

[10]http://pt.wiktionary.org/wiki/Categoria:Verbo

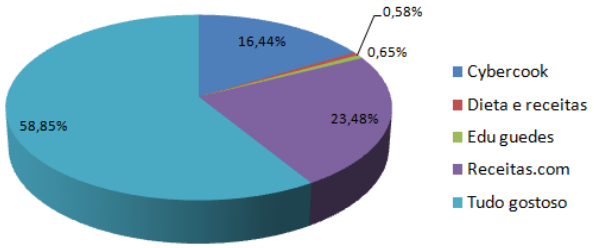

Figura 1: Composição de receitas da base de dados.

No gráfico da Figura 2 é explicitado o número de ingredientes por fonte de dados, tendo no eixo X as fontes de dados e no eixo Y a quantidade de ingredientes. Pode ser visualizado na Figura 2 que o número de ingredientes presentes na base de dados é algo proporcional ao número de receitas apresentadas em cada uma das fontes de dados. Naturalmente, como acontece com qualquer vocabulário, a medida em que a base cresce, o número de ingredientes presentes não cresce na mesma ordem. Por exemplo, a base Tudo Gostoso é quase três vezes maior que a base Receitas.com, no entanto possui pouco mais que o dobro de ingredientes em relação à base Receitas.com.

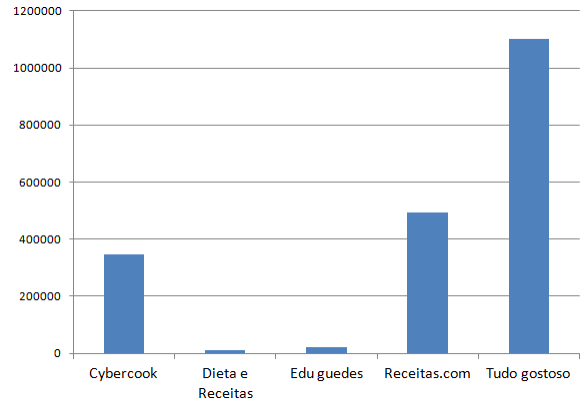

Figura 2: Número de ingredientes encontrados por fonte coletada.

A Figura 3 apresenta uma nuvem de termos com os ingredientes mais frequentes nas receitas. Verifica-se que os ingredientes mais utilizados foram "açúcar", "sal" e "ovo". O ingrediente "sal" ocorre muito frequentemente, sendo comum até mesmo em receitas doces. Já os outros dois ingredientes (açúcar e ovo) relacionam-se a várias receitas encontradas em categorias de bolos, tortas e doces.

Ainda sobre os ingredientes, visando entender o comportamento da base de ingredientes, analisou-se qual a distribuição estatística melhor se adapta aos dados. Para isso, utilizou-se do teste de Anderson Darling conforme [9], na qual a distribuição que mais se aproxima de cada função de distribuição é aquela que tem o menor valor do teste. Nesse escopo, utilizou-se o *software* EasyFit[11] para encontrar a distribuição e posteriormente gerar o gráfico da Função de

[11]http://www.mathwave.com/

Figura 3: Nuvem de termos dos ingredientes mais frequentes nas receitas.

Densidade de Probabilidade (PDF), conforme visualiza-se na Figura 4.

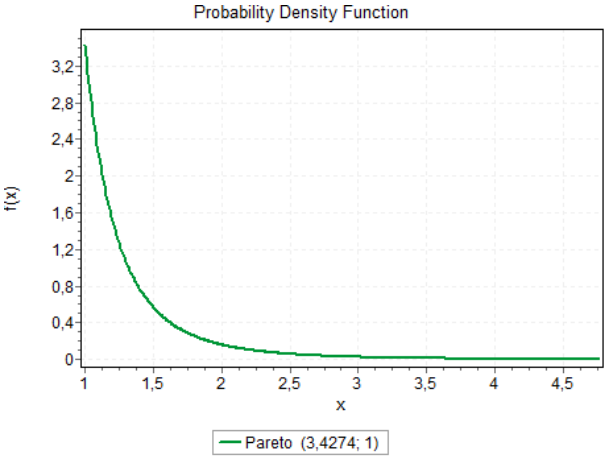

Figura 4: Distribuição de frequência de ocorrência de ingredientes nas receitas.

Verifica-se na Figura 4 que a distribuição que melhor se adapta à base de ingredientes é a distribuição de Pareto com os seguintes parâmetros: α= 3,4274 e β= 1. Conforme ressaltado em [4], [5] e [3], essa distribuição tem por característica a "cauda longa", onde verifica-se nesse caso que poucos ingredientes ocorrem com muita frequência e muitos ingredientes ocorrem poucas vezes.

Com exceção da fonte de dados Edu Guedes, todas as demais possuem em suas receitas avaliações realizadas por usuários que interagem com os *sites*. A Figura 5 apresenta uma sumarização das avaliações das receitas das demais fontes de dados, onde o eixo X representa os possíveis valores ao se avaliar uma receita, estando entre 1.0 e 5.0. Já o eixo Y representa a porcentagem de receitas avaliadas. Observa-se na Figura 5 uma semelhança no padrão das avaliações entre as as fontes de dados (Cybercook, Receitas.com e Tudo Gostoso). Em todas elas verifica-se que o crescimento é similar, onde se visualiza que até a avaliação 2.5, poucas são as receitas contidas nesse intervalo. A partir da avaliação 3.0 verifica-se um crescimento das avaliações chegando ao ápice na avaliação 4.5 para a fonte Cybercook, com cerca de 7% das receitas. Já as fontes Receitas.com e Tudo Gostoso têm como ápice a avaliação 5.0, onde se verificam cerca de 12% e 16% das receitas, respectivamente.

Pode-se perceber uma diferença entre as fontes Cybercook, Receitas.com e Tudo Gostoso, uma vez que as avaliações da fonte Receitas.com apresenta apenas avaliações para valores inteiros, de 1.0 a 5.0, diferentemente das outras duas. Finalmente, analisando as avaliações da fonte Dieta e Receitas, visualiza-se que até a avaliação 2.0 praticamente não há receitas avaliadas neste intervalo. Percebe-se que algumas receitas foram avaliadas em 2.5 e 3.0, entretanto, grande parte das receitas desta fonte foram avaliadas igual ou acima a 3.5, o que é um indicativo da qualidade média das receitas. A boa avaliação das receitas da fonte Dieta e Receitas pode estar associada ao fato de que as receitas presentes nesta fonte são relacionadas à dieta e alimentação saudável.

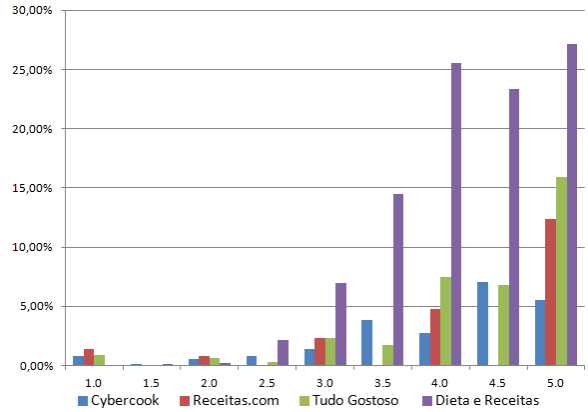

Figura 5: Avaliação das receitas realizadas por usuários.

Conforme supracitado, o tipo de informação presente em cada uma das fontes de dados se diferem. O gráfico da Figura 6 apresenta várias informações das fontes de dados, destacados como rótulos do eixo X do gráfico. No eixo Y têm-se os valores associados a cada uma das informações apresentadas no eixo X. Verifica-se no gráfico da Figura 6 que as fontes Dieta e Receitas e Cybercook apresentam apenas valores referentes ao número de votos, sendo que as demais fontes apresentam informações referentes a cinco atributos, mas que são diferentes entre si. Constata-se ainda que desses cinco atributos, três são em comum: curtidas do Facebook[12], tweets do Twitter[13] e recomendações do Google Plus[14]. Uma semelhança ainda maior é visualizada, uma vez que se observa que há um número maior de curtidas, seguido de recomendações, e por fim, tweets, para todas as três fontes.

Entre as fontes trabalhadas, três possuem comentários sobre as receitas. Nas fontes Tudo Gostoso e Cybercook, como o volume de receitas é maior, foram coletados apenas os dez comentários mais recentes. Já a fonte Dieta e Receitas, todos os comentários foram coletados. O total de comentários coletados foi de 218.316, sendo o Tudo Gostoso responsável por 74,07%, Cybercook 14,62% e Dieta e Receitas 11,31%. Verifica-se ainda que das três fontes que possuem comentários, duas oferecem a opção dos usuários avaliarem os comentários. A única fonte que não oferece essa opção é a Dieta e

[12]https://www.facebook.com
[13]https://twitter.com
[14]https://plus.google.com/

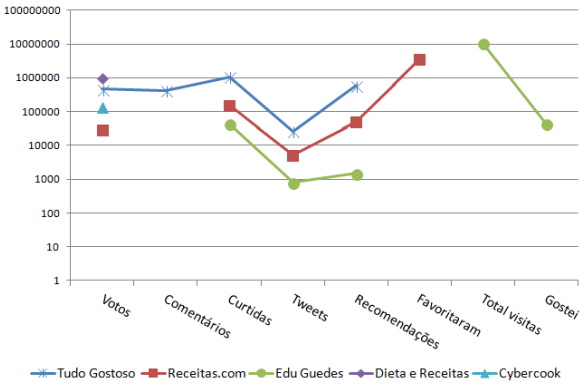

Figura 6: Interação dos usuários nas receitas.

Receitas. No gráfico da Figura 7 apresentam-se as avaliações dadas aos comentários, onde no eixo X encontram-se os possíveis valores para uma avaliação e no eixo Y a porcentagem de comentários avaliados. Analisando a Figura 7, verifica-se que no geral os comentários são bem avaliados, tendo o valor 4 e 5 juntos, aproximadamente 80% das avaliações realizadas para cada uma das fontes. Observa-se ainda que cerca de 3% dos comentários da fonte Cybercook não foram avaliados.

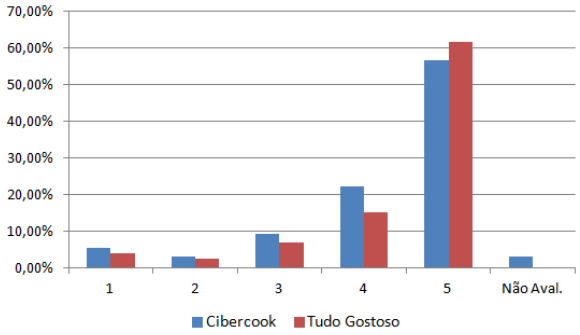

Figura 7: Avaliação dos comentários nas receitas realizadas por usuários.

Em todas as fontes de dados, ao inserir uma nova receita, observa-se a possibilidade de associá-la a uma ou mais categorias. Dessa forma, a base de dados conta com 247 categorias, sendo que essas se subdividem da seguinte forma: Tudo Gostoso apresenta 11 categorias; Receitas.com 62; Cybercook 17; Edu Guedes 69; e Dieta e Receitas 136. Observa-se ainda que há 48 categorias em comum entre duas ou mais fontes de dados.

O gráfico da Figura 8 apresenta as dez categorias que possuem mais receitas associadas, tendo no eixo X as categorias e no eixo Y a quantidade de receitas associadas as categorias. Observa-se na figura que a categoria com o maior número de receitas associadas é a "doces e sobremesas", com a presença de mais de 45.000 receitas. Analisa-se ainda que todas as dez categorias presentes possuem mais de 10.000 receitas associadas. Como há a possibilidade de associar uma receita a mais de uma categoria, há 406.045 associações entre receitas e categorias no total. Observa-se ainda que a soma das receitas associadas às dez categorias mais comuns resultam em 227.270 associações entre categorias e receitas, o que representa 55,97%, ou seja, as categorias presentes na Figura 8 representam mais da metade das associações entre as receitas. Por fim, analisa-se ainda que cinco das dez categorias presentes referem-se à fonte Tudo Gostoso, fato que pode ser explicado devido à quantidade de receitas presentes nessa fonte.

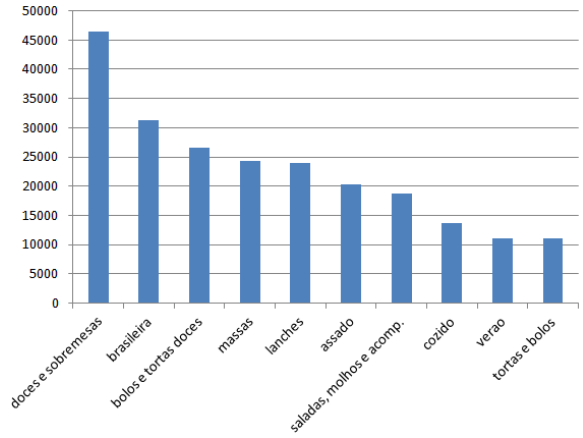

Figura 8: As 10 categorias mais comuns.

Por fim, o gráfico da Figura 9 apresenta o número de usuários que efetuaram a postagem de receitas gastronômicas ou comentários nos *sites* usados como fontes de dados, tendo no eixo X as fontes de dados e no eixo Y a quantidade de usuários. Ressalta-se que o número total de usuários identificados foi de 232.763, sendo que a fonte Edu Guedes apresenta apenas um usuário, sendo esse o próprio *chef* Edu Guedes. Verifica-se na Figura 9 que o número de usuários é relativamente proporcional ao número de receitas das fontes de dados, assim sendo, a fonte Tudo Gostoso apresenta aproximadamente 72,31% dos usuários da base de dados.

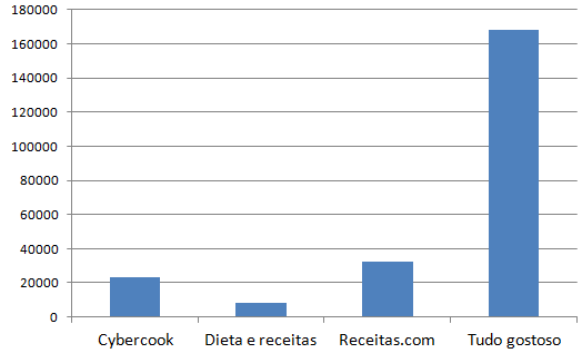

Figura 9: Número de usuários de cada fonte coletada.

4.2 Análise das ações verbais identificadas

Com o procedimento efetuado para encontrar os verbos presentes nas instruções de preparo, possibilitou-se que algumas análises fossem realizadas. Ressalta-se que foram ar-

mazenados os verbos no infinitivo. Na Tabela 1 são apresentados os verbos mais frequentes.

Tabela 1: Lista dos 10 verbos mais frequentes nas receitas.

Posição	Verbo	Frequência	#Receitas	%Receitas
1º	colocar	352.855	169.426	71,17%
2º	pôr	282.163	156.969	65,94%
3º	atar	241.356	144.638	60,76%
4º	misturar	197.577	125.285	52,63%
5º	deixar	172.459	107.955	45,35%
6º	levar	171.314	125.404	52,68%
7º	acrescentar	139.729	91.006	38,23%
8º	ficar	111.693	80.532	33,83%
9º	formar	103.374	75.604	31,76%
10º	reservar	76.059	58.060	24,39%

Ao analisar a Tabela 1, verifica-se que a maioria dos verbos presentes são responsáveis por uma ação típica que deve ser tomada para a realização da receita, como colocar, misturar, acrescentar, entre outros. Analisa-se ainda que o verbo mais frequente "colocar" encontra-se em aproximadamente 22% das instruções de preparo e em 71,17% das receitas.

Ressalta-se que é de interesse identificar a frequência de alguns verbos específicos que permitem identificar quais os processos de execução de determinada receita, visando analisar receitas que possam ser consideradas, por exemplo, mais saudáveis, levando-se em questão a forma de preparo. Na Tabela 2 são apresentados os cinco verbos mais frequentes que retratam como se deu o processo de preparação das receitas, apresentando também o volume de receitas associadas a esses verbos.

Tabela 2: Verbos relacionados à forma de preparo da receita.

Verbo	# Receitas	%Receitas
cozinhar	54.309	22,81%
ferver	29.441	12,37%
assar(forno)	123.683	51,96%
fritar	12.429	5,22%
refogar	6.380	2,68%

Visualiza-se na Tabela 2 o número de receitas que em seu modo de preparo possuem os verbos especificados. No entanto, para o verbo "assar", foram calculados não apenas as ocorrências do verbo, mas também as ocorrências da palavra "forno". Isso porque quando a palavra forno é usada nas instruções de preparo, esta refere-se à ação de assar algo. Ressalta-se assim, que quando não for identificado o verbo "assar", mas for identificada a palavra forno em instruções de preparo de uma mesma receita, essa receita pode ser considerada como um assado. Pode ser verificado, assim, que mais de 50% das receitas são levadas ao forno em seu modo de preparo. A ação "fritar", que pode ser considerada menos saudável, foi apresentada em apenas 5,22% das receitas. Verifica-se ainda que essas cinco características de preparo de uma receita representam juntas 95,04% das receitas.

Há a possibilidade de uma mesma receita apresentar mais de uma das ações encontradas na Tabela 2. A Tabela 3 retrata a porcentagem de receitas que apresentam interseções em mais de uma das ações em seu modo de preparo. Verifica-se na Tabela 3 que uma receita pode conter ações de "cozinhar" em uma etapa e também ações de "assar", por exemplo, em outra etapa. Percebe-se ainda que "cozinhar" e "ferver" são as ações que mais co-ocorrem em uma receita, constituindo 4,73% das receitas.

Tabela 3: Ações verbais que co-ocorrem em receitas.

Verbos	# Receitas	%Receitas
cozinhar + ferver	11.269	4,73%
cozinhar + fritar	4.922	2,07%
cozinhar + refogar	3.209	1,35%
cozinhar + assar	2.797	1,17%
fritar + ferver	2.397	1,01%
ferver + assar	1.699	0,71%
ferver + refogar	1.434	0,60%
fritar + refogar	1.026	0,43%
fritar + assar	259	0,11%
refogar + assar	133	0,06%

Ainda sobre a análise de verbos, a Figura 10 apresenta uma nuvem de palavras que ilustra por meio da frequência de ocorrências os 40 principais verbos. Quanto mais frequente o verbo, maior o tamanho da fonte. Assim, observa-se que os principais verbos encontrados são os mesmos presentes na Tabela 1.

Figura 10: Nuvem de termos com os verbos mais frequentes.

Por fim, ainda sobre os verbos, verificou-se a que distribuição de probabilidade os dados mais se aproximam, conforme pode ser visualizado no gráfico da Figura 11. Similarmente à distribuição da base de ingredientes a distribuição que melhor se adapta aos verbos é a distribuição de Pareto, com os parâmetros: α= 0,89416 e β= 1. Comparando as distribuições da base de ingredientes e dos verbos, verifica-se que há um decaimento mais acentuado da frequência de ocorrência dos verbos em relação aos ingredientes, evidenciando mais claramente a presença da "cauda longa" na análise de verbos.

5. CONCLUSÕES E TRABALHOS FUTUROS

Para o presente trabalho, realizou-se a coleta de 238.049 receitas gastronômicas, coletadas de cinco diferentes serviços especializados. Em seguida, identificou-se os verbos que são responsáveis pelas ações realizadas no processo de preparo de uma receita. Visando obter conhecimento amplo sobre receitas gastronômicas, efetuou-se uma análise da base de dados. Assim, verificou-se quais os principais ingredientes utilizados e as principais categorias associadas às receitas, bem como estudou-se a avaliação das receitas e informações

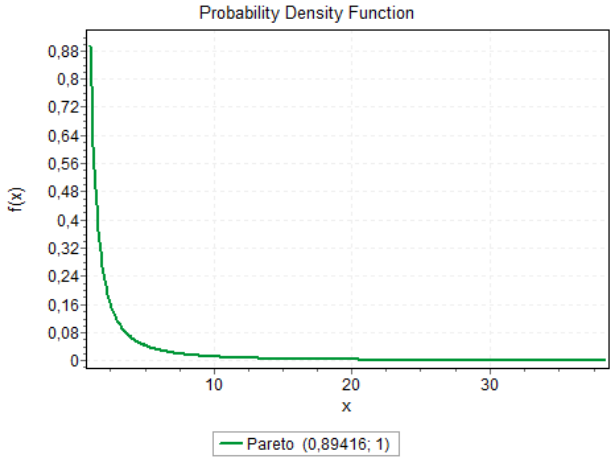

Figura 11: Distribuição de frequência de ocorrência de verbos nas receitas.

a respeito da interação dos usuários, entre outras características.

Constatou-se, por exemplo, que receitas fritas, que geralmente são receitas bastante consumidas, não se mostraram numerosas (5,22%). Já receitas assadas, consideradas mais saudáveis em relação à forma de preparo, apresentaram-se em 51,96% das receitas. Verificou-se também que os três principais ingredientes utilizados nas receitas foram: "açúcar", "sal" e "ovo", principalmente por serem ingredientes utilizados na preparação de receitas de diversas categorias, principalmente doces e sobremesas, que foi a categoria com maior número de receitas. Por fim, verificou-se o grande número de usuários interagindo com os *sites*, o que ilustra a importância de se estudar informações disponibilizadas no ambiente gastronômico da *Web*.

Como trabalhos futuros, almeja-se oferecer uma plataforma integrada (com receitas de diversas fontes) e nesta plataforma prover ao usuário a possibilidade de busca por uma receita levando em consideração a classe do prato. Para isso, primeiramente há a necessidade de categorizar as receitas de acordo com o prato que representam, de forma que o usuário terá a opção de escolher entre receitas de um determinado prato, e logo poder filtrar pela forma de preparo, por exemplo. Ainda sobre melhorar a procura de receitas por parte dos usuários, com este estudo, verificou-se a divergência das classificações das receitas em diferentes categorias de receitas. Assim, uma outra possibilidade a ser considerada, consiste em criar uma nova maneira de categorização das receitas que seja mais homogênea entre as diferentes fontes de dados.

Espera-se também encontrar uma base de dados de calorias dos ingredientes e a partir desta, identificar receitas menos calóricas ou mais saudáveis de acordo com definções fornecidas por especialistas da área de nutrição. Almeja-se ainda, identificar quais os ingredientes e/ou modos de preparo são fundamentais para o sucesso de uma receita e quais são apenas incrementos, visando assim oferecer a possibilidade de sugerir trocas de ingredientes mediante as preferências culinárias dos usuários, ou mesmo efetuando a substituição por ingredientes mais saudáveis.

6. AGRADECIMENTOS

Os autores agradecem à Universidade Federal de Ouro Preto, CAPES, FAPEMIG e CNPq por apoiarem o desenvolvimento desta pesquisa.

7. REFERÊNCIAS

[1] Y.-Y. Ahn, S. E. Ahnert, J. P. Bagrow, and A.-L. Barabási. Flavor network and the principles of food pairing. *Scientific reports*, 1, 2011.

[2] W. M. Ferreira, A. P. C. da Silva, F. Benevenuto, and L. H. Merschmann. Comer, comentar e compartilhar: Análise de uma rede de ingredientes e receitas. In *Proceedings of Brazilian Symposium on Collaborative Systems*, page 120. Sociedade Brasileira de Computação, 2013.

[3] T.-Y. Ko, C.-J. Tseng, H.-H. Chen, J.-J. Ding, and N. Babaguchi. Efficient dc term encoding scheme based on double prediction algorithms and pareto probability models. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE, 2013.

[4] H. Krishna and P. S. Pundir. Discrete burr and discrete pareto distributions. *Statistical Methodology*, 6(2):177–188, 2009.

[5] S. A. Mushtaq and A. A. Rizvi. Statistical analysis and mathematical modeling of network (segment) traffic. In *Emerging Technologies, 2005. Proceedings of the IEEE Symposium on*, pages 246–251. IEEE, 2005.

[6] M. Pimentel, M. A. Gerosa, D. Filippo, A. Raposo, H. Fuks, and C. J. P. Lucena. Modelo 3c de colaboração para o desenvolvimento de sistemas colaborativos. *Anais do III Simpósio Brasileiro de Sistemas Colaborativos*, pages 58–67, 2006.

[7] D. Schneider, J. de Souza, and K. Moraes. Multidões: a nova onda do cscw. *Proceedings of the SBSC & CRIWG-VIII Simpósio Brasileiro de Sistemas Colaborativos. Paraty, Brazil*, 2011.

[8] B. Sigurbjörnsson and R. Van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336. ACM, 2008.

[9] M. A. Stephens. Asymptotic results for goodness-of-fit statistics with unknown parameters. *The Annals of Statistics*, 4(2):357–369, 03 1976.

[10] M. Ueda, S. Asanuma, Y. Miyawaki, and S. Nakajima. Recipe recommendation method by considering the userś preference and ingredient quantity of target recipe. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, 2014.

[11] M. Ueda, M. Takahata, and S. Nakajima. Userś food preference extraction for personalized cooking recipe recommendation. In *Proc. of the Second Workshop on Semantic Personalized Information Management: Retrieval and Recommendation*, 2011.

[12] N. Yu, D. Zhekova, C. Liu, and S. Kübler. Do good recipes need butter? predicting user ratings of online recipes. In *Proceedings of the Cooking with Computer workshop at the International Joint Conference on Artificial Intelligence (IJCAI2013)*, 2013.