UNIVERSIDADE FEDERAL DE OURO PRETO

Análise de Receitas Visando a Descoberta de Conhecimento sobre Pratos Gastronômicos

Edwaldo Soares Rodrigues
Universidade Federal de Ouro Preto

Orientador: Álvaro Rodrigues Pereira Júnior

Dissertação submetida ao Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto para obtenção do título de Mestre em Ciência da Computação

Ouro Preto, outubro de 2015

Análise de Receitas Visando a Descoberta de Conhecimento sobre Pratos Gastronômicos

Edwaldo Soares Rodrigues
Universidade Federal de Ouro Preto

Orientador: Álvaro Rodrigues Pereira Júnior

R696a Rodrigues, Edwaldo Soares.
Análise de receitas visando a descoberta de conhecimento sobre pratos gastronômicos [manuscrito] / Edwaldo Soares Rodrigues. - 2015.
98f.: il.: color; grafs; tabs.

Orientador: Prof. Dr. Álvaro Rodrigues Pereira Júnior.

Dissertação (Mestrado) - Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Departamento de Computação. Ciência da Computação.
Área de Concentração: Recuperação e Tratamento da Informação.

1. Recuperação da Informação. 2. Mineração de dados (Computação). 3. Gastronômia. I. Pereira Júnior, Álvaro Rodrigues. II. Universidade Federal de Ouro Preto. III. Titulo.

CDU: 64.031.42:004

Catalogação: www.sisbin.ufop.br

MINISTÉRIO DA EDUCAÇÃO E DO DESPORTO
Universidade Federal de Ouro Preto
Instituto de Ciências Exatas e Biológicas – ICEB
Programa de Pós-Graduação em Ciência da Computação

decom
departamento de computação

Ata da Defesa Pública de Dissertação de Mestrado

Aos nove dias do mês de outubro de 2015, às 15 horas na Sala de Seminários do DECOM no Instituto de Ciências Exatas e Biológicas (ICEB), reuniram-se os membros da banca examinadora composta pelos professores: **Prof. Dr. Álvaro Rodrigues Pereira Júnior (presidente e orientador), Prof. Dr. Luiz Henrique de Campos Merschmann, Prof. Dr. Fabrício Benevenuto de Souza, Prof. Dr. José Renato Carvalho e Profa. Dra. Débora Maria Barroso Paiva**, aprovada pelo Colegiado do Programa de Pós-Graduação em Ciência da Computação, a fim de arguirem o mestrando **Edwaldo Soares Rodrigues**, com o título **"Análise de receitas visando a descoberta de conhecimento sobre pratos gastronômicos"**. Aberta a sessão pelo presidente, coube ao candidato, na forma regimental, expor o tema de sua dissertação, dentro do tempo regulamentar, sendo em seguida questionado pelos membros da banca examinadora, tendo dado as explicações que foram necessárias.

Recomendações da Banca:

(X) Aprovada sem recomendações

( ) Reprovada

( ) Aprovada com recomendações: _____

Banca Examinadora:

_____
Prof. Dr. Álvaro Rodrigues Pereira Júnior

_____
Prof. Dr. Luiz Henrique de Campos Merschmann

_____
Prof. Dr. Fabrício Benevenuto de Souza

_____
Prof. Dr. José Renato Carvalho

_____
Profa. Dra. Débora Maria Barroso Paiva

_____
Prof. Dr. Luiz Henrique de Campos Merschmann
Coordenador do Programa de Pós-Graduação em Ciência da Computação
DECOM/ICEB/UFOP
Ouro Preto, 09 de outubro de 2015.

Dedico este trabalho a Deus, sempre presente em minha vida, aos meus pais José *e* Adeilda (in memorian)*, a minha irmã* Eliane*, por estarem sempre ao meu lado e aos meus amigos, pela amizade e cumplicidade.*

Análise de Receitas Visando a Descoberta de Conhecimento sobre Pratos Gastronômicos

Resumo

Nos dias atuais, a internet 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 por meio de experiências pessoais. Um dos vários sistemas colaborativos existentes atualmente são os de compartilhamento de receitas gastronômicas. A área da 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 identificar relações como os principais ingredientes utilizados no preparo de um prato, que podem ser identificadas por meio do uso de técnicas de Mineração de Dados textuais. Nesse escopo, o presente trabalho propõe o desenvolvimento de uma metodologia de descoberta de conhecimento em receitas gastronômicas, usando receitas coletadas de diversas fontes de dados. Para isso, informações como os ingredientes, quantidades, unidades de medida, instruções de preparo e outras características associadas às receitas são descobertas. Com os resultados encontrados e avaliados por meio do estudo de caso e das experimentações apresentadas nesta dissertação, este trabalho representa um primeiro passo para o desenvolvimento de um serviço que, além de agregar receitas de diversas fontes, explora o conhecimento coletivo que pode ser descoberto ao se analisar centenas de milhares de receitas disponíveis na rede.

Recipes Analysis for Knowledge Discovery on Gastronomic Dishes

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 proposes the development of a methodology for knowledge discovery in gastronomic recipes, using data collected from various sources on the Web. Information such as list of ingredients, quantities, units of measurement, preparation instructions, among others, are discovered. With the results obtained and evaluated through a case study and by means of experiments for effectiveness presented in this thesis, this work represents the first step towards the development of a service that, in addition to aggregating recipes from various sources, it explores the wisdom of the crowds in order to extract collective knowledge from hundreds of thousands of recipes available on the web.

Declaração

Esta dissertação é resultado de meu próprio trabalho, exceto onde referência explícita é feita ao trabalho de outros, e não foi submetida para outra qualificação nesta nem em outra universidade.

Edwaldo Soares Rodrigues

Agradecimentos

Esse talvez seja um dos momentos mais difíceis, pois foram tantos que de uma forma ou de outra contribuíram para o alcance dessa vitória, e que por isso são merecedores da minha gratidão.

Primeiramente, gostaria de agradecer a Deus pelo dom da vida, e por estar sempre do meu lado, me guiando e abençoando, seja em momentos bons ou em momentos de dificuldades. Quantas vezes me peguei conversando com você ó meu Senhor, muitos pedidos o fiz e muitas graças me foram concedidas, e por isso agradeço imensamente por me conceder a graça de conquistar esta vitória de grande importância em minha vida!

Ter um lugar para ir é lar. Ter alguém para amar é família. Ter os dois é benção! É exatamente assim que penso, e pensamento que fiz valer durante esses anos de mestrado, afinal de contas a cada dificuldade encontrada o único lugar que queria ir e me sentia bem era em meu lar. Ter alguém para poder lhe apoiar, nem que seja por meio de uma simples frase como: Estou com saudades! É família. E agradeço imensamente aos meus pais *José Raimundo Rodrigues* e *Adeilda Soares Rodrigues*, fonte de minha inspiração, e a minha melhor amiga, minha irmãzinha *Eliane Soares Rodrigues* por sempre estarem do meu lado, me auxiliando e me dando forças para que eu concluísse essa etapa. Aos responsáveis por essa vitória o meu sincero e muito obrigado!

Não poderia deixar de mencionar aqui e agradecer a cada um dos meus familiares que sempre acreditaram e me incentivaram. A vocês o meu obrigado!

Agradeço aos meus amigos de toda a vida, da infância, da graduação, do trabalho, da república, enfim, todos aqueles que de alguma forma fazem parte desta conquista e que ficaram tão felizes quanto eu. Especialmente aos amigos do Barreiro e da república, uma vez que foram os que mais vivenciaram esse momento.

Foram muitas as dificuldades vivenciadas no mestrado, mas hoje vejo que estas eram

necessárias e me sinto muito feliz por tê-las superado. Mas não as superei sozinho, gostaria aqui de agradecer imensamente aos meus companheiros de percurso, especialmente ao *Alexandre*, *Felipe* e *Willyan* por acreditarem em mim e por me darem todo o apoio e respaldo para seguir em frente. Gostaria de agradecer também a outros tantos que se fizeram importantes nesta etapa, como o *Edwin*, *Leandro*, *Matheus*, *Walter* e *Wander*.

Por fim, mas não menos importantes, agradeço a todos os professores e funcionários da UFOP que estiveram envolvidos ao meu mestrado, principalmente ao meu orientador prof. *Álvaro Rodrigues* por confiar em meu trabalho e por me apoiar e encorajar nos momentos de dificuldades.

A todos o meu muito obrigado!

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 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: lazer, cultura, aprendizagem, negócios, entre tantos outros. A *Web* surgiu inicialmente com o intuito de compartilhar dados entre os membros dos diversos projetos de pesquisa em andamento no Conselho Europeu para Pesquisa Nuclear (CERN), conforme ressalta Monteiro (2001). No entanto, com o passar do tempo, percebeu-se que ali se encontrava uma ferramenta que apresentava um enorme potencial, uma vez que, cada vez mais pessoas possuíam acesso e estavam mais adeptas aos serviços prestados, além do crescimento do volume da informação pesquisável.

Desde os primórdios da *Web*, percebe-se que uma série de aplicações têm sido desenvolvidas no intuito de facilitar a realização de algumas tarefas que até então eram realizadas somente de forma presencial, ou até mesmo com um tempo de realização relativamente longo se comparado ao que se vê nos dias atuais. Algumas das aplicações que tiveram e continuam tendo grandes proporções de uso são emails, redes sociais, *sites* informativos em geral, sistemas colaborativos, entre outros.

Entre os inúmeros serviços que começaram a ser oferecidos, encontram-se as aplicações que possibilitam a interação entre os vários usuários da *Web*, mantendo nesse ambiente a constante troca de informações que contribui cada vez mais para a realização de tarefas cotidianas. Uma aplicação que tem crescido recentemente na *Web* são os sistemas colaborativos que, segundo Pimentel et al. (2006), 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 por Schneider et al. (2011) 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 salientam Sigurbjörnsson and Van Zwol (2008), 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.

Mediante ao crescimento do número de usuários que acessam a *Web* e ao aumento do conteúdo de dados pesquisáveis, percebe-se que há 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. Nesse contexto, surge o interesse da computação em áreas como Recuperação da Informação na *Web* e Mineração de Dados em utilizar essas informações disponibilizadas com o intuito de contribuir com a experiência do usuário em sua tarefa de obter conhecimento sobre receitas gastronômicas.

Atualmente, muitas são as opções de escolha de uma receita gastronômica para se cozinhar um determinado prato. De acordo com o dicionário da língua portuguesa Priberam[2], receita gastronômica é uma "fórmula que indica os ingredientes e o modo de preparar um prato". Já a Wikipedia[3], define receita gastronômica como "uma sequência de passos para a preparação de pratos e alimentos". Conforme as definições dadas,

[1] https://www.flickr.com
[2] http://www.priberam.pt/dlpo/receita
[3] https://pt.wikipedia.org/wiki/Receita_(culinária)

verifica-se que há uma diferenciação entre os termos receita e prato. Uma forma de diferenciar esses termos pode ser visualizada no seguinte exemplo: "Hoje Maria preparou Lasanha no almoço.". Nesta frase, verifica-se que o nome do prato preparado é "Lasanha". Entretanto, há diversas receitas de Lasanha, as quais poderiam ser utilizadas para preparar essa refeição. Pode ser observado, assim, que o conceito de prato está associado ao nome dado à refeição. Já o conceito de receita está relacionado às diferentes formas de preparo de um determinado prato.

A Figura 1.1 ilustra a situação exposta acima. Como pode ser visualizado, são várias as opções apresentadas para preparar um prato. Entretanto, pode ser difícil para o usuário ter que verificar entre várias, qual a receita ele irá preparar, uma vez que isso pode demandar tempo para analisar e escolher. Pode acontecer de a receita que ele escolheu possuir algum ingrediente que ele não tenha no momento em casa. Dessa forma, um desafio é identificar entre as várias opções de receitas disponíveis para um prato, uma receita que represente bem o prato escolhido, selecionando as receitas mais convenientes para o usuário.

Assim, a hipótese deste trabalho é que efetuar uma análise de dados de *sites* de receitas gastronômicas a fim de descobrir conhecimento sobre pratos, pode ser útil ao usuário que busca uma receita, tendo em vista que ele pode obter receitas que estejam mais alinhadas às suas necessidades ou preferências, bem como poderá ter maior autonomia no processo de cozinhar, uma vez que poderá escolher por receitas conforme a disponibilidade ou preferência por ingredientes ou mesmo de acordo com a forma de preparo desejada de um prato.

1.1 Objetivos

O objetivo principal deste trabalho consiste em desenvolver uma metodologia para descoberta de conhecimento sobre pratos gastronômicos, incluindo informações sobre receitas gastronômicas e seus ingredientes, quantidades e unidades de medida, bem como suas possíveis formas de preparo e outras informações associadas aos pratos, explorando a inteligência coletiva de especialistas que apresentaram suas receitas em variados serviços presentes na *Web*.

Os objetivos específicos do trabalho são apresentados a seguir:

- Realizar um estudo de caracterização das bases de dados coletadas e geradas a

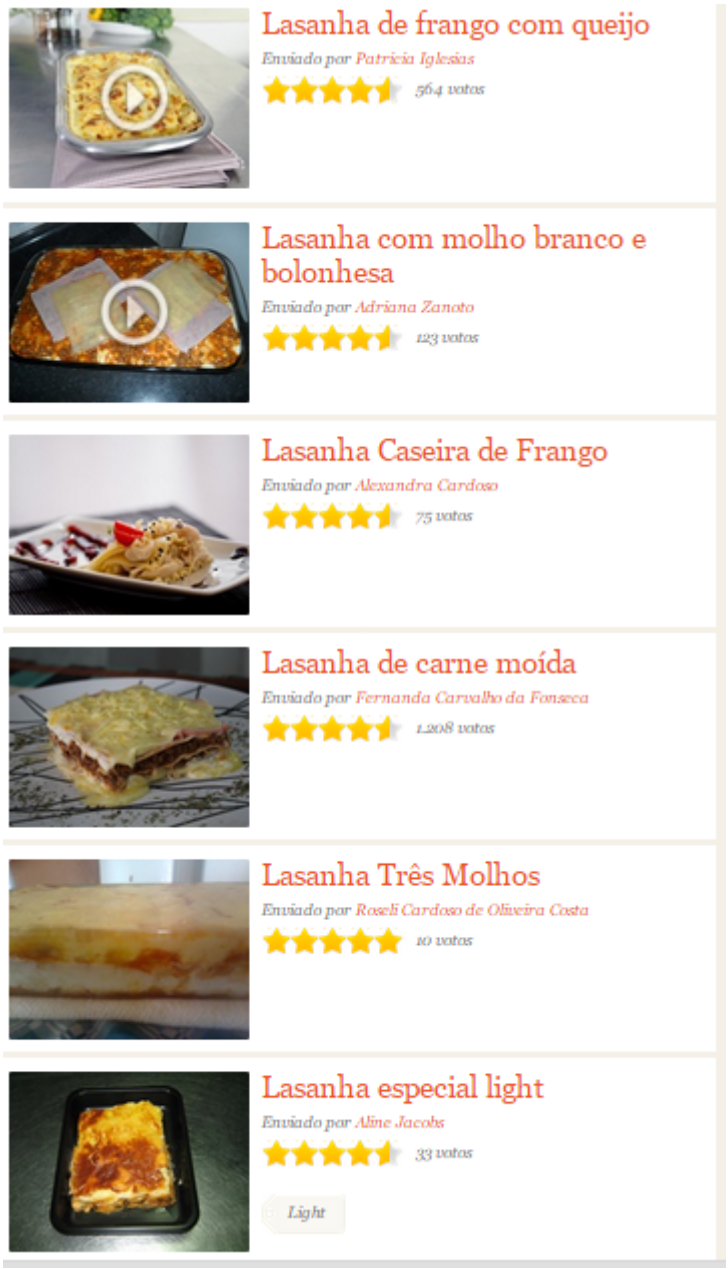


Figura 1.1: Algumas das possibilidades de receitas que podem ser usadas na preparação do prato "Lasanha".

partir da metodologia a fim de se ter um embasamento para o desenvolvimento da metodologia.

- Avaliar a eficácia da resposta dos métodos principais da metodologia, com o intuito de verificar a qualidade do processo de descoberta de conhecimento proposto.

- Propor um serviço de busca, seleção e visualização de receitas gastronômicas que

explore o conhecimento descoberto através da metodologia proposta.

- Disponibilizar publicamente parte do conhecimento descoberto, em forma de uma taxonomia de ingredientes usados em pratos da gastronomia brasileira.

1.2 Justificativa

O uso da *Web* tem crescido cada vez mais, bem como o acesso a *sites* de compartilhamento de receitas gastronômicas, contribuindo com a geração de dados, aumentando assim o número de receitas compartilhadas e oferecendo mais opções aos usuários. Apesar desse aumento, uma questão levantada sobre os *sites* de compartilhamento de receitas gastronômicas gira em torno da maneira como são apresentadas as informações, uma vez que, o tratamento dado é por receitas e não por pratos. Uma das vantagens em apresentar pratos gastronômicos, consiste em filtrar as melhores receitas e, assim, apresentá-las para a preparação de um prato, ou adequar um prato de acordo com as necessidades ou preferências do usuário.

Importa ressaltar ainda que as receitas dos serviços atuais disponíveis são dispostas de uma forma geralmente engessada, não oferecendo opções ao usuário de alterações na mesma, não permitindo assim a adaptação da receita com a realidade do usuário. Isso poderia ser mudado, de forma a disponibilizar a receita com base nos ingredientes que o usuário possui ou não possui em casa, bem como permitir que o usuário efetue manipulações nos ingredientes de uma receita conforme sua preferência culinária. A possibilidade de se escolher uma receita com base nos ingredientes é algo que já existe nos Estados Unidos. O Allrecipes[4] permite essa opção, entretanto, limitam em quatro o número máximo de ingredientes que podem ser buscados.

Outra questão abordada consiste no paradoxo entre o grande volume de receitas encontradas nos *sites* de receitas atuais e a qualidade das receitas disponibilizadas, conforme ressalta Lopes (2004), que discute o crescimento da *Web* com a falta de cunho científico, com a apresentação de informações irrelevantes ou até equivocadas. Nesse contexto, verificam-se no cenário das receitas gastronômicas casos onde a receita disponibilizada por algum usuário não apresenta boa qualidade, seja por falta de se expor as informações necessárias para sua preparação, ou mesmo por conter ingredientes que destoam da realidade daquela receita, não representando, assim, uma forma adequada

[4]http://allrecipes.com

ou eficiente de se preparar um prato específico.

Verifica-se ainda nos *sites* de receitas atuais que, geralmente, quando efetua-se uma busca por alguma receita, são retornadas as receitas mais populares. Acontece também casos onde os *sites* priorizam receitas que foram expostas por eles mesmos, valorizando-as em detrimento a receitas disponibilizadas por usuários. Outro fato relevante, consiste nas receitas que são expostas quando se abre o *site* de receitas, uma vez que essas tendem a ser mais visualizadas do que as que não são expostas. Desta forma, pode-se observar que a maneira como são ordenadas as receitas pode ser ruim, tendo em vista que podem ser ordenações tendenciosas, bem como limitando o acesso às receitas que poderiam ter maior qualidade ou relevância para o usuário.

No entanto, apesar do que fora supracitado, as receitas apresentam informações textuais que podem ser exploradas, com o objetivo de extrair informações semânticas. Isso pode ser visto no título da receita, que contribui com a descoberta do nome do prato, além das sentenças onde estão contidos os ingredientes, suas respectivas quantidades e unidades de medida, bem como informações como o número de porções, votos, comentários, avaliação recebida, entre outros. Até mesmo o grau de dificuldade para se preparar uma receita pode ser estimado a partir do emprego de técnicas de linguística computacional sobre o modo de preparo, identificando termos que possam contribuir para a identificação do grau de dificuldade, como assar, fritar, misturar, picar, entre outros.

Nesse contexto, visualiza-se a necessidade de se efetuar tratamento e análise sobre os dados disponibilizados em uma receita, visando oferecer serviços que façam uso do conhecimento encontrado. Dessa forma, seria possível identificar qual receita melhor representaria um prato, ou ainda analisar possíveis alterações no preparo da receita de acordo com particularidades do usuário, como permitir a adição ou remoção de ingredientes que o usuário possui em casa. Ou ainda, permitir que o usuário adapte a receita conforme suas preferências culinárias, deixando evidente a ele também caso uma adaptação desejada seja considerada equivocada para o preparo do prato com sucesso, mostrando ao usuário que em outras receitas do mesmo prato tais alterações não foram realizadas, utilizando-se do conhecimento coletivo para enfatizar essa posição.

Como motivação deste trabalho, o Apêndice A apresenta algumas das funcionalidades pretendidas, apresentadas por meio de algumas figuras que compõem o protótipo inicial da aplicação e alguns requisitos funcionais.

1.3 Organização da Dissertação

Os próximos capítulos estão divididos da seguinte maneira: no Capítulo 2 é realizada a descrição dos trabalhos relacionados a este. No Capítulo 3 é apresentada a metodologia proposta para a descoberta de conhecimento em receitas gastronômicas. No Capítulo 4 apresenta-se uma caracterização da base de dados de receitas, além de estudos sobre o conhecimento descoberto com o uso da metodologia, como uma análise dos ingredientes descobertos, análise do processo de extração de ingrediente principal, análise das ações verbais identificadas nas receitas coletadas, bem como uma caracterização dos pratos identificados, como parte de um estudo de caso. No Capítulo 5 são apresentados os resultados experimentais das etapas da metodologia que contam com processos de descoberta de conhecimento. Finalmente, no Capítulo 6 são apresentadas as conclusões deste trabalho, bem como sugestões de trabalhos futuros.

Capítulo 2

Trabalhos Relacionados

Os trabalhos relacionados estão organizados da seguinte forma. Na Seção 2.1, são apresentados trabalhos no escopo de análise e caracterização de receitas que, em sua grande maioria, visam realizar a coleta em um determinado site de receitas e, com base nos dados coletados, efetuar análise dos mesmos, identificando a co-ocorrência de ingredientes, entre outros. Na Seção 2.2, são explorados exemplos de propostas e de construção de sistemas de recomendação de receitas, utilizando-se de diversas técnicas. Por fim, na Seção 2.3, são apresentados trabalhos que apresentam adaptações em receitas, analisando a viabilidade de extrair ou inserir um ingrediente em uma receita ou mesmo identificar possíveis ingredientes substitutos, mantendo a qualidade da receita.

2.1 Análise e caracterização de receitas gastronômicas

O trabalho de Ahn et al. (2011) consiste na criação de 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.

Ferreira et al. (2013) 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 coocorrência 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.

Os autores Yu et al. (2013) realizam coleta 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 Zhang et al. (2008) propõem o desenvolvimento de uma ferramenta de busca de receitas gastronômicas. De forma a prover o desenvolvimento da ferramenta, utilizaram-se de dados semi-estruturados de receitas e de um dicionário de ingredientes do *Wordnet*[1]. Possuem como principais características a criação de uma representação estruturada de dados de receitas, utilizando-se do *Wordnet* para estabelecer similaridades de ingredientes das receitas e por fim rotular as receitas com participação mínima do usuário. No trabalho, utilizam o conceito de níveis de ingredientes, podendo ter os níveis 1 a 3, onde o primeiro diz que o ingrediente é o principal encontrado na receita, o segundo diz que o ingrediente é importante na receita e o terceiro considera ingredientes aromáticos, comumente utilizados para temperos. Além dos ingredientes, são ainda levados em consideração outras informações sobre as receitas, como: nome, processo de preparo, quantidade e unidade de medida associadas a um ingrediente, entre outras. Para se buscar uma receita o usuário entra com um ingrediente principal e o sistema traz algumas receitas que são relacionadas ao ingrediente de entrada.

Este trabalho também aborda a caracterização e análise das receitas coletadas, tendo como objetivo, a partir das análises, ter conhecimento amplo sobre os dados de forma a possibilitar a manipulação dos mesmos. Uma das principais diferenças em relação aos demais trabalhos relacionados, refere-se à diversidade dos dados, uma vez que foram utilizadas receitas de diversas fontes de dados.

[1]http://wordnet.princeton.edu/

2.2 Sistemas de recomendação de receitas gastronômicas

Os autores Svensson et al. (2005) desenvolveram um trabalho que tinha como objetivo a realização de recomendação de receitas. Para esse trabalho, foram utilizados mais de 300 usuários, que foram entrevistados no intuito de obter suas opiniões sobre receitas. A recomendação foi realizada de acordo com a avaliação positiva das receitas.

Os autores Mino and Kobayashi (2009) propõem a recomendação de receitas na realização de uma dieta dos usuários. Para isso, eles inovam em seu trabalho no que tange em recomendar receitas, evidenciando as atividades que os usuários realizam no dia a dia, levando-se em consideração sua rotina, se praticam esportes, se somente trabalham. Utiliza-se da programação linear para efetuar a recomendação das receitas, onde há restrições de carboidratos, lipídios, proteínas e sal. Em contrapartida, é estimulado o consumo de vegetais.

O trabalho apresentado por Wagner et al. (2011) propõem a qualificação no procedimento de cozinhar e um sistema onde recomendam receitas que são mais saudáveis, mediante a forma de preparo e aos ingredientes utilizados. O processo onde buscam qualificar os usuários foi inicialmente elaborado com base em estudos na maneira de como os usuários cozinhavam, por meio de questionários aplicados a um grupo de usuários. Com base nesse estudo, e tendo como premissas cozinhar de uma maneira mais correta e o fato de que usuários mais experientes na cozinha podem geralmente, elaborar receitas mais saudáveis, deu-se o desenvolvimento do sistema de recomendação, que leva em consideração o nível de conhecimento do usuário em cozinhar para recomendar uma determinada receita, além de dar dicas aos usuários mais leigos de como preparar a receita da melhor maneira possível.

Ainda, seguindo a linha de sistemas de recomendação visando a alimentação saudável, Geleijnse et al. (2011) propõe o desenvolvimento de um sistema de recomendação de receitas onde o fator que possui maior importância é o quanto saudável é a receita. O sistema foi criado com base em um estudo realizado com usuários, onde foram levantados alguns requisitos importantes, como por exemplo, ter um controle sobre as refeições feitas durante a semana. O sistema proposto apresenta uma interface bem simples de uso levando em consideração o cardápio da última semana, apresentando informações que mostram o quanto saudavelmente o usuário se alimentou recentemente, e assim são apresentadas sugestões de pratos ao usuário que pode escolher por meio dos ingredientes presentes no prato, ou ainda pela quantidade de calorias, tempo de preparo, entre outros.

Os trabalhos apresentados por Ueda et al. (2011) e Ueda et al. (2014) também 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 diferenciam no que tange a pontuação que é dada ao ingrediente quanto à preferência do usuário. No primeiro trabalho (Ueda et al., 2011), foi verificado somente se o usuário gostava ou não de determinado ingrediente, sem levar em consideração a ordem de preferência dos ingredientes.

Já o segundo trabalho (Ueda et al., 2014) trata a questão do gosto do usuário, elaborando um *ranking* dos ingredientes preferidos. Posteriormente, o usuário entraria com informações sobre seu histórico alimentar dos últimos dias no sistema de recomendação, de maneira que o sistema recomendaria receitas que levariam em consideração as preferências do usuário, além de evitar recomendar pratos que foram elaborados recentemente.

Os autores Freyne and Berkovsky (2010) também propuseram o desenvolvimento de um sistema de recomendação de receitas; no entanto, o trabalho de Freyne and Berkovsky (2010) se difere dos anteriores no que se refere à maneira como propõem a recomendação, uma vez que utilizam de informações dos usuários em relação aos ingredientes, os quais evidenciam a pontuação dada pelo usuário para cada ingrediente e, com base nessas informações, é recomendada uma receita.

Os autores Trevisiol et al. (2014) propõem o desenvolvimento de algoritmos para recomendação de pratos e menus (refeições completas) em restaurantes. Para isso, utilizaram-se de dados coletados (comentários dos usuários) na plataforma Yelp[2], referente a 11.537 empresas, mais de 40 mil usuários, totalizando mais de 229 mil comentários. Na sequência utilizaram-se de técnicas de Processamento de Linguagem Natural (PLN) para extrair o sentimento de cada uma das sentenças, além de identificar o nome de pratos. Foram implementados três sistemas de recomendação onde no primeiro são retornados os itens alimentares mais frequentes que pertencem aos perfis de usuários e restaurantes. No segundo é feita uma filtragem colaborativa que estima a semelhança de gostos do usuário a fim de avaliar os pratos pertencentes aos restaurantes. Por fim, o terceiro utiliza-se dos menus frequentes e considerados bons por meio da análise de sentimento. Como resultados, visualizaram que usar a análise de sentimentos foi um fator positivo, uma vez que aumenta a chance de retornar um prato ou menu que satisfaça as preferências do usuário. Visualizaram ainda que os dois últimos algoritmos

[2]http://www.yelp.com.br/

se mostraram melhores, trazendo resultados considerados mais pertinentes ao desejo do usuário.

O presente trabalho apresenta relação com os trabalhos sobre a recomendação de receitas, uma vez que com a utilização do conhecimento coletivo descoberto por meio da metodologia e com a informação sobre a busca do usuário, são ofertadas receitas que visam atender às necessidades de informação específicas de um dado usuário.

2.3 Adaptação de receitas gastronômicas

O trabalho dos autores Teng et al. (2012) consiste inicialmente na realização de uma coleta de receitas. Em seguida trabalham com o conceito de redes de ingredientes, criando redes de ingredientes que co-ocorrem e rede de ingredientes substitutos. Os ingredientes substitutos foram identificados em comentários realizados por usuários nas receitas. Finalmente, dado um par de receitas, identificam qual das receitas possivelmente seria mais bem avaliada, levando-se em consideração uma técnica de predição. Os autores concluem que uma receita pode ser fortemente influenciada por alguns dos ingredientes que a compõem.

Os trabalhos de Bridge and Larkin (2014) e Larkin and Bridge (2014) tem como objetivo criar receitas de sanduíches e identificar ingredientes substitutos em receitas de sanduíches. O primeiro trabalho tenta criar receitas de sanduíches através de um sistema desenvolvido, estabelecendo uma comparação das receitas geradas pelo sistema e receitas geradas por humanos. Já no segundo trabalho, inicialmente, os autores apresentam um estudo sobre alguns métodos que propõem a substituição de ingredientes em receitas. Além dos métodos apresentados, efetua-se uma comparação com o método criado no primeiro trabalho. Para a identificação dos ingredientes e suas categorias, utiliza-se de um *Thesaurus* de ingredientes da língua inglesa. Utilizam-se ainda do domínio de receitas de sanduíches para efetuar os testes entre os métodos estudados e o proposto, com o intuito de identificar o melhor método. Ressalta-se que o método proposto no trabalho leva-se em consideração conjuntos de ingredientes frequentes em uma receita. Para isso, utilizam-se do algoritmo Apriori. Finalmente, os autores concluíram que o método proposto apresenta nível de aceitação similar aos já existentes para o domínio de receitas estudado. Para chegar a essa conclusão, utilizaram-se de questionários aplicados a usuários e por fim utilizou-se de um teste estatístico que mostra que as diferenças entre os métodos não são estatisticamente significativas.

Os autores Cojan et al. (2011) apresentam um trabalho cujo intuito principal consiste em criar adaptações em receitas gastronômicas. Estes autores já vêm trabalhando a algum tempo nesse domínio de assunto como pode ser visto em: Badra et al. (2008), Badra et al. (2009) e Blansché et al. (2010). Já neste trabalho, os autores utilizam-se dos conhecimentos adquiridos dos anteriores, entretanto eles propõem algumas diferenças na maneira de propor as adaptações de receitas, introduzindo os conceitos da ontologia *formal concept analysis* (FCA) que permite um agrupamento de ingredientes de acordo com propriedades comuns, por exemplo: ingredientes frescos, ralados, picados entre outras. Esta técnica tem como uma de suas principais características a vantagem de melhorar o processo de busca, refinando as possibilidades. Uma segunda técnica aplicada neste trabalho refere-se ao uso de regras de associação visando identificar possíveis ingredientes substitutos em receitas.

Os autores dos trabalhos Belda and Gamonar (2014) e Gamonar and Brasil (2015) propõem o desenvolvimento de uma rede social de receitas gastronômicas. Ressalta-se que os autores são da área da Comunicação Social e tem como objetivo apenas propor a criação da rede supracitada. Para isso, inicialmente, estudam programas de culinárias apresentados em TVs abertas e fechadas, bem como avaliam dois *sites* de receitas gastronômicas (Tudo Gostoso[3] e Cybercook[4]). Por fim, analisam três redes sociais difundidas atualmente que são utilizadas como ambiente de exposição de receitas. Após esse estudo, fazem a modelagem de produto utilizando-se de processos de desenvolvimento ágil de *software*, criando ainda um modelo de negócios para o produto. Importa salientar que os trabalhos de Belda and Gamonar (2014) e Gamonar and Brasil (2015) se diferem tanto dos demais trabalhos relacionados, quanto deste, uma vez que apresentam uma proposta de criação de uma rede social de receitas. Entretanto, assemelha-se a este no ponto em que se deseja utilizar receitas de diversas fontes de dados e até mesmo de mídias diferentes. Outro ponto interessante a ser analisado nestes trabalhos, refere-se ao fato de serem trabalhos cujos requisitos e especificações do sistema terem sido identificados por pesquisadores da área da Comunicação Social, o que reforça a importância de pesquisas referentes ao domínio do presente trabalho.

O presente trabalho fornece a possibilidade do usuário escolher uma das diversas receitas de um determinado prato. Além disso, propicia maior autonomia ao usuário na escolha dos ingredientes que devem estar presentes em uma receita mediante suas necessidades ou preferências. Este trabalho assemelha-se aos demais no que tange as opções

[3]http://www.tudogostoso.com.br
[4]http://www.cybercook.com.br

de variações de receitas. Nos trabalhos relacionados há a possibilidade de adaptação de uma receita por meio dos ingredientes que esta pode vir a ter. Já neste trabalho, a variação ocorre entre as diversas receitas que podem ser utilizadas para representar um dado prato.

Capítulo 3

Metodologia de descoberta de conhecimento em receitas gastronômicas

Este capítulo apresenta a metodologia de descoberta de conhecimento em receitas gastronômicas. A metodologia desenvolvida propõe a coleta de receitas gastronômicas a partir de diferentes fontes de dados. Uma vez tendo as receitas gastronômicas coletadas, vários processos são executados com o intuito de descobrir conhecimento sobre as receitas, bem como processos usados na descoberta de conhecimento sobre pratos, utilizando-se de informações sobre as receitas. Finalmente, possibilita-se a consulta de pratos, no intuito de retornar receitas relevantes, entre as várias opções disponíveis para a preparação do prato pesquisado.

A Figura 3.1 apresenta a metodologia proposta para descoberta de conhecimento sobre pratos e receitas gastronômicas, que será apresentada ao longo deste capítulo. Verifica-se a presença de quatro bases de dados. A primeira é a base de dados de receitas, que é responsável pelo armazenamento de informações sobre as receitas. Além dos dados coletados, essa base armazena informações sobre as receitas as quais são descobertas como parte da metodologia. Constata-se também a presença da base de dados de ingredientes, que por sua vez armazena informações sobre os ingredientes, obtidas com a utilização da metodologia.

Visualiza-se também na Figura 3.1 a base de dados de pratos, que de uma maneira similar às bases supracitadas, armazena conteúdos associados aos pratos. A principal

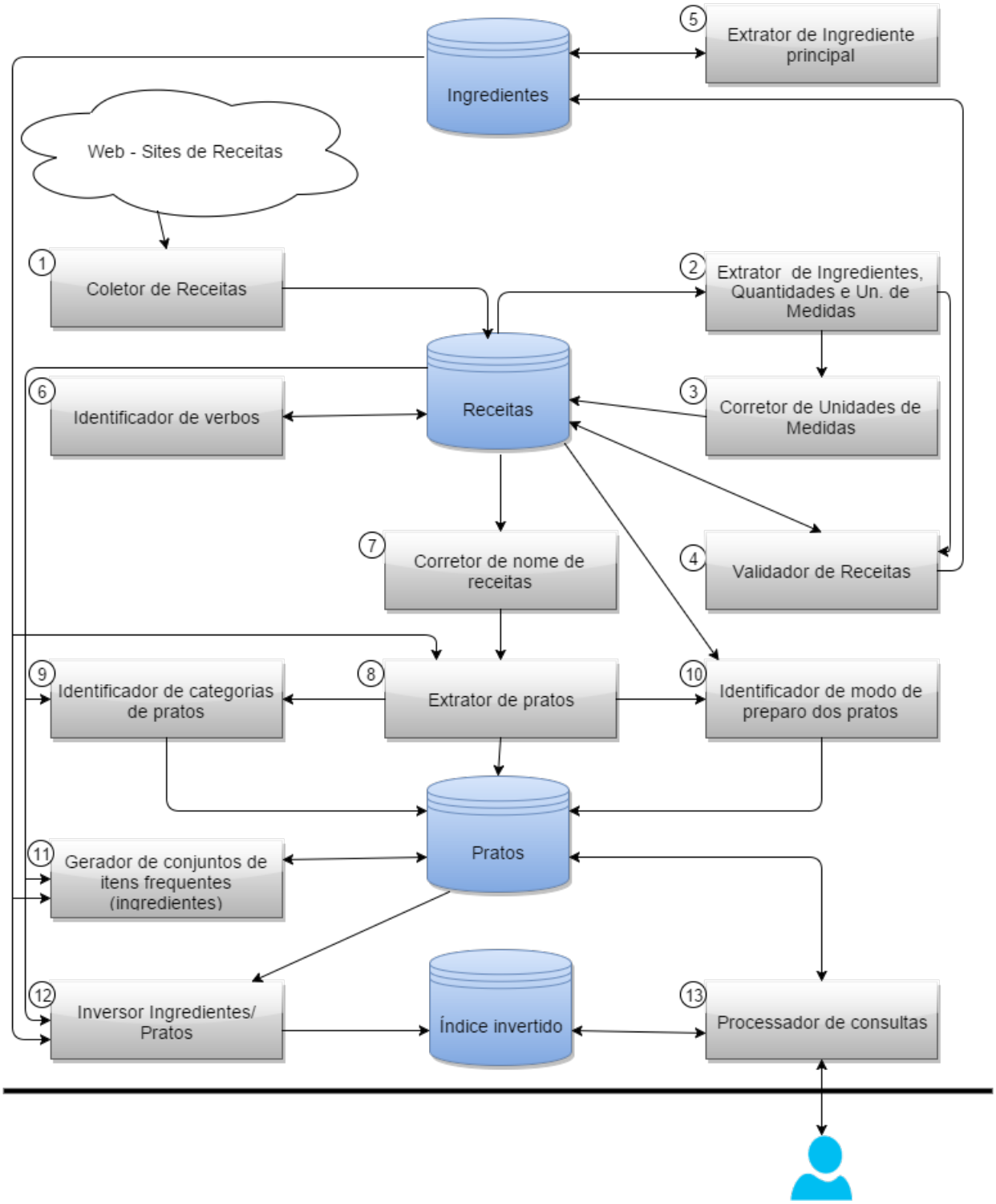


Figura 3.1: Arquitetura da metodologia de descoberta de conhecimento em receitas gastronômicas

diferença consiste no fato de que as informações acerca dos pratos não existem previamente. Assim, ressalta-se que todas as informações sobre os pratos originam-se de etapas da metodologia, onde a partir da descoberta de conhecimento sobre pratos, estas são descobertas e, posteriormente, persistidas. Por fim, verifica-se a base de dados de índice invertido. Essa base de dados, recebe dados sobre pratos e ingredientes, entretanto, os dados são armazenados na forma de um índice invertido de ingredientes, associando os respectivos pratos (e receitas) que contém cada ingrediente presente no índice (Baeza-Yates and Ribeiro-Neto, 2011). Similarmente ao que acontece com a base de dados de pratos, os dados aqui armazenados são provenientes de uma etapa da metodologia. A base de dados de índice invertido é criada com o objetivo de utilizá-la no momento da busca de receitas efetuada pelo usuário, uma vez que essa estrutura permite que os resultados sejam retornados em um menor espaço de tempo.

Conforme ressaltado anteriormente, a metodologia é dividida em algumas etapas, as quais serão apresentadas nas seções a seguir. A Seção 3.1 apresenta o coletor de receitas, que é responsável por coletar receitas gastronômicas em diversas fontes de dados, coletando informações como: ingredientes, instruções de preparo, avaliação da receita, entre outras, etapa 1 da figura. A Seção 3.2 apresenta o extrator de ingredientes, quantidades e unidades de medida, etapa 2 da figura. A Seção 3.3 apresenta o corretor de unidades de medida, que é responsável por estabelecer uma associação entre termos que indicam uma mesma unidade de medida, etapa 3 da figura.

Por sua vez, a Seção 3.4 apresenta o validador de receitas, onde para cada receita coletada, valida-se a receita por meio dos resultados retornados pelo extrator. Essa é a etapa 4 da figura. A Seção 3.5 apresenta o extrator de ingrediente principal, responsável por identificar entre os termos do ingrediente, qual o ingrediente principal, etapa 5 da figura. A Seção 3.6 apresenta o identificador de verbos, que identifica os verbos existentes nas instruções de preparo em cada uma das receitas, etapa 6 da figura.

Já a Seção 3.7 apresenta o corretor de nome de receita, que busca a padronização e correção ortográfica dos nomes das receitas, etapa 7 da figura. A Seção 3.8 apresenta o extrator de pratos, que é responsável por identificar, para cada receita, a qual(is) prato(s) a mesma se refere, etapa 8 da figura. A Seção 3.9 apresenta o identificador de categorias de pratos, que estabelece qual(is) categoria(s) melhor representa(m) um determinado prato, etapa 9 da figura. A Seção 3.10 apresenta o identificador de modo de preparo dos pratos, que identifica qual(is) modo(s) de preparo existente(s) para um determinado prato, bem como o modo de preparo predominante do prato, etapa 10 da figura.

Finalizando o processo apresentado pela metodologia, a Seção 3.11 apresenta o gerador de conjuntos de itens frequentes (ingredientes), que identifica os ingredientes que ocorrem frequentemente em receitas de um determinado prato, etapa 11 da figura. A Seção 3.12 apresenta o inversor de pratos/ingredientes, onde se indexa os registros de ingredientes, criando um índice invertido dos ingredientes em relação às receitas associadas a cada um dos pratos, etapa 12 da figura. Por fim, a Seção 3.13 apresenta o processador de consultas, que é responsável pela geração e apresentação dos resultados, etapa 13 da figura.

3.1 Coletor de receitas

O presente trabalho utiliza-se de diversas receitas gastronômicas disponibilizadas no ambiente *Web*. De forma a viabilizar o uso destas receitas, fez-se necessário o uso de um coletor. Todas as informações a respeito das fontes de dados escolhidas, bem como sobre a preparação dos dados coletados, são apresentadas nas Seções 3.1.1 e 3.1.2. Conforme pode ser visualizado na Figura 3.1, o coletor estabelece ligação com a *Web* para acessar *sites* de receitas, de onde são coletadas as receitas que em seguida são armazenadas na base de dados de receitas.

3.1.1 Serviços escolhidos como fontes de dados

Este trabalho apresenta receitas coletadas de cinco fontes de dados diferentes. 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. A seguir são apresentadas as fontes de dados escolhidas:

- Tudo Gostoso[1]: é um *site* brasileiro criado em 2005 que apresenta diversas receitas compartilhadas por usuários do *site*.

- Receitas.com[2]: receitas disponibilizadas por meio do *site* Globo.com, contendo receitas gastronômicas que são apresentadas em programas da emissora Globo,

[1]http://www.tudogostoso.com.br
[2]http://www.gshow.globo.com/receitas

como Mais Você e Estrelas. O *site* também permite o envio de receitas por usuários cadastrados que compartilham seus conhecimentos culinários.

- Edu Guedes[3]: mais um *site* onde se verifica apelo televisivo, apresentando receitas de um programa de culinária da rede de televisão Record. As receitas disponibilizadas são apresentadas unicamente pelo *chef* Edu Guedes.
- Cybercook[4]: *site* onde são apresentadas diversas 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[5]: *site* que apresenta receitas culinárias referentes a uma dieta chamada Dukan[6]. Todas as receitas presentes no *site* são especiais para dietas e disponibilizadas por usuários.

Para o processo de coleta, utilizou-se o *Crawler4j*[7], coletor que possui código aberto, desenvolvido em Java e que fornece uma interface simples para a tarefa de coleta na *Web*. Para a utilização do coletor em questão foi necessário analisar a estrutura das páginas a serem coletadas, de maneira a 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 *XML*, uma vez que esse formato oferece facilidades no manuseio dos dados. Ressalta-se que foi desenvolvida uma aplicação específica para efetuar a coleta de cada uma das fontes de dados.

3.1.2 Preparação dos dados coletados

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

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

liação média), data de postagem, descrição, tipo de cozinha (brasileira, italiana, entre outras), categoria (lanches, bebidas, bolos e tortas, entre outras), 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 se observa no serviço Receitas.com, onde há o número de pessoas que "favoritaram" determinada receita. O serviço Edu Guedes também apresenta algumas especificidades, como o numero de "gostei" e de visualizações do vídeo que segue a receita. Já o serviço Dieta e Receitas apresenta campos como tempo de cozimento, tempo de espera, quantidade de calorias, grau de dificuldade, quantidade tolerada 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 dados originais, da forma como coletados, foram mantidos para que depois possam ser usados como resultado a ser apresentado ao usuário.

3.2 Extrator de Ingredientes, Quantidades e Unidades de Medidas

As receitas coletadas apresentam as informações sobre os ingredientes em uma frase, aqui denominada simplesmente como sentença. Assim, para cada sentença, há a descrição do ingrediente, bem como informações referentes à sua quantidade e unidade de medida, quando estas existirem. Um exemplo de uma sentença pode ser visualizado na expressão "3 colheres de chá de açúcar".

Uma etapa fundamental da metodologia consiste da extração das informações encontradas em cada uma das sentenças. Para isso, deu-se o desenvolvimento de uma heurística para efetuar a extração dos ingredientes e suas quantidades e unidades de medida. Observa-se na Figura 3.1 que o extrator, etapa 2 da figura, recebe como entrada cada uma das sentenças presentes nas receitas coletadas. A heurística é executada duas vezes. A primeira execução tem a função de extrair um conjunto de unidades de medidas candidatas, que são repassadas para a etapa 3 da metodologia para validação das unidades de medida (Seção 3.3). Já a segunda execução tem a função de extrair os ingredientes, quantidades e unidades de medida, para que as receitas sejam validadas na

etapa 4 da metodologia (Seção 3.4).

O Algoritmo 3.1 apresenta como se dá a extração dos ingredientes e suas quantidades e unidades de medida. O algoritmo recebe como entrada uma dada sentença s e tem como saída os dados extraídos (ingrediente, quantidade e unidade de medida) para a sentença.

Algoritmo 3.1: **- Extrai_Ing_Quant_Med** - Algoritmo da heurística para identificar ingredientes, quantidades e unidades de medida.

```
   Entrada: s
   Saída: Res[ing, quant, med]
 1 início
 2 |  Res[ing, quant, med]
 3 |  se 1º termo de s for um caractere especial (-,*,.) então
 4 |  |  remove_carac_esp(s)
 5 |  se s ⊃ "expressões especiais" então
 6 |  |  resolve_exp_esp(s, Res[ing, quant, med])
 7 |  senão
 8 |  |  cont ← conta_de(s)
 9 |  |  se cont = 0 então
10 |  |  |  resolve_sem_de (s, Res[ing, quant, med])
11 |  |  senão se cont = 1 então
12 |  |  |  resolve_1_de (s, Res[ing, quant, med])
13 |  |  senão
14 |  |  |  resolve_2_de (s, Res[ing, quant, med])
15 |  |  |
16 |  |
17 fim
```

Para que o algoritmo fosse desenvolvido, primeiramente, foram realizadas algumas análises em diversas receitas coletadas, com o intuito de encontrar alguns padrões que acometiam com certa frequência. Ressalta-se que este procedimento de identificação dos padrões se deu simultaneamente ao desenvolvimento do algoritmo.

Inicialmente, o algoritmo verifica se o primeiro termo da sentença s representa um dos seguintes caracteres especiais: "-","*" ou ".". Em caso afirmativo, chama-se a função *remove_carac_esp*(), passando a sentença s como parâmetro. Em seguida, verifica-se se há a presença de uma das expressões especiais em s. A Tabela 3.1 apresenta as expressões especiais utilizadas no desenvolvimento da heurística. Se s apresentar uma das expressões especiais apresentadas na Tabela 3.1, então chama-se a função que resolve as sentenças que apresentam tais expressões, passando como parâmetro s e o vetor

de resultados (*Res*), onde serão armazenadas as saídas para ingrediente, quantidade e unidade de medida. Em caso negativo (não apresenta expressões especiais em *s*), então *cont* armazena o número de preposições "de" em *s*. Em seguida, verifica-se se o valor de *cont* é igual a 0. Em caso afirmativo, chama-se a função *resolve_sem_de*(), passando como parâmetro *s* e *Res*. Caso o valor de *cont* seja igual a 1, então chama-se a função *resolve_1_de*(). Finalmente, se o valor de *cont* for maior do que 1, então, chama-se a função *resolve_2_de*().

Tabela 3.1: Expressões especiais que são utilizadas no desenvolvimento da heurística.

Expressões especiais
quanto baste de
tempero a vontade
a gosto
a seu gosto
para untar
para polvilhar
para enrolar
para decorar

Ressalta-se que o processo de extração de ingredientes, quantidades e unidades de medida acontece em cada uma das funções, com exceção da primeira função que apenas remove o caractere especial encontrado (*remove_carac_esp*()) e, em seguida, passa o restante da sentença *s* adiante, para ser tratada em alguma outra parte do algoritmo. A Tabela 3.2 apresenta algumas sentenças e como elas são tratadas pelas funções presentes no Algoritmo 3.1. A tabela ilustra que para a função *remove_carac_esp*() nenhuma informação é extraída para ingrediente, quantidade e unidade de medida. A função inicialmente tenta identificar a presença de um dos caracteres especiais e, se encontrar, quebra a *string* da sentença *s* no primeiro espaço e considera-se apenas a parte da direita da *string* inicial. Na sequência, a sentença *s* sem o caractere especial passa por uma das funções destinadas à extração das informações.

A primeira função responsável pela extração de ingrediente, quantidade e unidade de medida é a *resolve_exp_esp*(). Pode-se verificar que as sentenças que são resolvidas por esta função, ilustradas na Tabela 3.2, apresentam apenas o ingrediente, não apresentando

Tabela 3.2: Alguns dos padrões existentes e resolvidos pelas funções do algoritmo.

Funções	Exemplos de Sentença	Classificação		
		Ingrediente	Quantidade	Un. Medida
remove_carac_esp()	* 1 beterraba picada	-	-	-
	- 1/2 xicara de leite	-	-	-
	. 2 kg de carne	-	-	-
resolve_exp_esp()	quanto baste de sal	sal	-	-
	alho a gosto	alho	-	-
	cebolinha a vontade	cebolinha	-	-
resolve_sem_de()	1 tomate picado	tomate picado	1	-
	2 cebolas picadas	cebolas picadas	2	-
	1 cenoura ralada	cenoura ralada	1	-
resolve_1_de()	1 e 1/2 copo de leite	leite	1 1/2	copo
	1/2 lata de milho	milho	1/2	lata
	1 xicara e 1/2 de oleo	oleo	1 1/2	xicara
resolve_2_de()	1 colher de cha de cafe	cafe	1	colher de cha
	2 colheres de sopa de mel	mel	2	colheres de sopa
	1 1/2 colher de cha de agua	agua	1 1/2	colher de cha

nem quantidade e nem a unidade de medida. A maneira como é feita a extração para as sentenças difere em alguns casos, uma vez que a apresentação das sentenças ocorre de maneira diferente. Por exemplo, nas sentenças resolvidas pela função *resolve_exp_esp*() ilustradas na Tabela 3.2, verifica-se que na primeira sentença tem-se inicialmente a expressão "quanto baste de" e na sequência vem a ocorrência do ingrediente. Já para as outras duas sentenças, verifica-se primeiramente a presença do ingrediente, seguido por uma expressão especial. Desta forma, cada sentença com expressões especiais distintas tem a sua maneira de ser resolvida.

Se uma sentença não apresentar uma das expressões especiais, ela obrigatoriamente será resolvida por meio de uma das três funções restantes (*resolve_sem_de*(), *resolve_1_de*() e *resolve_2_de*()). A diferença de padrão nessas funções é verificada mediante a frequência de ocorrência da preposição "de". A primeira função não apresenta em sua sentença a ocorrência da preposição supracitada. Já a segunda função tem uma ocorrência e por fim, na terceira função, a preposição ocorre mais de uma vez, conforme pode-se verificar nas ilustrações da Tabela 3.2.

Observa-se que as sentenças resolvidas pela função *resolve_sem_de*() não apresentam unidades de medida. Assim, para uma sentença com este padrão, quebra-se a *string* da sentença no primeiro espaço, de forma que as informações acerca da quantidade e ingrediente sejam recuperadas.

Quando uma sentença apresentar uma única ocorrência da preposição "de", esta será resolvida pela função *resolve_1_de*(). As sentenças resolvidas por esta função, conforme pode ser visto na Tabela 3.2, apresentam ingrediente, quantidade e unidade de medida. Verifica-se que há um padrão: primeiramente há a ocorrência da quantidade, seguida pela unidade de medida e por fim pelo ingrediente. Assim, através de tratamento de *strings*, quebra-se a sentença no primeiro espaço, obtendo a quantidade, e na preposição "de", recuperando as informações referentes à unidade de medida e ao ingrediente.

Normalmente, verifica-se nas sentenças a presença da quantidade antes do primeiro espaço, conforme explicado anteriormente. Entretanto, a Tabela 3.2 ilustra que para a primeira e terceira sentenças resolvidas pela função *resolve_1_de*(), há uma diferenciação em relação à maneira como são expostas as quantidades. Nestes casos, verificam-se dois padrões diferentes: no primeiro, a quantidade vem antes da unidade de medida, no entanto, a quantidade é composta pelo conectivo "e" entre os valores. Assim, quebra-se a *string* no primeiro espaço e verifica-se se a parte do lado direito do espaço apresenta a barra ('/'), que representa a ocorrência de uma quantidade em forma fracionada. Caso tenha, faz-se outra verificação: quebra-se a *string* que contém a barra no primeiro espaço e, posteriormente, verifica-se se o lado esquerdo possui o conectivo "e". Caso apresente, quebra-se a *string* do lado direito novamente no primeiro espaço e em seguida faz-se a junção do valor que compõe a quantidade.

Já o segundo padrão da função *resolve_1_de*() também apresenta quantidade composta por dois valores, entretanto, aqui estes encontram-se separados pela unidade de medida. Assim, inicialmente, quebra-se a *string* no primeiro espaço, depois toma-se a parte da direita e mais uma vez realiza uma quebra no primeiro espaço. Nesse momento, verifica-se se há a presença da barra ('/') na parte da direita; em caso afirmativo, quebra-se mais uma vez no primeiro espaço e, assim, concatena-se os valores referentes à quantidade e recuperam-se os valores da unidade de medida e do ingrediente.

Finalmente, a função *resolve_2_de*() resolve as sentenças que apresentam ocorrência da preposição "de" mais de uma vez, como é ilustrado na Tabela 3.2. Similarmente às sentenças resolvidas pela função *resolve_1_de*(), as sentenças resolvidas aqui também apresentam um padrão onde visualiza-se a quantidade seguida pela unidade de medida

e por fim o ingrediente. Entretanto, aqui há uma diferença em relação à unidade de medida. Percebe-se que uma das ocorrências da preposição "de" ocorre de forma a unir dois termos que compõem uma unidade de medida; e após a segunda ocorrência da preposição, encontra-se o ingrediente. Assim, quebra-se a *string* no primeiro espaço, obtendo a quantidade, e na segunda ocorrência da preposição "de", obtendo a ´parte referente à unidade de medida e ao ingrediente.

Similarmente aos casos apresentados anteriormente, onde as quantidades são compostas, para representar valores fracionados, percebe-se na Tabela 3.2 que a terceira sentença de exemplo resolvida pela função $resolve\_2\_de()$ também apresenta quantidade composta por dois valores, sendo que estes são separados por um espaço. Nesse caso, para extrair as informações, primeiramente quebra-se a *string* no primeiro espaço, em seguida verifica-se se há a presença da barra ('/') na outra parte da *string*. Caso tenha, quebra-se essa *string* no primeiro espaço e logo concatenam-se as duas partes que constituem a quantidade. Ressalta-se que esses padrões relacionados à quantidade podem ocorrer nas sentenças que são resolvidas pelas três funções principais da heurística: $resolve\_sem\_de()$, $resolve\_1\_de()$ e $resolve\_2\_de()$.

3.3 Classificador de unidades de medida

Com o desenvolvimento da heurística utilizada na etapa 2 da metodologia, por meio dos padrões identificados, verificou-se que há inúmeras maneiras de nomear uma unidade de medida. O resultado retornado pela heurística para a unidade de medida será chamado aqui de candidato a unidade de medida, uma vez que estes candidatos podem não ser uma unidade de medida válida. Outra questão a ser levantada sobre os candidatos a unidades de medida gira em torno da possibilidade de diversos candidatos se referirem a uma mesma unidade de medida. Ressalta-se que um candidato validado a unidade de medida neste trabalho poderá se tornar um apelido de unidade de medida ou nome principal de unidade de medida. A relação entre apelido de unidade de medida e nome principal de unidade de medida é verificada quando vários apelidos de unidade de medida referem-se a um mesmo nome principal de unidade de medida, conforme é ilustrado na Tabela 3.3.

Esta etapa da metodologia recebe como entrada os candidatos a unidades de medida extraídos por meio da heurística utilizada no extrator de ingredientes, quantidades e

Tabela 3.3: Exemplo de associação entre os apelidos de unidades de medida e o nome principal de unidade de medida.

Apelidos de Un. Med.	Nome principal de Un. Med.
colher de sopa	colher de sopa
colher sopa	colher de sopa
colher de (sopa)	colher de sopa
xic. chá	xícara de chá
xícara de chá	xícara de chá
chicara de chá	xícara de chá

unidades de medida, etapa 2 da metodologia. Como saída do processo, tem-se uma base de apelidos de unidades de medida e uma base de nomes principais de unidades de medida, os quais, na sequência, são inseridos na base de dados de receitas.

O Algoritmo 3.2 apresenta o procedimento utilizado na validação dos candidatos a unidades de medida, onde R é a entrada, que consiste no conjunto das receitas coletadas, S é o conjunto de sentenças em uma dada receita e M é a lista de candidatos a unidades de medida. M é uma estrutura de dados do tipo dicionário, onde para cada candidato a unidade de medida tem-se associada a sua frequência. Como saída do algoritmo, tem-se as unidades de medida validadas (podendo ser divididas em apelidos de unidade de medida AP e nome principal de unidade de medida NP).

O Algoritmo 3.2 inicialmente entra em um comando de repetição para ler cada receita r presente na lista de receitas R, onde para cada sentença s presente no conjunto de sentenças S de uma receita, chama o Algoritmo 3.1 (etapa 2 da metodologia), onde são extraídas de cada sentença s os dados referentes ao ingrediente, quantidade e unidade de medida. Em seguida, armazena-se na lista de candidatos a unidade de medida M, os dados extraídos para a unidade de medida ($Res[med]$). Cabe salientar que, nessa etapa da metodologia, as informações de ingrediente e quantidade não são utilizadas.

Na sequência, para cada candidato à unidade de medida m presente na lista de candidatos M, verifica-se se sua frequência de ocorrência é maior do que o limiar estabelecido l que no caso do presente estudo foi definido como $l = 10$.

Em seguida, ocorre a etapa de validação manual dos candidatos por um especialista, que decide se o candidato é um nome principal de unidade de medida ou um apelido de unidade de medida. Salienta-se que quando um candidato é classificado como um apelido

Algoritmo 3.2: Algoritmo de validação de unidades de medida.

Entrada: R
Saída: $NP \leftarrow$ base de dados de nomes principais de unidades de medida
$AP \leftarrow$ base de dados de apelidos de unidades de medida

```
1  início
2      para cada r ∈ R faça
3          para cada s ∈ S faça
4              Res[ing, quant, med] ← Extrai_Ing_Quant_Med(s)
5              M ← Res[med]
6      para cada m ∈ M faça
7          se frequência de m > l então
8              se valida_med(m) então
9                  se m ∈ NP então
10                     insere_NP(m)
11                 senão
12                     insere_AP(m)
13
14
15
16 fim
```

de unidade de medida, ele também é manualmente associado a um nome principal de unidade de medida, de forma a associar cada apelido a um nome principal. Em seguida as unidades de medida validadas (apelidos de unidades de medida e nome principal de unidade de medida) são armazenadas na base de dados de receitas por meio das estruturas NP e AP.

Pode-se dizer que o Algoritmo 3.2 conseguiu executar suas funções esperadas, identificando os candidatos válidos de unidades de medida e na sequência com a detecção manual, estabeleceu-se a classificação de um candidato válido em apelidos de unidades de medida ou nome principal de unidade de medida. Ao final do processo, verificava-se a presença de 170 nomes principais de unidades de medida e 5.482 apelidos de unidades de medida, para as bases de dados utilizadas neste trabalho.

3.4 Validador de Receitas

Com o extrator de ingredientes, quantidades e unidades de medida, viu-se a necessidade de criar um mecanismo de validação das receitas. Este processo de validação tem por

objetivo evitar que dados ruidosos afetem na qualidade da metodologia proposta. A etapa de validação das receitas recebe como entrada todos os resultados do extrator (etapa 2 da metodologia) e ainda informações sobre unidades de medidas validadas (etapa 3 da metodologia), que encontram-se armazenadas na base de dados de receitas. Após o processo de validação, caso a receita seja validada, são armazenados na base de dados de ingredientes os dados referentes aos ingredientes (ingrediente extraído) e são armazenados na base de dados de receitas os dados acerca das receitas (ingrediente, quantidade e unidade de medida extraídos, além dos demais dados de receitas, como nome, fonte onde foi coletada, entre outros). As receitas não validadas têm os seus dados descartados.

O funcionamento do validador de receitas pode ser visualizado por meio do Algoritmo 3.3, que recebe como entrada R, que representa a lista de receitas e AP, que consiste na lista de apelidos de unidades de medida retornada na etapa 3 da metodologia. Como saída do algoritmo tem-se as receitas validadas RV, que têm seus dados armazenados na sequência nas bases de dados de receitas e de ingredientes.

Algoritmo 3.3: Algoritmo de validação de receitas.

```
   Entrada: R, AP
   Saída: RV
 1 início
 2     para cada r ∈ R faça
 3         erro ← 0
 4         para cada s ∈ S faça
 5             Res[ing, med, quan] ← Extrai_Ing_Quant_Med(s)
 6             se med ⊄ AP então
 7                 descarta_receita(r)
 8                 erro ← 1;
 9                 break
10             senão
11                 I ← Res[ing]
12                 Q ← Res[quant]
13                 M ← Res[med]
14
15         se erro = 0 então
16             insere_receita(r, I, Q, M, RV)
17             insere_ing(I, RV)
18
19 fim
```

O Algoritmo 3.3 inicialmente entra em um comando de repetição para cada receita r presente na lista de receitas R, onde para cada sentença s presente no conjunto de sentenças S de uma receita, chama o Algoritmo 3.1 (etapa 2 da metodologia), onde são extraídos de cada sentença os dados referentes ao ingrediente, quantidade e unidade de medida. Em seguida, se o dado extraído para a unidade de medida não estiver presente na lista de apelidos de unidade de medida (AP), a receita analisada é descartada, na sequência a variável *erro* recebe o valor 1 e, é dado um comando (*break*) para sair do comando de repetição. Caso contrário, armazenam-se os dados retornados pelo extrator (ingrediente, quantidade e unidade de medida) para uma dada sentença.

Ressalta-se que estes dados são armazenados para todas as sentenças de uma receita onde a unidade de medida extraída pelo extrator foi validada. Posteriormente, verifica-se se o valor presente na variável *erro* é igual a 0. Em caso afirmativo, armazenam-se os dados referentes às receitas na base de dados de receitas (ingrediente I, quantidade Q e unidade de medida M extraídos, além dos demais dados de receitas, como nome e fonte onde foi coletada, entre outros), bem como armazena-se os dados de ingredientes na base de dados de ingredientes (ingrediente extraído I).

Com a utilização do validador de receitas obtêm-se dados mais puros, impedindo que dados ruidosos afetem na qualidade da base de dados. Salienta-se que para isso, foram utilizados apenas dos candidatos a unidades de medidas extraídos pelo extrator de ingredientes, quantidades e unidades de medida. Somente receitas que em todas as suas sentenças tiveram os candidatos a unidades de medida validados, foram utilizadas.

3.5 Identificador de ingrediente principal

Com a execução da heurística de extração de ingredientes, quantidades e unidades de medida, obtém-se os ingredientes. Entretanto, é classificado como ingrediente qualquer informação que venha posteriormente à unidade de medida. A Figura 3.2 apresenta os resultados da heurística para uma sentença como ilustração.

Entretanto, com a utilização da heurística, o ingrediente apresenta informações adicionais que ajudam no entendimento de como utilizá-lo, não apresentando apenas o ingrediente, uma vez que, além do ingrediente (cebola), apresenta-se também uma informação adicional (picadinha). Nesse ponto, faz-se necessário identificar de fato o ingrediente, aqui chamado de ingrediente principal. Para isso, desenvolveu-se o "Extrator de in-

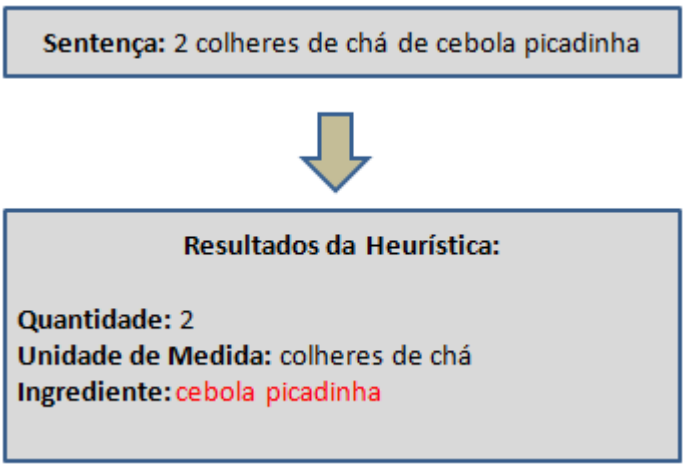


Figura 3.2: Ilustração de um resultado da heurística para uma dada sentença.

grediente principal", que recebe como entrada o nome de um ingrediente classificado pela heurística, aqui chamado de sentença de ingrediente, e retorna o ingrediente principal associando-o à sentença de ingrediente. Ambas as informações são armazenadas na base de dados de ingredientes, como pode ser verificado na arquitetura apresentada pela Figura 3.1.

O extrator de ingrediente principal possui três fases. Na primeira fase, consideram-se todas as sentenças de ingredientes distintas que possuem apenas um termo como um ingrediente principal. Na segunda fase, identificam-se manualmente os ingredientes principais nas sentenças de ingredientes levando-se em consideração a frequência de ocorrência das sentenças de ingredientes. Finalmente, a terceira fase utiliza-se dos ingredientes principais identificados nas fases anteriores, de forma a tentar associá-los às sentenças de ingredientes até então não associadas a um ingrediente principal.

O processo de identificação de ingrediente principal pode ser visualizado por meio do Algoritmo 3.4, que tem como entrada SI, que consiste na lista de sentenças únicas de ingredientes, contendo também a frequência de cada uma das sentenças. Como saída do algoritmo tem-se os ingredientes principais IP identificados, de cada sentença. O algoritmo apresenta a divisão por fases do processo de identificação de ingredientes principais.

Na primeira fase do processo de identificação de ingrediente principal, para cada uma das sentenças de ingredientes $si \in SI$, inicialmente armazena-se em *cont* o número de termos de si. Em seguida, se *cont* for igual a 1, ou seja, se a sentença de ingrediente possuir apenas um termo, considera-se esta sentença de ingrediente como um ingrediente

Algoritmo 3.4: Algoritmo de identificação de ingrediente principal.

```
   Entrada: SI: lista de sentenças de ingredientes únicas
   Saída: IP: lista de ingredientes principais
 1 início
 2     ——fase 1 da identificação de ingredientes principais——
 3     para cada si ∈ SI faça
 4         cont ← conta_numero_termos(si)
 5         se cont = 1 então
 6             IP ← si
 7             SI ← SI − si
 8
 9     ——fase 2 da identificação de ingredientes principais——
10     para cada si ∈ SI faça
11         se si.freq ≥ l então
12             ingp ← especialista_identifica_ingp(si)
13             se ingp ⊂ IP então
14                 IP(ingp).freq+ = 1
15             senão
16                 IP ← ingp
17
18         SI ← Si − si
19     ——fase 3 da identificação de ingredientes principais——
20     para cada si ∈ SI faça
21         k ← 4
22         repita
23             NG ← extrai_ngrama(si, k)
24             para cada ng ∈ NG faça
25                 se ng ⊂ IP então
26                     IP(ing).freq+ = 1
27                     k ← 0
28                     break
29
30         até k > 0 ;
31 fim
```

principal. Desta forma, cria-se uma nova instância de ingrediente principal para cada uma das sentenças de ingredientes distintas que apresentam um único termo.

Já na segunda fase do processo de identificação de ingrediente principal, conforme visualiza-se no Algoritmo 3.4, busca-se identificar os ingredientes principais que possuem mais de um termo nas sentenças de ingredientes, levando-se em consideração a frequência

de ocorrência das sentenças de ingredientes. Por exemplo, constatou-se que a sentença de ingrediente "leite condensado" ocorreu em aproximadamente 50 mil sentenças; em contrapartida, verificou-se que a sentença de ingrediente "zimbro moído" ocorreu apenas uma vez. Nesta fase, sentenças de ingredientes que ocorreram com uma frequência maior que um dado limiar l (no caso deste trabalho, $l = 50$), são passadas para o especialista identificar o ingrediente principal, de forma manual. Se o ingrediente principal encontrado pelo especialista já estiver presente em IP, então sua frequência é incrementada em IP. Caso contrário, o novo ingrediente principal é inserido em IP.

Finalmente, a terceira e última fase do processo de identificação de ingredientes principais consiste na utilização dos ingredientes principais já identificados nas fases anteriores. O processo trabalha com o conceito de janelamento, de forma a permitir que cada parte da *string* da sentença de ingrediente seja buscada em IP, na tentativa de encontrar um ingrediente principal em alguma parte da sentença de ingrediente.

A fase 3 do Algoritmo 3.4 inicia com a variável k sendo inicializada com o valor $k = 4$, que é o maior tamanho de janela possível, já que não há ingredientes principais em IP com mais do que quatro termos. A função $extrai\_ngrama()$ retorna em NG a lista de n-gramas de tamanho k presentes na senteça si. Caso um n-grama seja encontrado em IP, nenhum outro n-grama de si será analisado, uma vez que o ingrediente principal de si já foi encontrado.

A Figura 3.3 apresenta uma ilustração do processo de janelamento presente na terceira fase do Algoritmo 3.4. Visualiza-se na figura que, inicialmente, tenta-se identificar na sentença de ingrediente, um ingrediente principal com número de termos igual a quatro. Esse procedimento foi efetuado para todos os n-gramas de tamanho 4 "cebola roxa cortada em" e "roxa cortada em tiras". No exemplo dado, o número de possibilidades foi igual a duas. Uma vez não encontrado um ingrediente principal na sentença de ingrediente, realizou-se a mesma verificação, agora com um janelamento de tamanho três. Similarmente ao que fora feito anteriormente, para cada três termos da sentença, verifica-se se o n-grama é um ingrediente principal. O janelamento de três termos possibilitou três verificações diferentes.

Com o janelamento de tamanho três, também não foi possível identificar um ingrediente principal. Assim, passou-se para o janelamento de tamanho dois. Nesse momento, logo na primeira verificação, identificou-se o ingrediente principal "cebola roxa". Caso não fosse encontrado nenhum ingrediente principal com o janelamento de tamanho dois, seriam analisados individualmente cada um dos termos da sentença de ingrediente, vi-

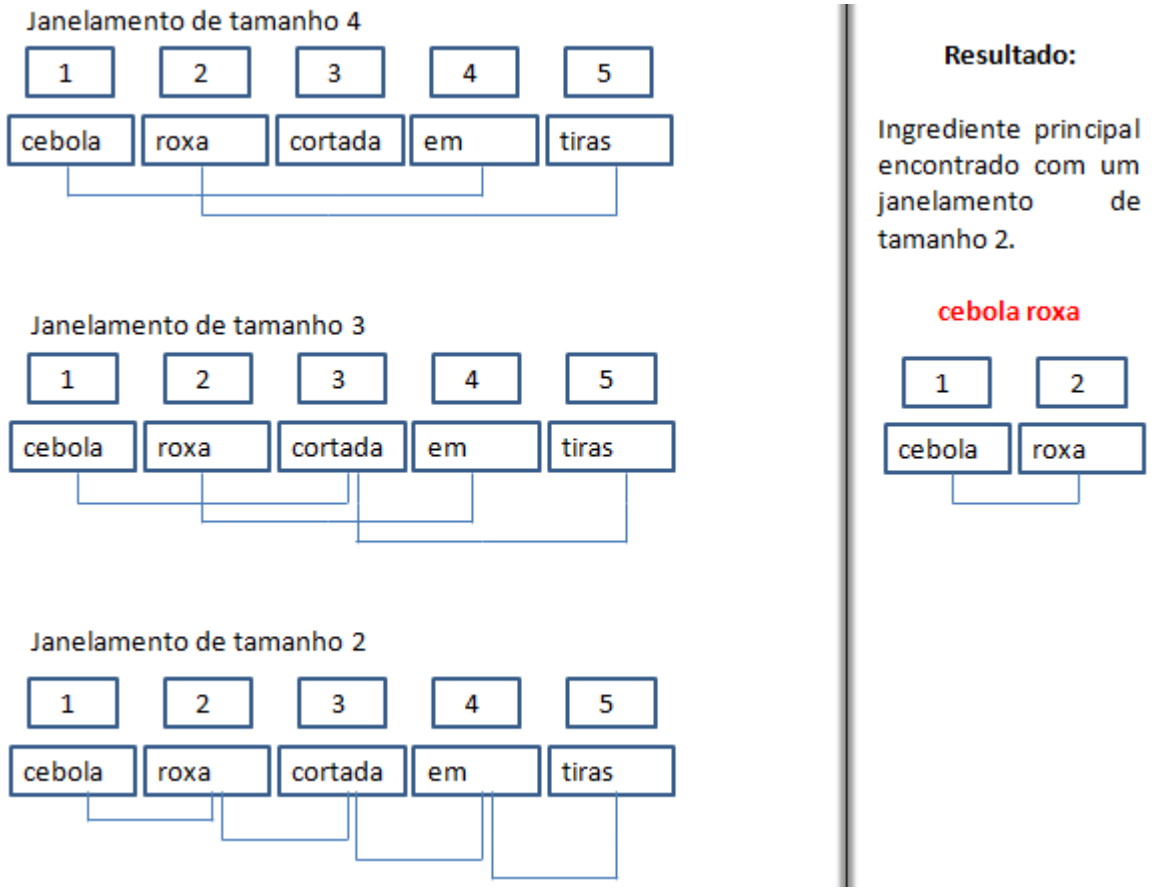


Figura 3.3: Exemplo de janelamento.

sando analisar se algum dos termos representaria um ingrediente principal. Ressalta-se que este processo foi realizado para todas as sentenças de ingredientes que até então não possuíam associação com um ingrediente principal.

Cabe salientar que o Algoritmo 3.4 também é responsável por associar cada sentença de ingrediente $si \in SI$ ao ingrediente principal encontrado. Dessa forma, ao final do processo de identificação de ingrediente principal, cada sentença $s \in S$ de cada receita $r \in R$ tem um ingrediente principal associado, com exceção daquelas sentenças onde não foi possível encontrar o ingrediente principal. Esta associação da sentença de ingrediente ao ingrediente principal não é explicitada no Algoritmo 3.4 apenas a título de simplificação do algoritmo.

3.6 Identificador de verbos

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. Para identificar as ações verbais, a etapa do identificador de verbos recebe como entrada as instruções de preparo de cada uma das receitas encontradas na base de dados de receitas, e tem como saída os verbos encontrados, os quais são armazenados na base de dados de receitas. De forma a permitir a identificação das ações verbais nas instruções de preparo das receitas, primeiramente, buscou-se na *Web* uma lista de verbos no infinitivo com suas conjugações. Foi encontrada uma lista de verbos no Wikicionário[8], contendo 77.647 conjugações verbais. Em seguida os verbos foram armazenados na base de dados de receitas.

Uma vez tendo recebido as 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.

3.7 Corretor de nome de receitas

Uma das informações que possui importância para este trabalho é o nome das receitas, uma vez que a partir desta informação é que será possível descobrir os pratos de uma determinada receita. Entretanto, antes de utilizar o processo de descoberta de pratos, faz-se necessário efetuar a correção dos nomes das receitas. Isso é justificável, uma vez que há diferentes formas (nomes de receitas) para referenciar a um mesmo prato, ou mesmo por consequência de erros ortográficos, o que comprometeria o processo posteriormente na descoberta dos pratos. Desta forma, há duas correções que podem acontecer: a padronização da escrita do nome de uma receita e a correção ortográfica. A Figura 3.4 ilustra os dois tipos de erros que podem existir em um nome de receita.

Visando corrigir os erros apresentados, o corretor de nomes de receitas recebe como entrada, por meio da base de dados de receitas, o nome da receita, tendo como saída

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

o nome corrigido da receita, passando essa informação para o processo de extração de pratos, que é apresentado adiante na Seção 3.8.

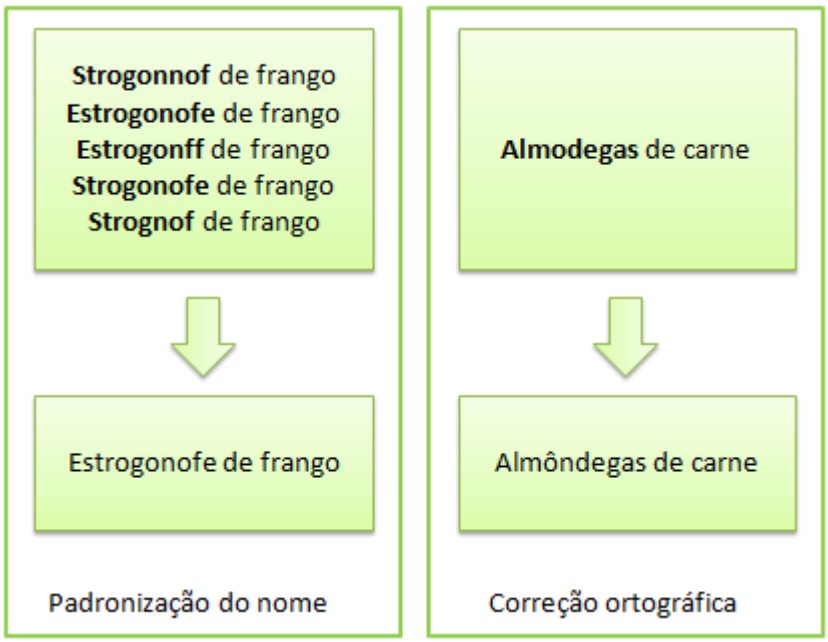


Figura 3.4: Exemplos de erros ortográficos e padronização em nomes de receitas.

O corretor de nomes de receitas utiliza de meta-busca a partir do Google, o que é feito por meio de sua *Application Programming Interface* (API) de coleta[9]. A resposta dada pelo Google para as consultas que consistem dos nomes de receitas da base de dados de receitas pode acontecer de duas maneiras diferentes: uma resposta que explicita que o nome da receita está errado e outra que mostra que o nome da receita está certo.

Uma das respostas dada pelo Google é aquela onde há o erro no nome da receita consultada. Para esses casos, o Google inicialmente apresenta a frase: "Exibindo resultados para (nome sugerido para a receita)" e na sequência expõe a frase "Nenhum resultado encontrado para (nome da receita pesquisada)". Na sequência da página de resposta emitida pelo Google são apresentados os resultados da pesquisa para o nome da receita sugerida como nome certo. A Figura 3.5 ilustra como o Google apresenta os resultados para a consulta de nome de receita "etrognofe de frango", onde o nome da receita não é de fato um nome válido.

A outra maneira de exposição da resposta dada pelo Google acontece quando a receita pesquisada está grafada com o nome certo. Nesse caso, ao receber a consulta com o nome

[9]https://developers.google.com/api-client-library/

Figura 3.5: Resposta da consulta no Google para uma receita com nome errado.

da receita, o Google apenas apresenta os resultados para a devida consulta. A Figura 3.6 ilustra como os resultados são apresentados para a consulta "estrogonofe de frango".

Figura 3.6: Resposta da consulta no Google para uma receita com nome certo.

Por meio da resposta dada pelo Google para uma determinada consulta, verifica-se se o nome da receita está certo ou errado, ou mesmo se o nome é considerado padrão ou não para uma dada receita. Se a resposta dada contiver uma sugestão de reformulação da consulta, conforme ilustra a Figura 3.5, então o nome da receita é corrigido e passa a ter o nome sugerido pelo Google. Assim, no exemplo da Figura 3.5, o nome da receita que antes era "etrognofe de frango" passa a ser "estrogonofe de frango". Em contrapartida, se a resposta dada não contiver uma sugestão, considera-se que o nome da receita está

correto e nenhuma alteração acontece no nome da receita.

A escolha de utilizar o Google como meta-buscador tem como uma das principais justificativas a confiabilidade. Outro ponto a ser levantado é que o número de nomes distintos de receitas não é tão grande, viabilizando, assim, a consulta por meio do Google de forma que sejam respeitados os limites de coleta. A partir do momento que se tem uma base de nomes de receitas de qualidade, seria possível utilizá-la com o intuito de extrair o mesmo tipo de informação. Isso pode ser feito com a utilização da distância de edição de Levenshtein (Levenshtein, 1966), que calcula o custo para duas palavras se tornarem iguais a partir de operações sobre caracteres. Entretanto, utilizar da distância de edição sem ter uma base de dados de qualidade estabelecida previamente pode ser ineficaz, uma vez que pode permitir um maior índice de dados ruidosos, se comparado com o uso do Google, conforme apresentado.

3.8 Extrator de Pratos

Os *sites* de compartilhamento de receitas gastronômicas existentes atualmente oferecem a opção de busca por receitas, abrindo espaço para que uma receita que não apresente boa qualidade, em relação a poucas e insuficientes informações sobre instruções de preparo, ou mesmo com a utilização de ingredientes que destoam da realidade daquela receita seja apresentada, não atendendo assim ao usuário quanto à sua necessidade. Entretanto, o usuário possui autonomia para buscar pela receita que julgar melhor, contudo, ao fazer isso, o usuário irá demandar tempo e habilidade culinária para escolher entre as opções.

O presente trabalho propõe que a forma de pesquisa seja por pratos e não por receitas. Ao oferecer a busca por pratos, seriam analisadas informações como instruções de preparo e os principais ingredientes sobre receitas de um mesmo prato para, com as informações encontradas, oferecer uma receita que atenda melhor às características do prato desejado, com o propósito de atender de forma rápida e eficiente à necessidade de informação do usuário.

No entanto, para viabilizar essa opção, necessita-se que se descubra a qual(is) prato(s) uma determinada receita se refere. Nesse contexto, foi necessário desenvolver uma heurística, extrator de pratos, com esse objetivo. O extrator de pratos recebe como entrada o nome da receita, após a correção efetuada na etapa 7 da metodologia, além dos ingredientes principais, advindos da base de ingredientes, e devolve como saída o(s)

prato(s) para uma dada receita e seu(s) nível(is), armazenando essas informações na base de dados de pratos, além de repassá-las para as duas próximas etapas da metodologia, as quais serão apresentadas adiante.

Cabe destacar que um dado nome de receita pode ser classificado como sendo diferentes pratos, sendo um de cada nível de especificidade diferente. A Tabela 3.4 ilustra esta possibilidade, onde um nome de receita é classificado em quatro pratos em níveis diferentes. A heurística de extração de pratos a ser apresentada adiante trabalha com até quatro níveis de classificação.

Tabela 3.4: Exemplo de pratos e seus níveis para a receita "macarrão com frango desfiado".

Nome Prato	Nível Prato
macarrão	1
frango	2
macarrão com frango	3
macarrão com frango desfiado	4

O Algoritmo 3.5 apresenta o pseudocódigo da heurística desenvolvida para a extração de pratos, onde a entrada R consiste nas receitas validadas, e como saída tem-se os pratos P identificados e seus respectivos níveis, para cada receita $r \in R$. Observa-se no Algoritmo 3.5 que, inicialmente, para cada receita ($r \in R$), recupera-se o nome da receita n e conta-se o número de termos presente no nome da receita, armazenando essa informação em *cont*.

Na sequência do Algoritmo 3.5, começam as diversas verificações que visam identificar a qual padrão o nome da receita pertence. Quando o padrão é identificado, é chamada a função que resolve aquele padrão, passando para essa função o nome da receita (n) e o vetor P. Como se pode observar, a heurística apresenta diversas funções as quais são responsáveis por efetuar a identificação dos pratos gastronômicos e seus níveis, de acordo com os padrões encontrados. Por exemplo, a primeira verificação realizada analisa se os dois primeiros termos do nome da receita estão contidos em algumas das expressões especiais utilizadas no desenvolvimento da heurística. A Tabela 3.5 apresenta alguns padrões existentes e que são resolvidos pelas funções presentes no algoritmo.

A primeira função responsável pela extração de pratos e seus níveis é a *resolve_exp_esp()*.

Algoritmo 3.5: Heurística para encontrar pratos.

```
   Entrada: R
   Saída: P
1  início
2      para cada r ∈ R faça
3          n ← retira_nome_receita(r)
4          cont ← conta_quantidade_termos(n)
5          se 2 primeiros termos de n ⊃ "delícia de" | "o melhor" | "festival de" |
           "como fazer" então
6              resolve_exp_esp(n, P)
7          senão se cont = 3 && n ⊃ ("de" | "a" | "ao" | ... | "super") então
8              resolve_3_termos(n, P)
9          senão se cont = 2 então
10             resolve_2_termos(n, P)
11         senão se n ⊅ espaço (" ") então
12             resolve_espaço(n, P)
13         senão se n ⊃ ("e" | "com") no segundo termo então
14             resolve_e_com(n, P)
15         senão se n ⊃ ("da" | "do" | "a moda" | "por" | "by") então
16             resolve_autor(n, P)
17
18 fim
```

Para que uma receita utilize essa função, os dois primeiros termos do nome da receita n devem estar contidos nas expressões especiais utilizadas no desenvolvimento da heurística. Pode-se verificar que a sentença que é resolvida por esta função, apresenta uma das expressões especiais antes do nome do prato principal. Dessa forma, primeiramente quebra-se a *string* na expressão especial, neste caso "o melhor". Assim, o prato de nível 1 será o restante do nome da receita, nesse caso, "feijão". Considera-se ainda, como prato de nível 2, o nome completo da receita.

Caso o nome de receita n não entre no primeiro padrão, verifica-se o segundo padrão. Para que o nome de uma receita entre no segundo padrão o valor de *cont* tem que ser igual a três e n deve conter alguma das seguintes expressões: "de", "a", "ao", "na", "para", "no", "via", "e", "com" "típico" e "super". Caso isso ocorra, chama-se a função *resolve_3_termos()*. Conforme verifica-se na Tabela 3.5, quando ocorre esse caso, são identificados 2 níveis de pratos, sendo que o primeiro é somente a primeira palavra do nome da receita e o segundo é o nome completo da receita.

Na sequência, se o valor de *cont* for igual a dois, então chama-se a função *re-*

Tabela 3.5: Alguns dos padrões existentes e resolvidos pelas funções da heurística para extrair pratos

Função	Nome da Receita	Pratos	Níveis
resolve_exp_esp()	o melhor feijão	feijão	1
		o melhor feijão	2
		-	-
		-	-
resolve_3_termos()	bolo de milho	bolo	1
		bolo de milho	2
		-	-
		-	-
resolve_2_termos()	abacate recheado	abacate	1
		abacate recheado	2
		-	-
		-	-
resolve_espaço()	lasanha	lasanha	1
		-	-
		-	-
		-	-
resolve_e_com()	frango com batata palha	frango	1
		batata	2
		frango com batata	3
		frango com batata palha	4
resolve_autor()	bolo de fubá com queijo da ana	bolo	1
		bolo de fubá	2
		bolo de fubá com queijo	3
		bolo de fubá com queijo da ana	4

solve_2_termos(). Verifica-se na Tabela 3.5 que o nome de receita resolvida por essa função apresenta apenas dois termos (abacate recheado). Neste caso, há duas possibilidades: a primeira é vista quando o primeiro termo do nome da receita consiste em um ingrediente; já a segunda é vista quando o primeiro termo não é um ingrediente. Assim, quebra-se a *string* no espaço entre os dois termos, em seguida, consulta-se na base de ingredientes principais se o primeiro termo consiste em um ingrediente. Em caso afirmativo, dois pratos são visualizados, sendo no exemplo dado pela Tabela 3.5, "abacate" como prato de nível 1 e "abacate recheado" como prato de nível 2. Se por exemplo o nome da receita fosse "misto quente", apenas um prato seria identificado, sendo o

próprio nome da receita, uma vez que o termo "misto" não representa um ingrediente.

Seguindo o Algoritmo 3.5, verifica-se se há a presença de espaço no nome da receita n. Caso não haja, chama-se a função *resolve_espaço()*, conforme apresentado na Tabela 3.5, constata-se que há apenas um nível de prato, sendo este o próprio nome da receita.

Na sequência, a próxima função é a *resolve_e_com()*, que é responsável por resolver nomes de receitas que tenham em seu segundo termo um dos seguintes conectivos: "e" e "com". Quando surgir em nomes de receitas que se encaixem neste padrão, primeiramente quebra-se a *string* em cada um dos espaços. Na sequência, verifica-se se o terceiro termo do nome está presente na base de dados de ingredientes principais. Em caso afirmativo, o prato de nível 1 será o primeiro termo do nome que, no exemplo dado na Tabela 3.5, seria "frango". O prato de nível 2 consiste no terceiro termo do nome "batata"; já o prato de nível 3 seria composto pelos três primeiros termos "frango com batata"). Por fim, o prato de nível 4 é todo o nome da receita.

Finalmente, caso o nome da receita não tenha sido encontrado em nenhum dos padrões anteriores, verifica-se se o nome da receita n possui uma das seguintes expressões: "do", "da", "a moda", "por" ou "by". Em caso afirmativo, a função *resolve_autor()* se responsabiliza por extrair o(s) prato(s). A Tabela 3.5 apresenta um exemplo de nome de receita com essa característica. Observa-se que os nomes de receitas resolvidos por essa função apresentam o autor da receita. Assim, para efetuar a extração do(s) prato(s) e seu(s) nível(is), primeiramente divide-se a *string* do nome da receita em uma das expressões citadas anteriormente. Em seguida, a parte da esquerda da *string* é enviada como parâmetro para uma das funções apresentadas anteriormente, a qual coincide com os padrões da *string* enviada, e assim a função devida (que possui o padrão da *string* enviada) a resolve. E o nome todo da receita é considerado o prato de nível mais específico, neste caso, o prato de nível 4.

3.9 Identificador de categorias de pratos

Um dos conceitos primordiais deste trabalho está associado à relação entre receitas e pratos, onde um prato é representado por uma ou mais receitas. Ao estabelecer essa relação, muitos atributos relacionados às receitas em si precisam ser processados para se adaptarem ao conceito de pratos. Um destes atributos é a categoria do prato. Todas as receitas apresentam classificações em categorias, de forma que, uma receita pode

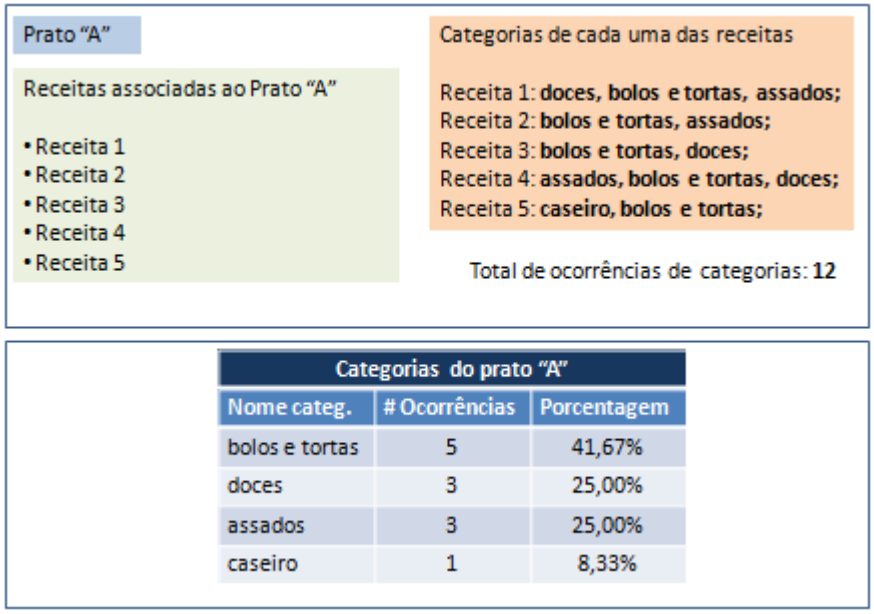
apresentar uma ou mais categorias associadas. Para as fontes Tudo Gostoso e Edu Guedes, cada receita apresenta apenas uma categoria relacionada. Entretanto, para as demais fontes, uma receita pode se relacionar a várias outras categorias, como pode ser visualizado na Figura 3.7.

Figura 3.7: Exemplo de categoria(s) em receitas.

A questão em aberto nesse ponto é a seguinte: como estabelecer uma associação entre as categorias das diversas receitas de um mesmo prato, de forma a categorizar o prato de uma forma geral? Visando responder este questionamento é que se dá origem ao identificador de categorias de pratos. Conforme observa-se na Figura 3.7, o identificador de categorias de pratos recebe como entrada o resultado da etapa anterior, extrator de pratos, contendo o nome dos pratos juntamente com as demais informações das receitas (base de dados de receitas) que se associam àquele prato. Como saída, é enviada à base de dados de pratos, a lista de categorias que se refere a um determinado prato.

Para definir categorias para um prato, realizou-se um agrupamento de todas as receitas relacionadas a um determinado prato e, na sequência, identificaram-se todas as categorias para cada uma das receitas relacionadas ao prato. Após essa identificação, verifica-se qual das categorias é a predominante, além das outras possíveis categorias. Entretanto, definiu-se que serão consideradas apenas as categorias que apresentarem no

mínimo 10% das associações entre categorias e receitas. A Figura 3.8 ilustra como se dá o processo de categorização de um determinado prato em relação à categoria.



Categorias do prato "A"		
Nome categ.	# Ocorrências	Porcentagem
bolos e tortas	5	41,67%
doces	3	25,00%
assados	3	25,00%
caseiro	1	8,33%

Figura 3.8: Exemplo de categorização de um prato em relação à categoria do prato.

A ilustração da Figura 3.8 mostra que entre as ocorrências de categorias nas receitas do prato "A", a categoria que mais ocorreu foi "bolos e tortas", ocorrendo em 41,67% das associações entre receitas e categorias. Dessa forma, considera-se esta como a categoria principal do prato "A". Entretanto, conforme citado anteriormente, considera-se ainda todas as categorias que possuem no mínimo 10% das ocorrências, sendo assim, neste caso, ainda são levadas em consideração as categorias "doces" e "assados", associando-as ao prato "A" como categorias complementares à categoria principal "bolos e tortas".

3.10 Identificador de modo de preparo dos pratos

Uma informação que pode ser considerada relevante para um usuário escolher um determinado prato consiste na identificação do modo de preparo deste prato. Um dos motivos pode estar relacionado à questão do quanto saudável um prato pode ser. Nos dias atuais, tem-se difundido a ideologia de alimentar-se saudavelmente, seja por motivos que se relacionam à saúde, ou mesmo por motivos de dietas com intuito de manter a boa forma física ou por alguma restrição alimentar, conforme é salientado por Sichieri et al.

(2000). Dessa forma, a maneira como se dá o preparo de um prato pode influenciar na escolha do mesmo por parte do usuário.

Tendo em vista essa necessidade, faz-se necessário classificar os pratos mediante seu modo de preparo. Para isso, utilizou-se das ações verbais que representam possíveis formas de preparo de receitas. Assim, estabeleceram-se algumas formas de preparo de um prato com base nas ações verbais que permitiam identificar quais os processos de execução do mesmo a partir de suas receitas. A Tabela 3.6 apresenta os verbos mais frequentes e que foram associados às formas de preparo das receitas, definidas neste trabalho de forma manual pelo especialista.

Tabela 3.6: Ações que ocorrem simultaneamente.

Formas de preparo	Verbos associados
Assada	Assar
	Gratinar
	Untar
	Dourar
	Torrar
	Grelhar
	Tostar
	Rosar
	Corar
Cozida	Cozinhar
	Ferver
	Aferventar
	Aquentar
	Requentar
	Cozer
Frita	Fritar
Refogada	Refogar
Crua	—

Visualiza-se na Tabela 3.6 que foram definidas cinco categorias relacionadas à forma de preparo de pratos. Dentre essas, quatro são categorizadas levando-se em consideração a presença de ações verbais que condizem com a forma de preparo apresentadas na tabela.

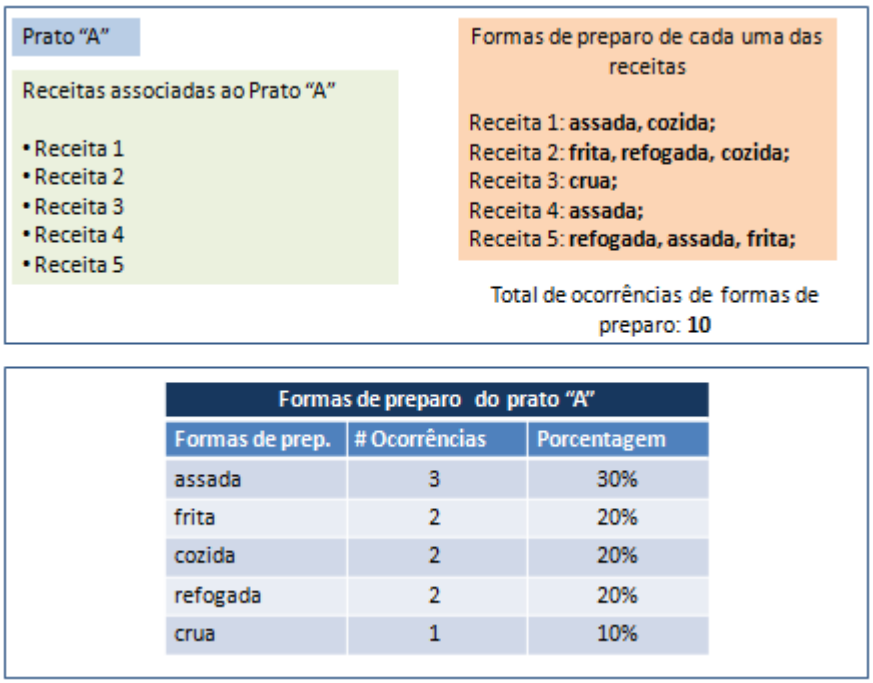
Já a quinta categoria de preparo (crua), ao contrário, é classificada quando não há a presença das ações verbais identificadas nas demais categorias de preparo.

O identificador de modo de preparo dos pratos recebe como entrada o resultado do extrator de pratos, contendo o nome do prato juntamente com as demais informações das receitas (base de dados de receitas) que se associam àquele prato. Como saída, é enviada à base de dados de pratos a lista dos possíveis modos de preparo que se referem a um determinado prato.

A maneira como se deu a categorização para os pratos mediante a forma de preparo é similar à forma como se deu a categorização dos pratos em relação às categorias, conforme pode ser visualizado na Seção 3.9. Primeiramente, verifica-se para cada uma das receitas qual(is) o(s) modo(s) de preparo existente(s). Essa verificação é bem simples, onde se analisa em cada uma das instruções de preparo de uma determinada receita se há a presença dos verbos identificados como fundamentais na identificação de um modo de preparo, visualizados na Tabela 3.6. Desta forma, constata-se que uma receita pode apresentar um ou mais modos de preparo. Isso pode ser visto em casos onde há verbos que associam uma receita a processos diferentes de preparo, como assar e cozinhar.

Assim, uma vez tendo para cada receita a informação de como esta foi preparada, é possível agora efetuar a categorização do modo de preparo também para os pratos. Para isso, contabiliza-se para todas as possíveis receitas de um determinado prato qual(is) a(s) forma(s) de preparo está(ão) presente(s). Desta maneira, é possível identificar o modo de preparo predominante de um determinado prato, bem como escolher uma receita que represente este prato de acordo com suas possíveis formas de preparo. A Figura 3.9 ilustra como esse processo acontece.

Analisando a Figura 3.9, visualiza-se que agora um prato pode ter uma forma de preparo que é usada na maioria das receitas. É possível ainda verificar que o usuário pode ter a opção de escolher um prato que possa ser preparado de diferentes maneiras. Para isso, basta apresentar uma receita que ofereça em sua forma de preparo o que o usuário deseja. Por exemplo, se o usuário desejar uma receita que tenha como forma de preparo o processo de assar, será apresentado ao usuário receita(s) que apresente(m) tal processo, onde nesse caso poderia ser: Receita 1, Receita 4 ou Receita 5.



Formas de preparo do prato "A"		
Formas de prep.	# Ocorrências	Porcentagem
assada	3	30%
frita	2	20%
cozida	2	20%
refogada	2	20%
crua	1	10%

Figura 3.9: Exemplo de categorização de um prato em relação à forma de preparo.

3.11 Gerador de conjuntos de ingredientes frequentes

Cada uma das receitas apresentam especificidades que podem ser observadas por meio dos ingredientes que as compõem, ou mesmo mediante a forma de preparo do prato, através das instruções de preparo. Para isso, faz-se necessário identificar os conjuntos de ingredientes frequentes para cada um dos pratos que contenha no mínimo duas receitas associadas. O componente responsável por esta tarefa é o gerador de conjuntos de ingredientes frequentes, que recebe dados das bases de dados de receitas, de ingredientes e de pratos, e que tem como saída os possíveis conjuntos de itens frequentes para cada um dos pratos, armazenando esses dados na base de dados de pratos.

A necessidade de encontrar os conjuntos de ingredientes frequentes em um determinado prato é justificada pela importância destes ingredientes para a preparação do prato, visando utilizar do conhecimento coletivo. Desta forma, podem ser identificados os ingredientes, mais utilizados em conjunto nas diversas receitas de um dado prato.

Inicialmente, foram preparadas as bases de dados de ingredientes de receitas para cada um dos pratos. Estas bases foram geradas da seguinte maneira: cada receita associada a um prato representava uma transação diferente. Portanto, cada transação

era composta pelos ingredientes principais presentes em cada receita daquele prato. A Tabela 3.7 apresenta uma ilustração de como foram preparadas as base de dados.

Tabela 3.7: Exemplo de base de dados para um determinado prato contendo 5 receitas.

Receitas	Ingredientes
Receita 1	{ing 1, ing 2, ing 3, ing 4, ing 5}
Receita 2	{ing 3, ing 7, ing 2, ing 4, ing 6}
Receita 3	{ing 1, ing 6, ing 3, ing 2}
Receita 4	{ing 1, ing 3, ing 4, ing 6, ing 7, ing 8}
Receita 5	{ing 2, ing 5, ing 7, ing 9, ing 8, ing 10}

A identificação de conjuntos de ingredientes frequentes pode ser resolvida utilizando a tarefa de mineração de dados: análise de regras de associação, que segundo Camilo and Silva (2009) é uma das tarefas mais conhecidas em Mineração de Dados, tendo como exemplo clássico de aplicação, o problema da análise de cesta de compra. De acordo com Agrawal et al. (1993), a tarefa de análise de regras de associação é dividida em duas etapas. A primeira etapa é responsável por gerar os conjuntos de itens frequentes. Para a geração destes conjuntos de itens frequentes, é estabelecido um valor o qual deve ser respeitado, chamado de suporte. Assim, apenas os conjuntos que tenham ocorrido no mínimo o percentual definido para o suporte são gerados. A segunda etapa consiste na mineração dos conjuntos de itens frequentes, chegando-se assim às regras de associação. Neste momento, verifica-se a presença de mais uma métrica, chamada confiança, que visa avaliar se as regras geradas são relevantes.

Ressalta-se que este trabalho utiliza-se apenas dos conjuntos de ingredientes frequentes gerados na primeira etapa da análise por regras de associação. Isso porque o interesse aqui não é identificar conjuntos de ingredientes que, combinados, frequentemente levam ao uso de um outro conjunto de ingredientes no preparo de um prato, mas sim, o conjunto de ingredientes que co-ocorrem com frequência nas receitas que compõem o prato. O algoritmo Eclat (Zaki et al., 1997) foi utilizado para geração dos conjuntos de ingredientes frequentes.

Para a execução do algoritmo Eclat, utilizou-se do pacote Arules (Hahsler et al., 2005), disponível no *software* R-Project[10]. O algoritmo é executado para cada uma

[10]https://www.r-project.org/

das bases de dados geradas, para cada prato, escrevendo os conjuntos de ingredientes frequentes na base de dados de pratos, conforme visualiza-se na 3.1.

O algoritmo Eclat pode receber três atributos em sua configuração: suporte, minlen e maxlen. O suporte, conforme supracitado, estabelece uma porcentagem mínima de ocorrências de um determinado conjunto de ingredientes frequentes diante do total de trasações da base de dados. O minlen define o número mínimo de itens que devem estar contidos em cada conjunto de itens frequentes e por fim, o maxlen define o número máximo de itens que devem estar contidos em cada conjunto de itens frequentes. Neste trabalho, apenas dois parâmetros foram utilizados: o suporte e o minlen. O suporte mínimo utilizado foi 0,01, o que significa que se o conjunto de ingredientes ocorrer em pelo menos 1% das receitas de um dado prato, este conjunto de ingredientes será avaliado. Já o minlen (tamanho da lista de ingredientes) recebeu o valor 2 como parâmetro de configuração, o que significa que os conjuntos de ingredientes frequentes terão no mínimo dois ingredientes.

A Tabela 3.8 apresenta o formato de como que os arquivos de saída contendo os resultados dos conjuntos de ingredientes frequentes são apresentados.

Tabela 3.8: Exemplo de como encontram-se alguns dos conjuntos de ingredientes frequentes referentes ao prato Almôndega.

Ingredientes contidos no conjunto	Minlen (tamanho da lista)	Suporte
{cebola, sal}	2	0,57
{alho, cebola}	2	0,46
{cebola, ovo}	2	0,43
{alho, cebola, sal}	3	0,39
{alho, carne moída, cebola, sal}	4	0,22

Finalmente, após a geração dos conjuntos de ingredientes frequentes, verifica-se que podem haver diversos conjuntos de ingredientes frequentes para um dado prato, com diferentes valores de suporte. Desta forma, há necessidade de decidir qual conjunto escolher, para que os ingredientes do conjunto frequente sejam apresentados como sendo importantes de serem usados no preparo do prato em questão. Para isso, utiliza-se do maior resultado entre a multiplicação do suporte do conjunto de ingredientes frequentes e do número de ingredientes, em tamanho de lista, que compõe o conjunto de ingredientes frequentes. A Equação 3.1 apresenta o cálculo, que é realizado para cada conjunto de

ingredientes frequentes de um prato.

$$r = sup \times num\_ing, \tag{3.1}$$

onde r é o resultado calculado, *sup* é o suporte do conjunto de ingredientes frequentes e *num_ing* representa o número de ingredientes que compõe o conjunto de ingredientes frequentes.

Após a aplicação da equação, escolhe-se o conjunto de ingredientes que obteve maior valor r e, desta forma, será o conjunto de ingredientes que representa os principais ingredientes de um dado prato. Ressalta-se que se utiliza a Equação 3.1, uma vez que esta dá importância similar para as duas medidas que são importantes na escolha de um conjunto de ingredientes, que são o suporte e o número de ingredientes presentes no conjunto de ingredientes, não priorizando, assim, nenhuma das medidas isoladamente.

3.12 Inversor de Ingredientes/Pratos

Esta seção apresenta a etapa Inversor de Ingredientes/Pratos, que é responsável por criar um índice invertido de ingredientes em relação a pratos. Esta etapa recebe como entrada informações das bases de dados de ingredientes, receitas e pratos. Como saída, tem-se uma base de dados de índice invertido. A importância de criar um índice invertido de ingredientes em relação a pratos é visualizada, uma vez que, permite que o usuário efetue buscas de pratos por meio de seus ingredientes. Assim, uma receita será retornada mediante aos ingredientes que o usuário definir em sua consulta. Dessa forma, verificou-se a necessidade de uso de uma estrutura de dados consolidada na área de recuperação de informação que é o arquivo invertido (Zobel et al., 1998).

Os documentos de entrada no índice invertido são compostos por: identificador do prato, identificador da receita e finalmente a lista de ingredientes que compõem a receita. Desta forma, verifica-se que cada documento possui a associação entre uma receita e seu prato. A Tabela 3.9 ilustra os documentos de entrada no inversor de ingredientes/pratos. Observa-se que cada receita de cada prato representa um documento diferente a ser indexado. A Figura 3.10 ilustra o processo de geração do índice invertido de ingredientes.

Geralmente a indexação em um índice invertido acontece textualmente. Entretanto,

Tabela 3.9: Exemplo de documentos a serem processados pelo Inversor de Ingredientes/Pratos.

Id. Doc.	Id. Prato	Id. Receita	Ingredientes
1	147657	145282	{abacaxi, canela, açúcar}
2	147658	21317	{abacaxi, gema, leite, clara}
3	147658	59878	{abacaxi, leite condensado, gema}
4	147653	24257	{abacate, água, leite em pó, açúcar}
5	147655	73600	{abacaxi, canela, cravo da índia}

verifica-se na Tabela 3.9 e também na Figura 3.10 que há a presença dos identificadores numéricos dos pratos e das receitas. Esse procedimento ocorre devido a facilidade encontrada no processo de consulta, uma vez que, quando o usuário efetuar uma consulta por um determinado prato além de incrementar a consulta com ingredientes de sua preferência, os resultados retornados já estarão com a associação estabelecida entre as receitas retornadas com o prato desejado. Em contrapartida, o fato de armazenar os identificadores de receitas e pratos faz com que se aumente a quantidade de dados indexados, demandando de mais espaço para isso, entretanto, obter os resultados de uma consulta em tempo rápido também é de vital importância em uma aplicação, e optou-se por essa vantagem em detrimento à economia de espaço para indexação.

A Figura 3.10 ilustra a construção de um vocabulário de ingredientes com os identificadores numéricos de receitas e pratos utilizando-se dos documentos ilustrados na Tabela 3.9. Observa-se que existem 5 documentos (receitas) e cada termo destes documentos são indexados. Em seguida efetua-se a contagem de quantos documentos possuem cada termo e quais documentos se associam a eles. Desta forma, o processo de busca consiste na inserção de um termo e são retornados os documentos que possuem este termo. Por exemplo, na figura apresentada se for efetuada a consulta pelos termos “abacaxi” e “canela” são retornados dois documentos, sendo eles o primeiro e o último, visto que os termos procurados estão presentes em ambos. Entretanto, se a consulta possuir o interesse de buscar receitas de um dado prato mediante seus ingredientes, bastaria consultar pelos ingredientes associado ao identificador do prato e assim somente receitas do prato desejado com os ingredientes procurados são retornadas. Isso pode ser feito com os seguintes termos na consulta: “abacaxi”, “canela” e “147657”, assim somente uma receita seria retornada como resultado. Com a possibilidade de efetuar uma busca já com a assoaciação estabelecida entre ingredientes, receitas e pratos, o tempo de resposta

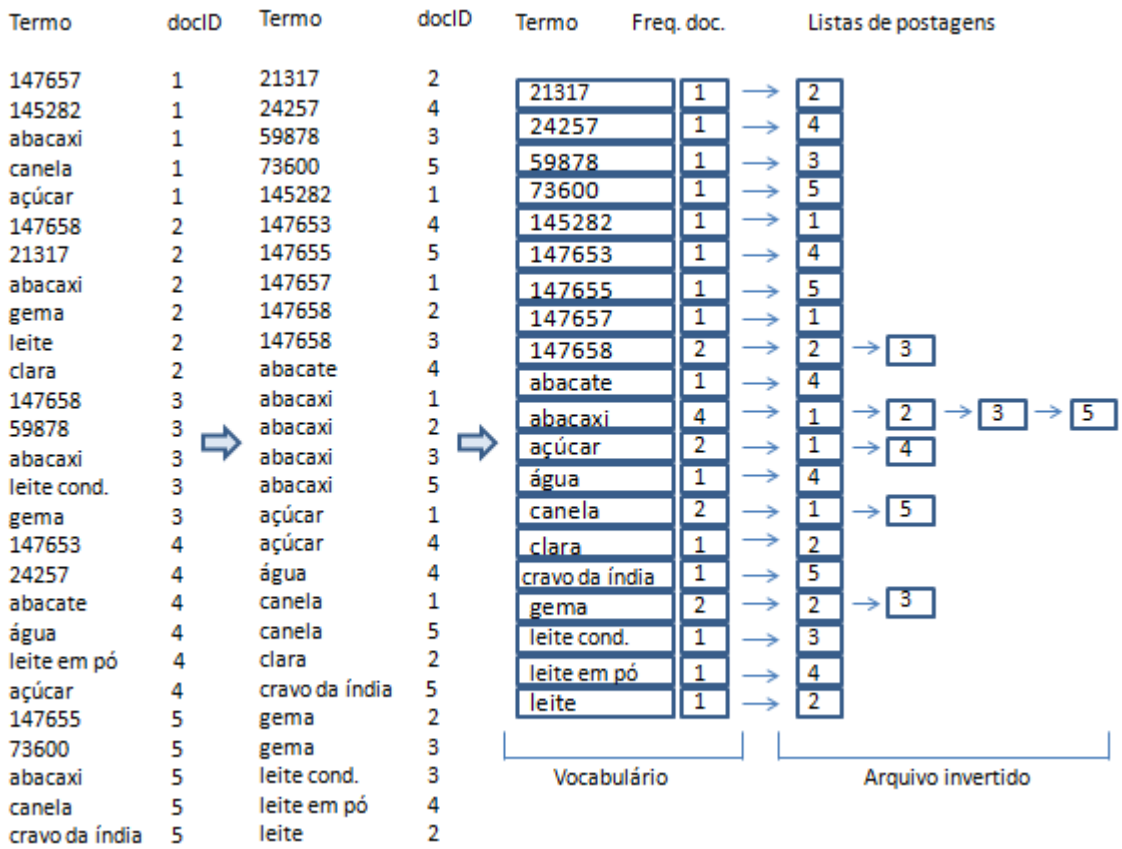


Figura 3.10: Exemplo de construção do índice invertido.

da consulta é minimizado.

Para a criação do índice invertido, utilizou-se de uma ferramenta para indexação e pesquisa de dados, que possui interface *Web* para consultas, que é o Apache Solr, conforme pode ser visto em Grainger et al. (2014) e Kuć (2013). Ressalta-se que o Apache Solr é baseado no Apache Lucene (Hatcher et al., 2004) e (Gospodnetic and Hatcher, 2005). O Apache Solr possui algumas características que contribuem para a indexação e busca como: permite a indexação de diversos tipos de dados, filtros avançados de busca, busca facetada e fonética, extensibilidade, entre outras.

3.13 Processador de consultas

Esta seção apresenta a etapa Processador de Consultas, que é responsável por receber a consulta do usuário em busca de uma receita e por fim entregar receitas como resultado ao usuário. Como se pode ver pela Figura 3.1, o processador de consultas recebe

como entrada a consulta do usuário, além das informações das bases de dados de pratos e também do índice invertido. Como saída, tem-se a interação com o usuário, onde é apresentado o resultado de uma consulta realizada, podendo permitir opções para usuários adaptar a consulta.

O usuário pode interagir com o processador de consultas por meio de três formas, conforme pode ser visualizado no Apêndice A. Na primeira opção, o usuário apenas escolhe receber receitas de um dado prato, sem que se faça alguma inserção ou remoção de ingredientes. Na segunda opção, o usuário pode optar por receber receitas de um dado prato, entretanto, aqui ele pode ainda solicitar que tenha ou não determinados ingredientes conforme sua necessidade ou preferência. Já a terceira opção consiste em uma busca onde o usuário entra apenas com os ingredientes e na sequência ele escolhe uma das possíveis categorias e por fim, escolhe um dos pratos associados à categoria escolhida.

Para a primeira opção, logo na tela inicial do sistema, o usuário apenas informa o nome do prato que está buscando. Na sequência, ele pode informar ao sistema que quer receber as receitas, imediatamente, ou pode ainda incrementar a busca, adicionando algum ingrediente. A cada manipulação de ingredientes realizada pelo usuário, o mesmo recebe uma informação referente à porcentagem de receitas do prato desejado que possuem os ingredientes escolhidos. Quando ele optar por um prato e desejar receber imediatamente as receitas sem que haja manipulação de ingredientes, o sistema informa a porcentagem de receitas do prato desejado que possuem os ingredientes solicitados para aquele prato.

Quando o usuário optar pela primeira opção de consulta, verifica-se que um determinado conjunto de ingredientes frequentes pode ser representado por várias receitas de um dado prato. Nesse momento, surge a necessidade de criar um *ranking* de receitas, de forma a apresentar as receitas mais bem ranqueadas ao usuário. Para isso, primeiramente, verifica-se a necessidade de aplicar um peso à receita conforme a fonte de dados. Este peso está relacionado ao número de receitas que cada fonte de dados apresenta. A intuição é que as bases de dados com mais receitas tendem a ser melhores do que as bases com menos receitas.

Para cada uma das fontes de dados há uma informação que melhor representa a qualidade das receitas. Assim, definiu-se que para as fontes de dados Tudo Gostoso, Cybercook e Dieta e Receitas, a métrica a ser usada é o número de votos da receita. Já para a fonte Receitas.com a métrica a ser usada é o número de pessoas que favoritaram

a receita. Por fim, para a fonte de dados Edu Guedes, a métrica escolhida é o número de curtidas do Facebook. Finalmente, com as métricas definidas, apresenta-se a Equação 3.2, utilizada para gerar os resultados da busca de uma dado prato, ordenando as receitas daquele prato de forma a compor um *ranking* de receitas.

$$r1 = val\_met \times \sqrt{num\_rec}, \tag{3.2}$$

onde $r1$ é o resultado para uma determinada receita para a primeira opção de busca, *val_met* é o valor da métrica utilizada (número de curtidas, votos, entre outras) e *num_rec* representa o peso dado à fonte de dados da receita. Salienta-se que a multiplicação pela raiz quadrada de *num_rec* acontece com o objetivo de suavizar os valores do número de receitas de cada fonte de dados.

Uma vez feito esse cálculo para cada uma das receitas que possuem os ingredientes presentes no conjunto de ingredientes frequentes, estabelece-se o *ranking*, dando maior relevância às receitas que possuem maior valor de *r1*.

A segunda forma consiste na busca de receitas de um determinado prato, porém incluindo os ingredientes que devem estar presentes nas receitas retornadas como resultado. Logo na tela inicial do sistema o usuário escolhe pelo prato desejado, bem como adiciona ingredientes que ele gostaria que estivessem presentes nas receitas. Na sequência, o sistema apresenta ao usuário a porcentagem de receitas do prato desejado que apresenta os ingredientes solicitados. Esta opção de consulta permite ao usuário maior autonomia, na escolha de uma receita. Com a autonomia dada ao usuário, há a possibilidade de não encontrar na base de dados receitas com os ingredientes solicitados, entretanto, neste caso, será reportada esta situação, bem como sugeridas receitas com o maior número possível de ingredientes solicitados, permitindo também que o usuário refaça sua lista de ingredientes desejados.

Quando o usuário optar pela segunda opção de consulta, similarmente à primeira opção, faz-se necessário a criação de um *ranking* de receitas. No entanto, aqui a geração do *ranking* é feita de uma maneira diferente. Como aqui o usuário possui maior autonomia, o ranqueamento das receitas deve priorizar as receitas que têm maior similaridade entre seus ingredientes e os ingredientes desejados pelo usuário. A Equação 3.3 apresenta a fórmula do ranqueamento das receitas para o cenário apresentado.

$$r2 = \frac{num\_ing\_des}{num\_ing\_rec} \times val\_met, \tag{3.3}$$

onde $r2$ é o resultado para uma determinada receita para a segunda opção de busca, *num_ing_des* é o número de ingredientes desejados pelo usuário, *num_ing_rec* representa o número total de ingredientes presentes na receita e *val_met* é o valor da métrica utilizada. Com este cálculo, prioriza-se as receitas que tem maior percentual de ingredientes pesquisados levando-se em consideração ainda o valor da métrica de cada uma das receitas. Ressalta-se a equação não leva em consideração as fontes de dados das receitas, isso ocorre porque aqui como o usuário entra com os ingredientes, então pressupõe-se que ele deseja obter uma receita mais próxima dos ingredientes solicitados, independente de que fonte venha a receita.

A terceira opção de consulta do usuário consiste, inicialmente, na escolha dos ingredientes que o usuário gostaria de utilizar, dando-lhe autonomia para escolher apenas os ingredientes que possui em casa ou mesmo de acordo com suas preferências culinárias. Em seguida, o usuário escolhe uma das categorias possíveis, de forma a refinar sua busca, direcionando-a de acordo com a categoria escolhida. Finalmente, o usuário escolhe um dos pratos associados à categoria desejada. Similarmente à segunda opção de consulta, aqui pode haver a possibilidade da busca do usuário não encontrar resultados, uma vez que ele insere apenas os ingredientes que ele deseja ter nas receitas. No entanto, nesse caso, o sistema pode apresentar receitas que contenham similaridade entre os ingredientes escolhidos, ou mesmo permitir que o usuário refaça sua consulta. O usuário recebe também a informação em porcentagem de receitas do prato desejado que possuem os ingredientes selecionados.

Com a opção pela terceira forma de consulta, onde inicialmente se escolhe os ingredientes e na sequência a categoria e por fim um prato da categoria desejada, também verifica-se a necessidade de estabelecer *rankings* de receitas. O *ranking* aqui é criado utilizando-se das equações dos *rankings* das duas primeiras formas de consulta. Assim, leva-se em consideração a similaridade dos ingredientes das receitas retornadas em relação aos ingredientes desejados, bem como a relevância das fontes de dados de receitas. A Equação 3.4 apresenta a fórmula do ranqueamento das receitas para este cenário.

$$r3 = \frac{num\_ing\_des}{num\_ing\_rec} \times val\_met \times \sqrt{num\_rec}, \tag{3.4}$$

onde $r3$ é o resultado para uma determinada receita para a terceira opção de busca, *num_ing_des* é o número total de ingredientes desejados pelo usuário e *num_ing_rec* representa o número total de ingredientes presentes na receita, *val_met* é o valor da métrica utilizada e, por fim, *num_rec* representa o peso dado à fonte de dados da receita (conforme já apresentado acima).

Salienta-se que para todas as opções de consulta realizadas pelo usuário, ele pode ainda optar por receber receitas de acordo com seu modo de preparo, conforme definido na Seção 3.10. Para isso, são apresentadas as opções possíveis de preparo para o prato desejado. Finalmente, são apresentadas as receitas do prato desejado retornadas pelo sistema ao usuário, em forma de *rankings*. Vale ressaltar que o resultado retornado pelo sistema, se possível, traz receitas de todas as fontes de dados. Isso somente não acontecerá, se não houver resultados para uma determinada fonte de dados.

Capítulo 4

Estudo de Caso

Este capítulo apresenta um estudo de caracterização da base de dados de receitas coletadas na, Seção 4.1. Em seguida, são apresentados estudos relacionados aos resultados das etapas mais importantes da metodologia de descoberta de conhecimento em receitas gastronômicas, na Seção 4.2.

4.1 Caracterização das bases de dados coletadas

Esta seção apresenta uma análise sobre as receitas coletadas, visando identificar algumas das principais características presentes na base de dados de receitas.

A Figura 4.1 apresenta a composição da base de dados de acordo com a porcentagem de receitas coletadas de cada uma das fontes de dados utilizadas. Observa-se na figura que quase 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 uma, menos de 1% das receitas. Ressalta-se que essas fontes foram identificadas como importantes por causa de suas características, como a exposição dada às receitas em um programa televisivo, como acontece com a fonte Edu Guedes, e devido à característica de ser focado em dieta e alimentação saudável, o que se observa em Dieta e Receitas. O número total de receitas coletadas foi de 288.537 receitas.

Com exceção da fonte de dados Edu Guedes, todas as demais fontes possuem em suas receitas avaliações realizadas por usuários que interagem nos *sites*. A Figura 4.2 apresenta uma sumarização das avaliações das receitas das demais fontes de dados, onde

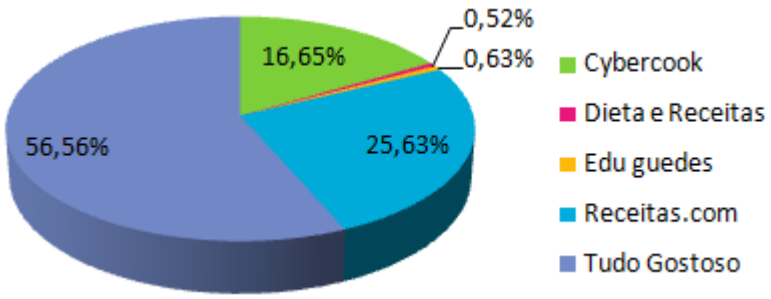


Figura 4.1: Porcentagem de receitas coletadas para cada uma das fontes de dados.

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.

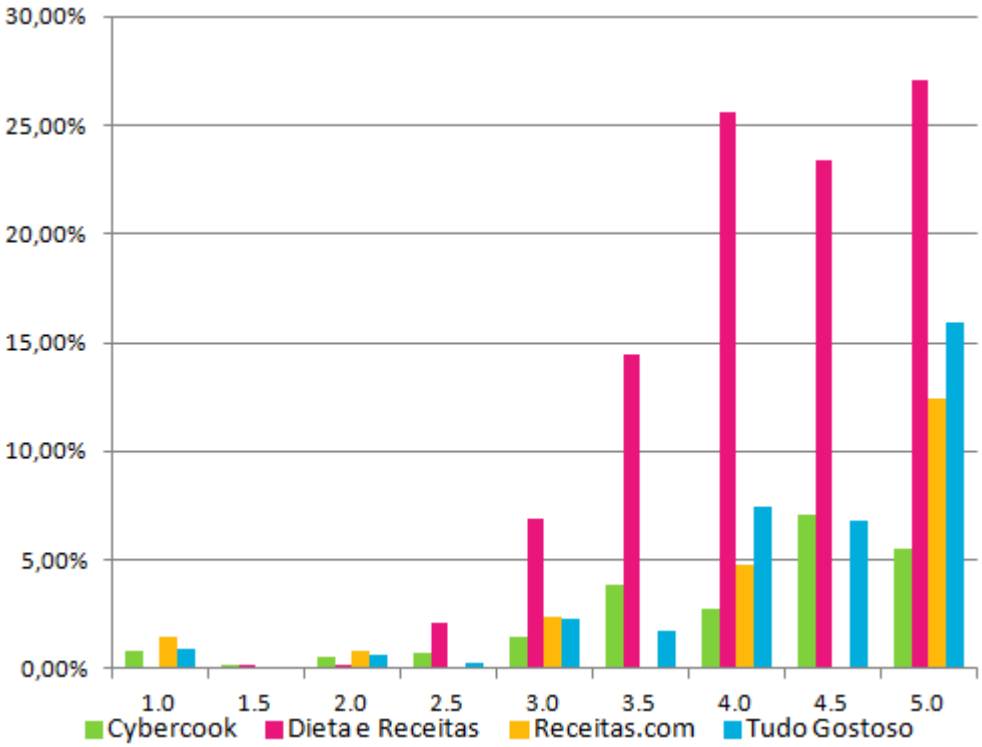


Figura 4.2: Sumarização das avaliações das receitas realizadas pelos usuários.

Observa-se na Figura 4.2 uma semelhança no padrão das avaliações entre 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.

Conforme supracitado, o tipo de informação presente em cada uma das fontes de dados se diferem. O gráfico da Figura 4.3 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 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[1], tweets do Twitter[2] e recomendações do Google Plus[3]. 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.

Com o intuito de analisar uma possível relação entre o número de votos e de comentários nas receitas, pegou-se as 10 receitas com maior número de votos e tentou identificar se realmente há uma relação. Buscou-se ainda analisar possíveis relações associadas aos valores atribuídos pela interação com os dados das redes sociais(curtidas, tweets e recomendações). A Tabela 4.1, lista as dez receitas ranqueadas de acordo com a quantidade de votos que estas receitas receberam. Importa salientar que para esse estudo foram utilizadas somente as receitas referentes à fonte Tudo Gostoso, uma vez que esta fonte apresenta aproximadamente 60% das receitas da base de dados.

Percebe-se, na Tabela 4.1, a relação entre a quantidade de votos e de comentários, onde o valor destes são bem próximos, para as dez receitas que possuem o maior número

[1]https://www.facebook.com
[2]https://twitter.com
[3]https://plus.google.com/

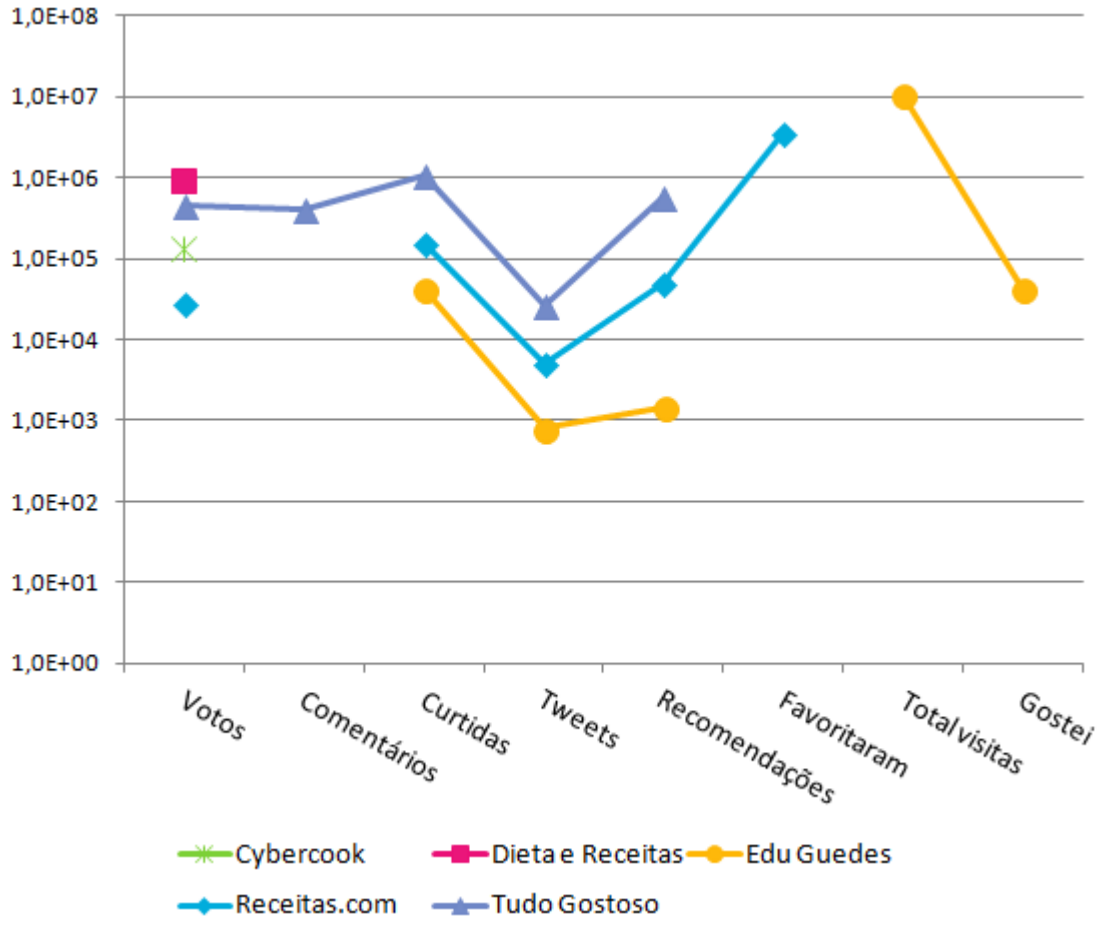


Figura 4.3: Interação dos usuários por meio de características das receitas.

Tabela 4.1: Top 10 receitas ranqueadas pelo maior número de votos.

Receitas	#Votos	#Coment.	#Curtidas	#Tweets	#Recom.
Bolo choc. molhadinho	5.413	5.156	5.401	94	560
Bolo gelado	5.114	4.771	10.709	89	552
Torta liquidificador	4.170	4.097	8.539	55	230
Fricasse frango	3.027	2.934	9.051	60	194
Panqueca de carne moída	2.883	2.489	7.950	77	539
Danononinho caseiro	2.846	2.885	6.394	75	778
Pudim de leite condensado	2.765	2.689	6.164	91	235
Pão caseiro	2.702	2.934	8.087	42	564
Bolinho de chuva	2.596	2.417	8.095	98	478
Bolo de fubá maria	2.537	2.506	3.586	36	193

de votos. Buscou-se, ainda, associar essa relação do número de votos e comentários com os dados da interação das redes sociais. Entretanto, verificou-se que os dados das redes sociais não apresentam valores próximos às outras características das receitas, como se percebe no número de curtidas na primeira receita, que fica abaixo de quase todas as demais receitas apresentadas, mas se verifica que, na maioria dos casos, as receitas que apresentam mais votos e comentários, são aquelas que tendem a ter uma maior quantidade de curtidas, tweets e recomendações. Verifica-se ainda uma maior interação de usuários do Facebook, seguido do Google Plus e Twitter, corroborando as análises efetuadas sobre as características das receitas por meio da interação dos usuários, conforme pode ser analisado pela Figura 4.3.

Entre as fontes trabalhadas, três possuem comentários sobre as receitas (Tudo Gostoso, Cybercook e Dieta e Receitas). Para as fontes Tudo Gostoso e Cybercook, como o volume de receitas é maior, foram coletados apenas os dez comentários mais recentes. Já para a fonte Dieta e Receitas, todos os comentários foram coletados. O total de comentários coletados referentes a receitas validadas 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 Receitas. O gráfico da Figura 4.4 apresenta 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, 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.

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 4.5 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

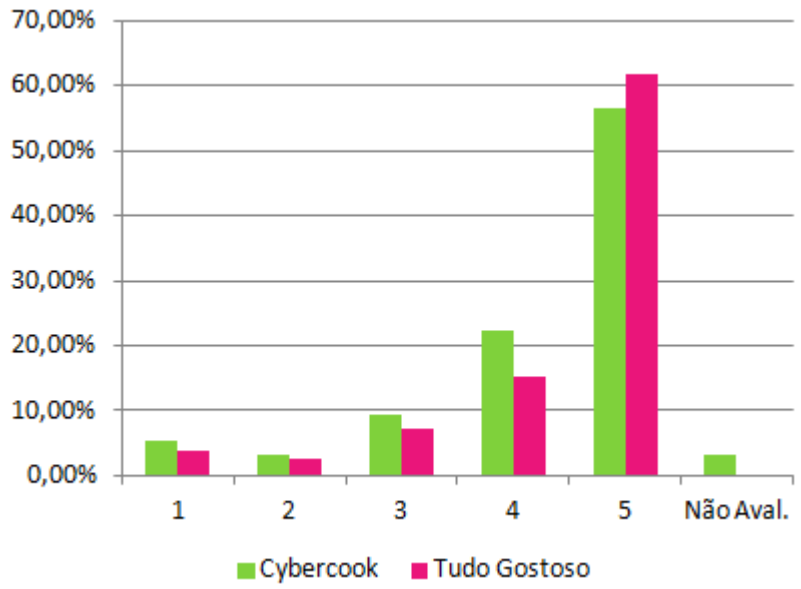


Figura 4.4: Sumarização da avaliação dos comentários realizados pelos usuários.

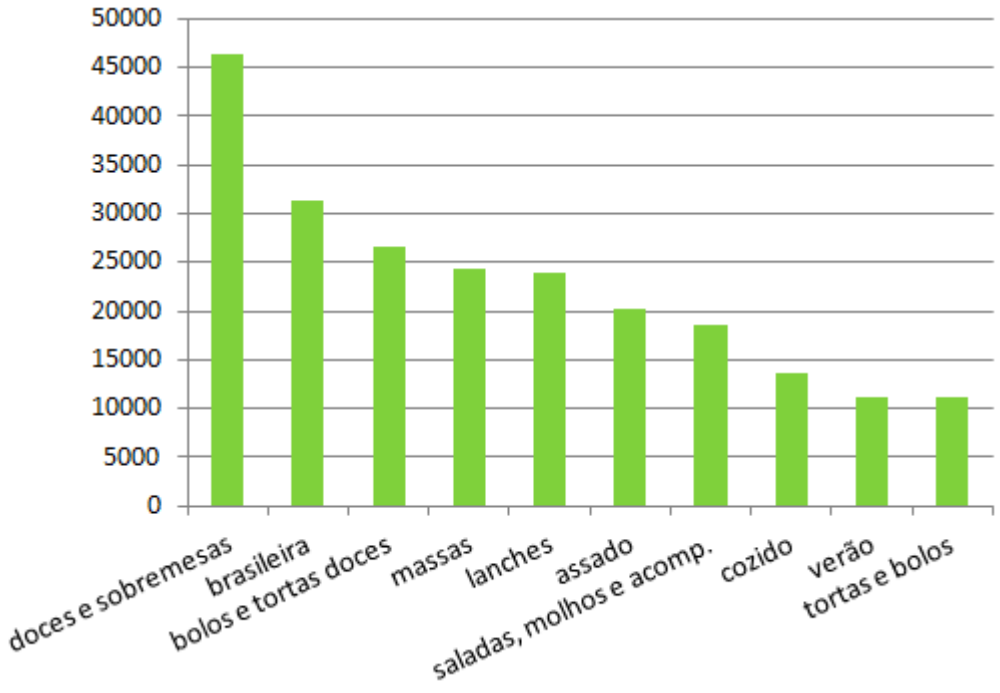


Figura 4.5: As 10 categorias mais comuns.

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

4.5 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.

Por fim, o gráfico da Figura 4.6 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 4.6 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.

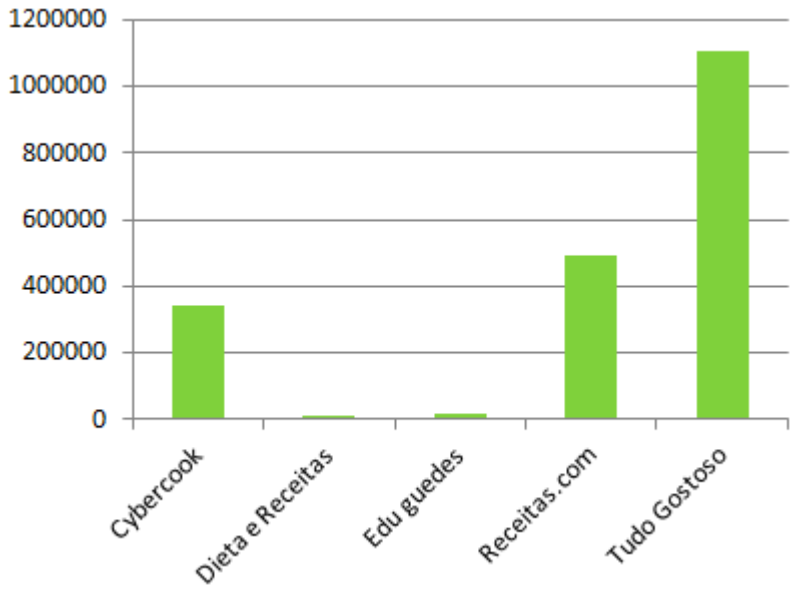


Figura 4.6: Número de usuários identificados que postaram receitas ou comentários.

4.2 Conhecimento descoberto com o uso da metodologia

Esta seção apresenta estudos relacionados aos resultados de algumas etapas da metodologia de descoberta de conhecimento em receitas gastronômicas. A Seção 4.2.1 apresenta um estudo sobre a quantidade de receitas que foram validadas pelo validador de receitas

(Seção 3.4), elucidando a porcentagem de receitas validadas em cada uma das fontes de dados e também de uma maneira geral. A Seção 4.2.2 apresenta um estudo sobre a descoberta de conhecimento sobre os ingredientes. Em seguida, a Seção 4.2.3 apresenta uma análise sobre a extração de ingredientes principais. A Seção 4.2.4 apresenta uma análise das ações verbais identificadas, apresentando os principais verbos encontrados nas instruções de preparo das receitas, bem como alguns verbos que influenciam no modo de preparo de uma receita. A Seção 4.2.5 apresenta uma caracterização dos pratos encontrados após a execução do extrator de pratos. Por fim, a Seção 4.2.6 apresenta uma análise com algumas estatísticas acerca dos conjuntos de ingredientes frequentes gerados.

4.2.1 Validador de receitas

Esta seção apresenta um estudo sobre as quantidades e porcentagens de receitas inseridas na base de dados de receitas após a execução do validador de receitas, que pode ser visualizado na Seção 3.4. Uma vez tendo os dados inseridos na base de dados de receitas, verificou-se a porcentagem de receitas inseridas, identificando-se assim, também, a porcentagem de receitas que não foram inseridas por não atender a restrição proposta no validador de receitas. A Tabela 4.2 apresenta, dentre outras informações, a porcentagem de inserção das receitas para cada uma das fontes de dados e ainda em um contexto geral.

Tabela 4.2: Quantidade e porcentagem de receitas validadas pelo Validador de receitas para cada fonte de dados (T.G.: Tudo Gostoso, Rec.com: Receitas.com, Cyber: Cybercook, E.G: Edu Guedes e D.Rec.: Dieta e Receitas).

	T.G.	Rec.com	Cyber.	E.G.	D. Rec.	Total
#receitas	163.210	73.960	48.039	1.824	1.504	288.537
#rec. ins.	140.094	55.901	39.144	1.541	1.369	238.049
%rec. ins.	85,84%	75,58%	81,48%	84,48%	91,02%	82,50%
#rec. não ins.	23.116	18.059	8.895	283	135	50.488
%rec. não ins.	14,16%	24,42%	18,52%	15,52%	8,98%	17,50%

Ao analisar os valores apresentados na Tabela 4.2, verifica-se que 82,23% das receitas foram validadas, totalizando 238.049 receitas. Desta forma, verifica-se que cerca de

17% das receitas não foram validadas, sendo portanto descartadas. Essa ação faz-se necessária, tendo em vista a importância de se trabalhar com dados com o mínimo de ruído possível. Pode-se verificar também que a fonte de dados Tudo Gostoso (que apresenta maior número de receitas) teve mais de 85% das receitas validadas. Percebe-se ainda que a fonte Receitas.com foi a que apresentou maior valor de receitas não validadas, aproximando-se de 25%, o que pode ser explicado devido às variações de padrão de publicação das receitas, uma vez que algumas receitas não apresentavam o padrão de tópicos, onde cada linha representa uma sentença contendo: quantidade, seguida pela unidade de medida e finalmente o ingrediente. Em parte das receitas as informações sobre os ingredientes e suas quantidades e unidades de medida eram apresentadas por meio de texto corrido, não seguindo o padrão de tópico.

A Figura 4.7 apresenta a composição da base de dados de acordo com a porcentagem de receitas que foram validadas pelo validador de receitas de cada uma das fontes de dados utilizadas. Similarmente à composição da base de dados de receitas coletadas (conforme Figura4.1), observa-se que que quase 60% das receitas validadas que compõem a base de dados são extraídas da fonte Tudo Gostoso. Observa-se ainda que as fontes Dieta e Receitas e Edu Guedes apresentam cada uma, menos de 1% das receitas.

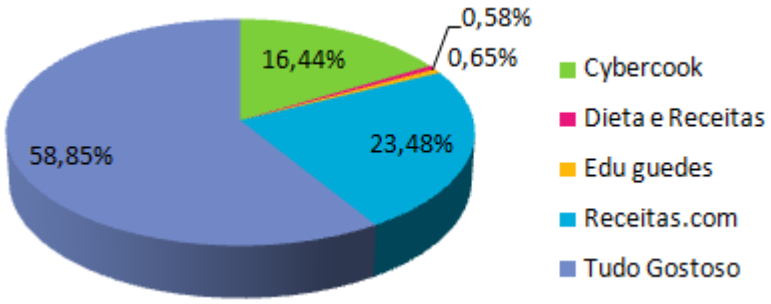


Figura 4.7: Porcentagem de receitas validadas para cada uma das fontes de dados.

4.2.2 Análise dos ingredientes descobertos

Esta seção apresenta uma análise sobre os ingredientes presentes nas receitas após o procedimento de identificação de ingredientes, quantidades e unidades de medida (Seção 3.2) e após a validação das receitas (Seção 3.4).

Na grande maioria das receitas coletadas e posteriormente validadas, para todas as

fontes de dados, verifica-se um padrão na maneira em que são expostos os ingredientes, assim como para as instruções de preparo. Esse padrão é facilmente visualizado, apresentando tais informações em forma de tópicos. Assim, cada sentença que contém o ingrediente, quantidade e unidade de medida é representada por tópicos. Similarmente às sentenças de ingredientes, cada instrução de preparo também é apresentada por tópicos. O número total de sentenças que contém ingrediente, quantidade e unidade de medida encontrados nas receitas da base de dados de receitas 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.

No gráfico da Figura 4.8 é explicitado o número de ingredientes encontrados 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 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 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.

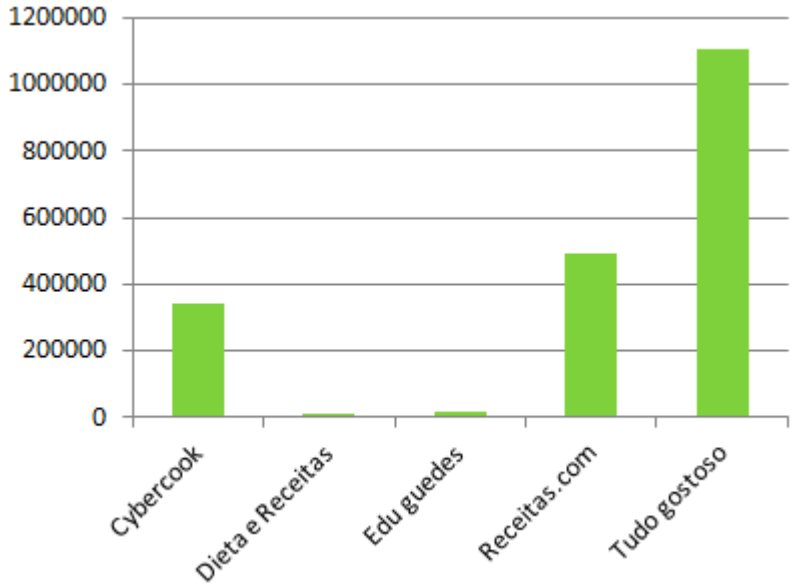


Figura 4.8: Número de ingredientes por fonte de dados.

A Figura 4.9 apresenta uma nuvem de termos com os ingredientes mais frequentes nas receitas de toda a base de dados. 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.

Figura 4.9: Ingredientes mais frequentes encontrados nas receitas.

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 Stephens (1976), 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[4] para encontrar a distribuição e posteriormente gerar o gráfico da Função de Densidade de Probabilidade (PDF), conforme visualiza-se na Figura 4.10.

Verifica-se na Figura 4.10 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 Krishna and Pundir (2009), Mushtaq and Rizvi (2005) e Ko et al. (2013), 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.

4.2.3 Análise do processo de extração de ingrediente principal

Uma das etapas da metodologia de descoberta de conhecimento em receitas gastronômicas consiste na identificação dos ingredientes principais, conforme visualiza-se na Seção 3.5, onde se verifica que o processo de identificação de ingredientes principais é composto por

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

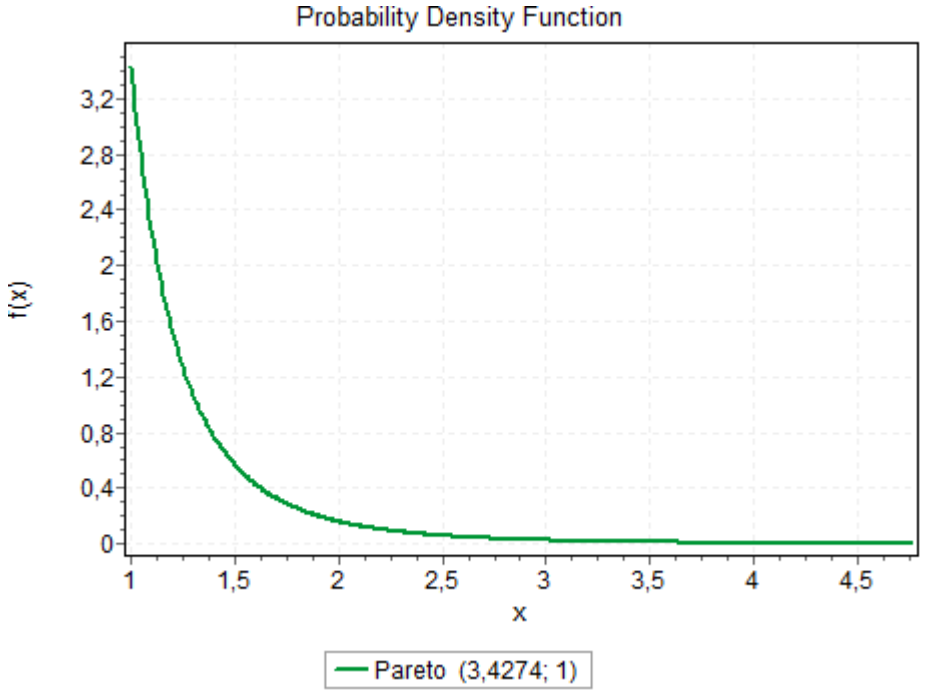


Figura 4.10: Função de Densidade de Probabilidade (PDF) da base de dados de ingredientes.

três fases. Vale salientar que o total de sentenças de ingredientes presentes na base de dados de ingredientes é de 1.971.405, conforme pode ser visualizado na Seção 4.2.2. A Tabela 4.3 apresenta os resultados de identificação de ingredientes principais para cada um das três fases do processo de extração de ingrediente principal.

Tabela 4.3: Número de sentenças de ingredientes associadas a um ingrediente principal por cada fase da identificação de ingrediente principal.

Procedimentos	#Sentenças de ingredientes
Primeira Fase	865.836
Segunda Fase	809.585
Terceira fase	208.506
Total	1.883.927

Após a execução das três fases para a identificação dos ingredientes principais, constatou-se que 95,57% das sentenças de ingredientes foram associadas a um ingrediente principal. Importa salientar que os 4,43% das sentenças de ingredientes que não foram associadas a

um ingrediente principal, referem-se a sentenças que apresentam menos de 50 ocorrências de frequência. Ressalta-se ainda que, deste total, cerca de 50% são de sentenças que ocorreram apenas uma vez. Isso mostra que, apesar de existir um número razoável de sentenças de ingredientes que não foram associadas a um ingrediente principal, as sentenças não associadas possuem pouca representatividade.

4.2.4 Análise das ações verbais identificadas

Com o procedimento efetuado para encontrar os verbos presentes nas instruções de preparo (Seção 3.6), possibilitou-se que algumas análises fossem realizadas. O total de verbos encontrados nas instruções de preparo foi 1.415. Ressalta-se que foram armazenados os verbos no infinitivo. Na Tabela 4.4 são apresentados os 10 verbos mais frequentes.

Tabela 4.4: 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 4.4, 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 permitam 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 4.5 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 4.5: 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%

Destaca-se que para o verbo "assar", foram calculadas 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 se refere à ação de assar algo. Pode ser verificado, assim, que mais de 50% das receitas são levadas ao forno em seu modo de preparo. A ação "fritar", que é tipicamente considerada menos saudável, foi apresentada em apenas 5,22% das receitas.

Há a possibilidade de uma mesma receita apresentar mais de uma das ações encontradas na Tabela 4.5. A Tabela 4.6 retrata a porcentagem de receitas que apresentam interseções em mais de uma das ações em seu modo de preparo. Verifica-se na tabela 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.

Ainda sobre a análise de verbos, a Figura 4.11 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.

Por fim, visando entender o comportamento da base de verbos, analisou-se qual a distribuição estatística melhor se adapta aos dados, conforme apresentado pela Figura 4.12. Similarmente a análise de distribuição de ingredientes, conforme visualiza-se na Seção 4.2.2, 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

Tabela 4.6: 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%

Figura 4.11: Nuvem de termos com os verbos mais frequentes.

verbos e dos ingredientes, 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 base de verbos. Esse comportamento era esperado, uma vez que os verbos mais frequentes são aqueles muito comuns da linguagem ou que estão relacionados à prática da culinária (como os verbos ferver, misturar, deixar, pôr, entre outros).

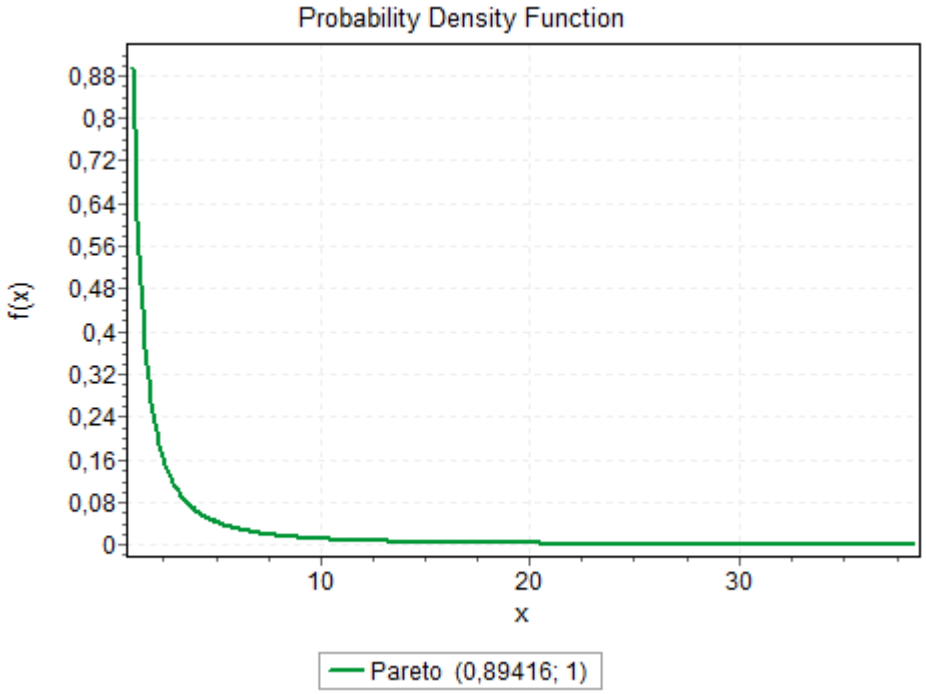


Figura 4.12: Função de Densidade de Probabilidade (PDF) de ocorrência de verbos nas receitas.

4.2.5 Caracterização dos pratos encontrados

Esta seção apresenta uma análise efetuada sobre os pratos encontrados com a utilização do extrator de pratos, que pode ser visualizado na Seção 3.8.

Conforme apresentado na Seção 3.8, onde foi apresentado o processo de extração dos pratos, utilizou-se do nome de receitas para assim identificar os pratos e seus níveis. Após a execução do extrator de pratos para todas as receitas, chegou-se a um resultado de 150.852 pratos. A Tabela 4.7 ilustra a quantidade de pratos, subdivididos em níveis.

Tabela 4.7: Número de pratos por cada nível de prato.

Nível Prato	#Pratos
1	11267
2	52383
3	84002
4	3200

Conforme se verifica na Tabela 4.7, os níveis que apresentam maior quantidade de pratos são os níveis 2 e 3. Isso pode ser explicado, devido a maior especificidade dos pratos encontrados nestes níveis em relação aos de nível 1. Em contrapartida, o nível 4 apresenta a menor quantidade de pratos, embora apresente maior especificidade. Entretanto, neste caso, verifica-se que os pratos são muito específicos e assim, poucas receitas contemplam esse nível de especificidade. Dessa forma, pode-se concluir que, os pratos de níveis pouco ou muito específicos (nível 1 pouco específico e nível 4 muito específico), tendem a não ter uma quantidade elevada de pratos.

O processo de geração de conjuntos de ingredientes frequentes visualizado na Seção 3.11 utiliza-se das receitas associadas aos pratos, com o intuito de identificar ingredientes frequentes em receitas de um mesmo prato. Para isso, um prato deve possuir ao menos 2 receitas, constituindo assim uma base de dados de pratos. Dessa forma, a Tabela 4.8 apresenta o número de bases de dados subdivididas por nível do prato que foram enviadas ao gerador de conjuntos de ingredientes frequentes.

Tabela 4.8: Quantidade de pratos por cada nível de prato enviadas ao gerador de conjuntos de ingredientes frequentes.

Nível Prato	#Pratos
1	4301
2	14236
3	10278
4	313

A Figura 4.13 apresenta os 10 pratos de nível 1 que possuem mais receitas, onde no eixo X apresentam-se os nomes dos pratos e no eixo Y a quantidade de receitas associadas aos pratos. Pode-se perceber na figura que há uma discrepância muito grande em relação ao número de receitas associadas aos pratos mais frequentes. Ao passo que o prato “bolo” ocorre em quase 40 mil receitas, 7 dos demais pratos do topo possuem menos de 10 mil receitas.

A Figura 4.14 apresenta os 10 pratos de nível 2 que possuem mais receitas associadas, onde no eixo X apresentam-se os nomes dos pratos e no eixo Y a quantidade de receitas associadas aos pratos. Observa-se que aqui há uma proximidade maior entre o número de receitas que se associam em cada um dos dez pratos mostrados. Pode ser visualizada ainda uma grande diferença entre os números de receitas, apresentadas nos principais

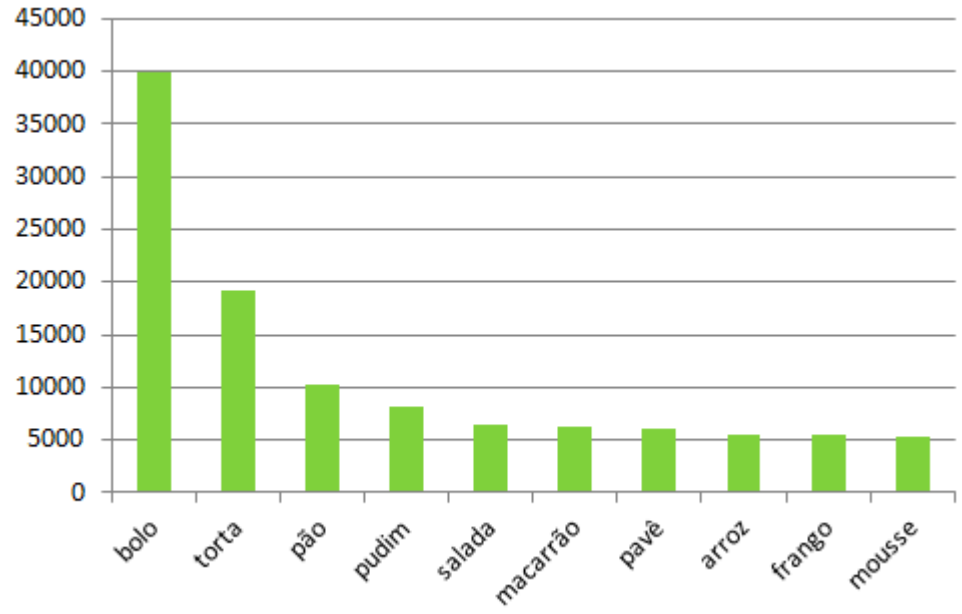


Figura 4.13: Os 10 pratos de nível 1 que apresentam maior número de receitas.

pratos do nível 1 em comparação com o nível 2.

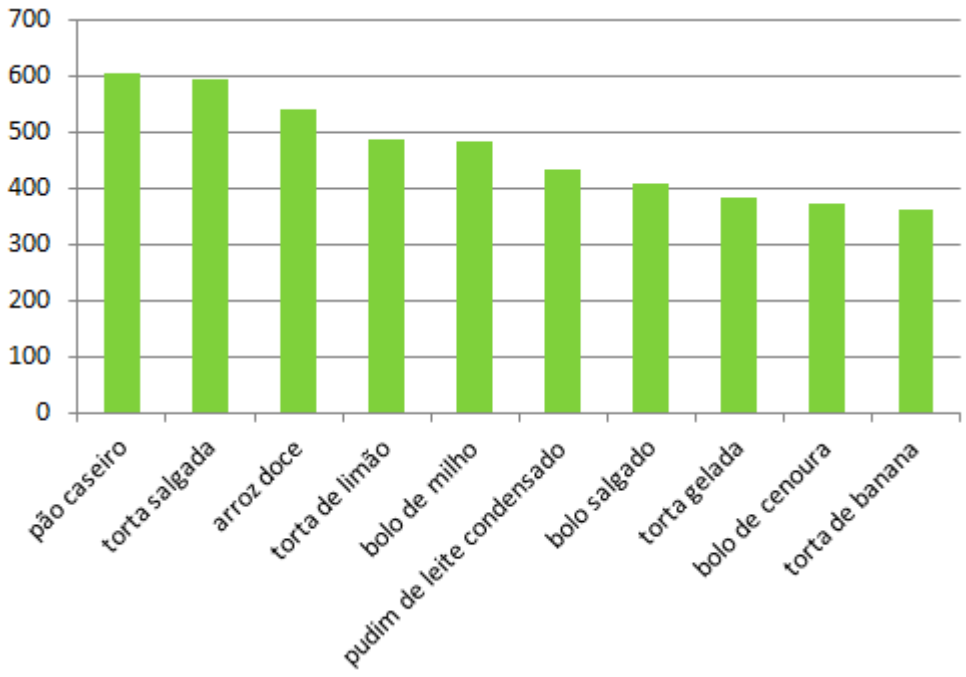


Figura 4.14: Os 10 pratos de nível 2 que apresentam maior número de receitas.

De maneira similar, a Figura 4.15 apresenta os 10 pratos de nível 3 que possuem mais receitas, tendo no eixo X os nomes dos pratos e no eixo Y a quantidade de receitas asso-

ciadas aos pratos. Verifica-se que há um comportamento mais homogêneo. Com exceção do primeiro prato, os demais possuem números de receitas similares, diferentemente do que acontece com os pratos de nível 1.

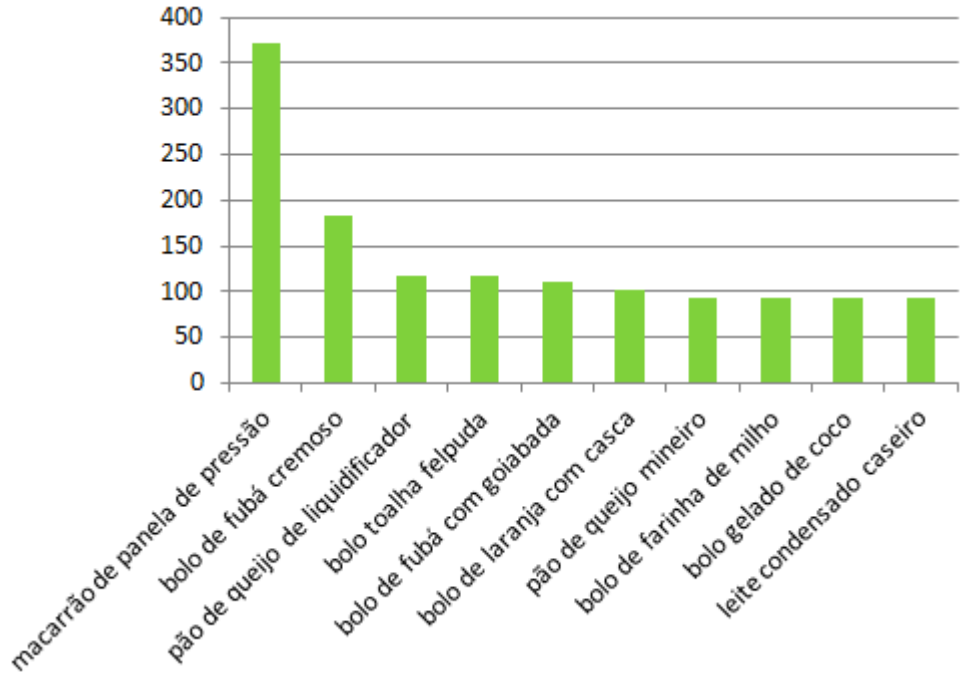


Figura 4.15: Os 10 pratos de nível 3 que apresentam maior número de receitas.

Por fim, chegando aos pratos de nível 4, a Figura 4.16 apresenta os 10 pratos com mais receitas associadas, tendo no eixo X os nomes dos pratos e no eixo Y a quantidade de receitas associadas aos pratos. Visualiza-se que não há uma dissimilaridade com relação às frequências. Entretanto, verifica-se neste nível que o número de receitas associadas aos pratos é baixo. Isso pode ser visto com o prato de nível 4 que apresenta o maior número de receitas, sendo este o "macarrão com carne moída", que possui apenas 13 receitas, entretanto o baixo número de receitas associadas aos pratos de nível 4 é justificável, uma vez que neste nível a especificidade do prato é maior.

Apesar das análises apresentadas anteriormente ilustrarem o comportamento dos principais pratos de cada um dos níveis, elas não evidenciam o comportamento de toda a base de dados de pratos. Dessa forma, a Figura 4.17 apresenta o número de receitas por pratos para cada um dos níveis. A figura apresenta no eixo X cinco possíveis valores representando a quantidade de receitas presentes nos pratos. Visualiza-se que há a presença do valor 1, do valor 2 e depois de 3 intervalos de valores: 3 a 5, 6 a 100 e por fim, acima de 100. Ressalta-se que não há uma proporção nos valores deste eixo devido

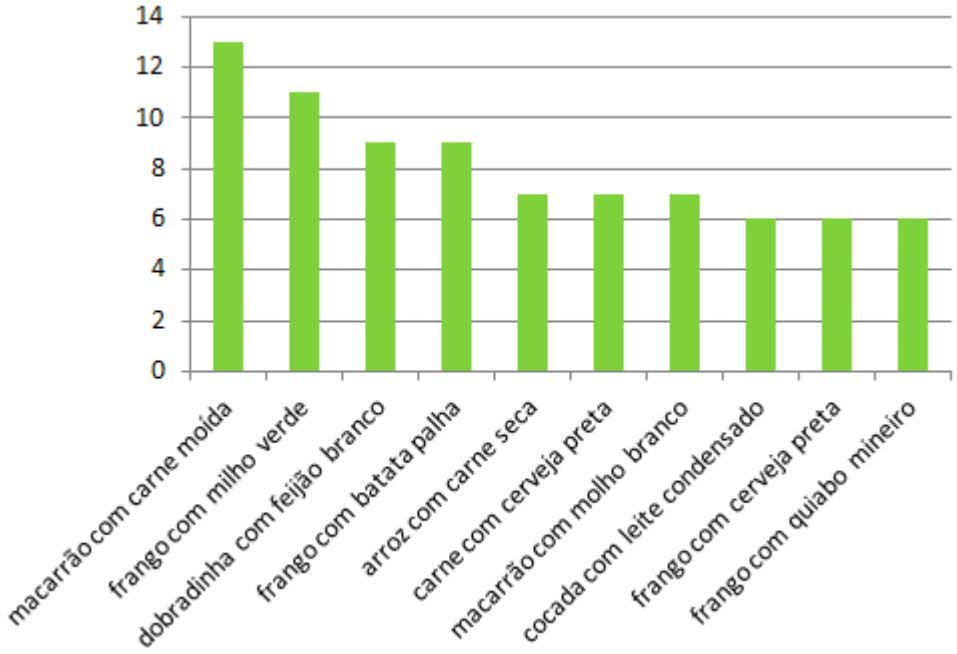


Figura 4.16: Os 10 pratos de nível 4 que apresentam maior número de receitas.

ao fato de que grande parte dos pratos apresentam poucas receitas associadas. Já o eixo Y consiste no número de pratos.

Analisando o gráfico, percebe-se que para todos os níveis, a maioria dos pratos são representados apenas por uma receita, onde se percebe que para os níveis 2 e 3 o número de pratos representado por uma única receita é aproximadamente 50 mil pratos. Percebe-se também um alto número de pratos que são representados apenas por duas receitas. A figura mostra ainda que para o intervalo de 3 a 5, o número de pratos representado por receitas deste intervalo diminui em relação ao número de pratos representado apenas por 2 receitas. Na sequência, o número de pratos representado por receitas entre o intervalo de 6 a 100 é similar ao apresentado no intervalo anterior (3 a 5). Entretanto, percebe-se que no intervalo de 6 a 100, são poucos os pratos de nível 4, valor levemente superior a 10. Finalmente, para a última faixa de receitas de um prato, que é representado por mais de 100 receitas, visualiza-se um decaimento no número de pratos em todos os níveis de pratos, com exceção do nível 4, uma vez que neste nível nenhum prato é associado a mais de 100 receitas.

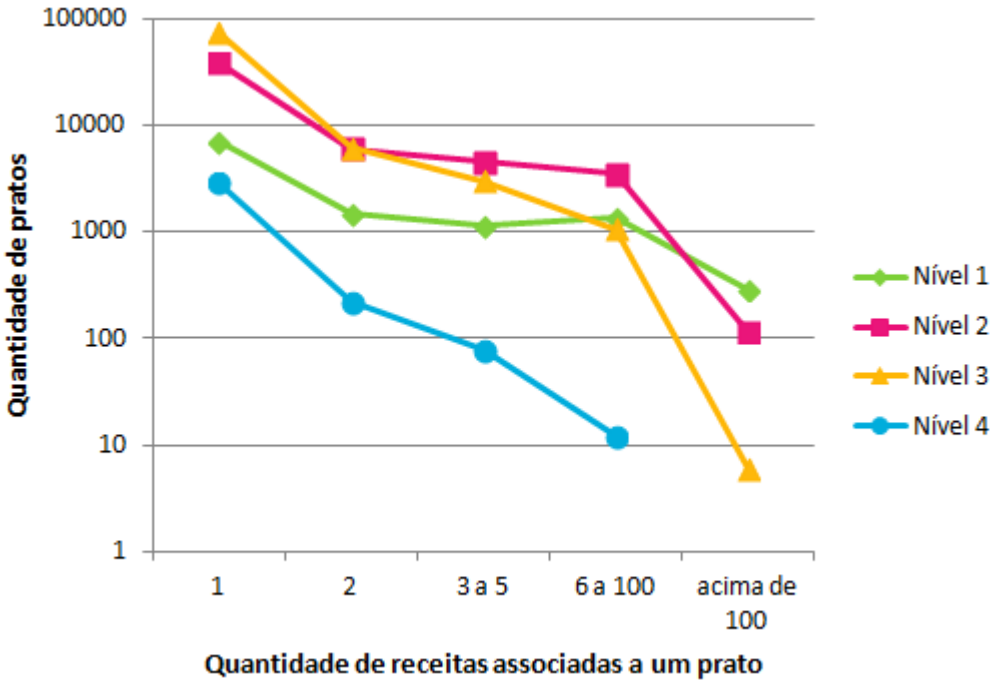


Figura 4.17: Número de receitas associadas aos pratos.

4.2.6 Análise dos conjuntos de ingredientes frequentes gerados

Esta seção apresenta uma análise efetuada sobre os conjuntos de ingredientes frequentes gerados com a utilização do gerador de conjuntos de ingredientes frequentes, que pode ser visualizado na Seção 3.11, onde foram geradas bases de dados para pratos que continham no mínimo duas receitas associadas. Assim, totalizou 29.128 bases de dados, compreendidas em todos os níveis de pratos. Após a geração das bases, deu-se início a identificação dos conjuntos de ingredientes frequentes.

A Tabela 4.9 apresenta a média simples e o desvio padrão do suporte e do tamanho da lista (número de ingredientes presentes no conjunto de ingredientes frequentes). Ressalta-se que estas informações foram obtidas para os conjuntos de ingredientes frequentes gerados para cada um dos níveis de pratos.

Observa-se que a média simples e o desvio padrão para o número de ingredientes é decrescente na hierarquia dos pratos, ou seja, os pratos de nível 1 apresentam maiores valores que os demais níveis de pratos. Diferentemente, nas informações acerca do suporte, os valores são em grande maioria crescentes a partir do prato de nível 1. Isso pode ser justificado, pelo número de receitas presentes nos pratos, levando-se em consideração

Tabela 4.9: Informações sobre os conjuntos de ingredientes frequentes gerados para cada um dos níveis de pratos.

	Tamanho da lista		Suporte	
	Média Simples	Desvio Padrão	Média Simples	Desvio Padrão
Prato N1	2,3592	0,9135	0,5871	0,0816
Prato N2	2,1968	0,5723	0,6495	0,2563
Prato N3	2,0700	0,3314	0,7800	0,2500
Prato N4	2,0300	0,1674	0,7900	0,2600

seus níveis, uma vez que, os pratos de nível 1 apresentam um número maior de receitas associadas, número que tende a diminuir nos pratos de nível 2 e consequentemente de níveis 3 e 4, conforme pode ser visualizado nos gráficos das Figuras 4.13, 4.14, 4.15 e 4.16, apresentados na seção anterior.

De forma similar, realizou-se um estudo, sobre os conjuntos de ingredientes frequentes, não fazendo distinção entre os níveis de pratos. A Tabela 4.10 apresenta os resultados encontrados. Visualiza-se na tabela que há informações sobre o tamanho da lista e suporte dos conjuntos de ingredientes frequentes. Foram utilizadas, além das métricas usadas anteriormente (média simples e desvio padrão), as métricas: valor máximo, valor mínimo e moda. De modo geral, pôde-se observar que os conjuntos de ingredientes frequentes obtiveram um valor alto de suporte (0,6885), o que indica que os conjuntos de ingredientes frequentes são uma boa maneira de identificar os principais ingredientes presentes nas diversas receitas de um prato. Observa-se ainda que a média do tamanho da lista foi de 2,1758, valor que pode ser considerado baixo. Esse comportamento ocorre porque há muito mais combinações de dois ingredientes do que combinações com muitos ingredientes.

Tabela 4.10: Informações sobre os conjuntos de ingredientes frequentes de modo geral, para todas as bases de dados geradas.

	Tamanho da lista	Suporte
Média Simples	2,1758	0,6885
Desvio Padrão	0,5756	0,2645
Valor Máximo	19	1,0000
Valor Mínimo	2	0,0952
Moda	2	1

Capítulo 5

Resultados experimentais

Neste capítulo são apresentados os resultados experimentais de algumas das principais etapas da metodologia de descoberta de conhecimento em receitas gastronômicas. A Seção 5.1 apresenta os resultados da heurística para identificar ingredientes e suas quantidades e unidades de medida. Em seguida, na Seção 5.2 são apresentados os resultados da terceira fase utilizada no processo de extração de ingrediente principal. Por fim, na Seção 5.3 são apresentados os resultados da heurística para encontrar pratos.

5.1 Resultados experimentais da heurística para identificar ingredientes e suas quantidades e unidades de medida

Com o desenvolvimento da heurística para identificar ingredientes e suas quantidades e unidades de medida, viu-se a necessidade de avaliar os resultados encontrados e verificar a eficácia da heurística. Para isso, inicialmente foi efetuada a configuração dos experimentos, onde três métricas foram utilizadas: precisão, revocação e F1. De acordo com Baeza-Yates and Ribeiro-Neto (2011), precisão é a fração dos documentos recuperados que são relevantes, conforme apresentado pela Equação 5.1, onde *tp* (*true positive*) consiste em itens relevantes que foram retornados e *fp* (*false positive*) representa os itens não relevantes que foram retornados erroneamente.

$$p = \frac{tp}{tp + fp} \tag{5.1}$$

Já a revocação, é a fração dos documentos relevantes que foram retornados, conforme apresentado pela Equação 5.2, onde *tp* (*true positive*) consiste em itens relevantes que foram retornados, e *fn* (*false negative*) representa os itens relevantes que, erroneamente, não foram retornados.

$$r = \frac{tp}{tp + fn} \tag{5.2}$$

Finalmente, a métrica F1 é a combinação da precisão com a revocação com o intuito de obter um valor balanceado que leve em consideração ambas as métricas, conforme verifica-se na Equação 5.3, onde p representa o valor da precisão e r representa o valor da revocação.

$$F1 = 2 \times \frac{1}{\frac{1}{p} + \frac{1}{r}} \tag{5.3}$$

Após a definição das métricas a serem utilizadas na experimentação, viu-se a necessidade de efetuar amostragens, uma vez que o número alto de receitas inviabilizava a avaliação para toda a base de dados. Desta forma, duas amostragens foram realizadas. Ressalta-se que se fez necessário o uso de duas amostragens porque as receitas da primeira amostragem foram utilizadas na fase de implementação, onde se buscou identificar os padrões das receitas para elaboração da heurística. Neste ponto, as receitas desta amostragem foram utilizadas de forma a prover o conhecimento necessário para a construção da heurística. Diferentemente, as receitas que compõem a segunda amostragem não foram utilizadas na fase de implementação, sendo utilizadas apenas para a avaliação da eficácia da heurística. As amostras são compostas por 50 receitas selecionadas aleatoriamente, sendo 10 de cada uma das fontes de dados.

Cada receita tem suas especificidades e entre essas, encontra-se o número de sentenças onde se encontram os ingredientes, quantidades e unidades de medida. Apesar do número de receitas que compõem as amostragens serem iguais, o número de sentenças

nas amostragens são diferentes. A Tabela 5.1 apresenta a quantidade de sentenças em cada uma das fontes de dados e também em um contexto geral, para as duas amostragens. Apesar de terem sido usadas apenas 50 receitas em cada uma das amostras, o número de sentenças fica na ordem de 450 sentenças, o que é um valor representativo para a elaboração e avaliação da heurística.

Tabela 5.1: Número de sentenças que compõem as amostragens para cada fonte de dados e no total.

	Primeira Amostragem	Segunda Amostragem
Cybercook	77	82
Dieta e Receitas	75	49
Edu Guedes	140	129
Receitas.com	76	130
Tudo Gostoso	79	67
Total	447	457

Após a realização das amostragens, deu-se início a análise da eficácia da heurística. Na Tabela 5.2, são expostos os resultados encontrados. Analisam-se nos resultados obtidos, que os valores encontrados atingiram mais de 97% de precisão e mais de 99% de revocação, obtendo consequentemente F1 acima de 98% para todas as informações extraídas (ingredientes, quantidades e unidades de medida). Entretanto, ressalta-se que esses valores foram obtidos a partir da execução da heurística para a amostra a qual suas receitas foram utilizadas na detecção dos padrões, o que pode influenciar positivamente nos resultados. Por sua vez, a Tabela 5.3 apresenta os resultados da heurística para a segunda amostragem.

Tabela 5.2: Execução da heurística usando a primeira amostragem.

	Ingr.	Quant.	Un. Med.
Precisão	97,01%	98,60%	97,04%
Revocação	100,00%	99,80%	99,17%
F1	98,48%	99,19%	98,09%

Tabela 5.3: Execução da heurística usando a segunda amostragem.

	Ingr.	Quant.	Un. Med.
Precisão	97,03%	99,07%	95,68%
Revocação	99,78%	98,84%	99,68%
F1	98,39%	98,95%	97,64%

Observa-se nos resultados obtidos que, apesar destas receitas não terem sido utilizadas como objeto de estudo na detecção dos padrões, os resultados não ficaram muito abaixo em relação aos resultados da primeira amostragem. A precisão foi superior a 95% para a unidade de medida, aproximadamente 97% para os ingredientes e chegando a mais de 99% para a quantidade. A revocação aproximou-se de 100% tanto para os ingredientes quanto para as unidades de medida; já para as quantidades aproximou-se de 99%. Por fim, a F1 ficou acima de 97% em todos os casos.

Em suma, constata-se que os resultados obtidos com a realização da heurística são considerados bons para o presente trabalho, uma vez que é aceitável ter uma pequena porcentagem de ruído nas informações encontradas.

5.2 Resultados experimentais da terceira fase da extração de ingrediente principal

O processo de extração de ingrediente principal, conforme pode se verificar na Seção 3.5, é dividido em três fases. A terceira fase consiste na utilização dos ingredientes principais já identificados, buscando associá-los às sentenças de ingredientes que ainda não possuíam associação aos ingredientes principais. Esta fase trabalha com o conceito de janelamento, onde para cada sentença de ingrediente quebra-se a *string* em cada espaço e procura-se, por meio do janelamento, associar esta sentença de ingrediente a um ingrediente principal. Após a execução desse procedimento, viu-se a necessidade de verificar sua eficácia, utilizando-se da métrica acurácia, que segundo Manning et al. (2008) consiste na fração de decisões (relevante/não relevante) que estão corretos, conforme apresentado pela Equação 5.4.

$$a = \frac{tp + tn}{tp + fp + fn + tn}, \tag{5.4}$$

onde *tp* (*true positive*) consiste em itens relevantes que foram retornados, *tn* (*true negative*) representa os itens relevantes que não foram retornados, *fp* (*false positive*) representa os itens não relevantes que foram retornados erroneamente, e por fim, *fn* (*false negative*) consiste nos itens relevantes que, erroneamente, não foram retornados.

Para esta experimentação também fez-se necessário efetuar uma amostragem. Dessa forma, utilizou-se de uma amostra contendo 500 sentenças de ingredientes selecionadas aleatoriamente. Em seguida, analisou-se para cada uma destas sentenças de ingredientes se o ingrediente principal associado à mesma, realmente era o ingrediente principal que deveria ser associado. Finalmente, após a execução sobre a amostra, verificou-se que a terceira fase utilizada na extração de ingrediente principal obteve acurácia de 88,4%.

Salienta-se que o processo de extração de ingrediente principal (Seção 3.5) em uma de suas fases, utiliza-se de processamento manual na identificação dos ingredientes, priorizando as sentenças de ingredientes com maior frequência de ocorrência. Desta forma, o valor da acurácia encontrada na terceira fase do extrator de ingredientes principal é considerado razoável e suficiente para várias aplicações, incluindo a proposta deste trabalho. No entanto, caso uma acurácia maior seja necessária, a avaliação manual de mais ingredientes levaria a um aumento natural da acurácia.

5.3 Resultados experimentais para a heurística para encontrar pratos

Com o desenvolvimento da heurística para encontrar pratos, viu-se a necessidade de avaliar os resultados encontrados e verificar a eficácia da heurística. A métrica utilizada também foi a acurácia, apresentada na Equação 5.4.

Para a avaliação, duas amostragens foram geradas. Ressalta-se que fez-se necessário duas amostragens, uma vez que as receitas da primeira amostragem foram utilizadas na fase de implementação, onde se buscou identificar os padrões das receitas para a elaboração da heurística. Neste ponto, as receitas desta amostragem foram utilizadas de forma a prover o conhecimento necessário para a construção da heurística. Diferen-

temente, as receitas que compõem a segunda amostragem não foram utilizadas na fase de implementação. A primeira amostragem é composta por 700 receitas, ao passo que a segunda amostragem possui 100 receitas. Salienta-se que a diferença na quantidade de receitas entre as amostragens, se dá devido ao fato de que as receitas utilizadas na primeira amostragem foram utilizadas também na identificação dos padrões e, consequentemente, permitindo a criação da heurística.

Após a realização das amostragens, deu-se início à análise da eficácia da heurística. A princípio, utilizou-se da primeira amostragem realizada. A Tabela 5.4, apresenta os resultados encontrados. Analisa-se nos resultados obtidos, que em todas as receitas há a presença do prato de nível 1, apresentando acurácia acima de 98%. O prato de nível 2 pode ser caracterizado em 662 receitas, com acurácia próxima de 97%. O prato de nível 3 verifica-se em 337 receitas com acurácia acima de 98%. Finalmente, o prato de nível 4 ocorre em apenas 12 receitas, com acurácia de 100%. Entretanto, esses valores foram obtidos a partir da execução da heurística para a amostra a qual suas receitas foram utilizadas na detecção dos padrões, o que pode influenciar positivamente nos resultados.

Tabela 5.4: Exucução da heurística para encontrar pratos usando a primeira amostragem.

	Prato N1	Prato N2	Prato N3	Prato N4
# Receitas	700	662	337	12
Acurácia	98,57%	96,53%	98,52%	100,00%

Tabela 5.5: Exucução da heurística para encontrar pratos usando a segunda amostragem.

	Prato N1	Prato N2	Prato N3
# Receitas	100	94	47
Acurácia	95,00%	90,43%	93,62%

Por fim, avaliou-se a heurística de descoberta de pratos para a segunda amostragem. A Tabela 5.5 apresenta os resultados. Observa-se que em todas as receitas usadas nessa amostragem, há a presença do prato de nível 1, com acurácia de 95%. Observa-se ainda que o prato de nível 2 ocorre em 94 receitas, apresentando acurácia acima de 90%. Já o prato de nível 3 ocorre em 47 receitas, apresentando acurácia acima de 93%.

Finalmente, cabe destacar que na Tabela 5.5 não foi apresentada a porcentagem de acerto da heurística para o prato de nível 4. Isso ocorre, uma vez que nas receitas usadas nesta amostragem, apenas duas, possuiam pratos de nível 4. Assim, devido a pouca relevância do número de receitas, optou-se por não verificar a acurácia para o prato de nível 4.

Capítulo 6

Conclusões e Trabalhos Futuros

Com o surgimento da *Web* e sua difusão mundial, verifica-se que inúmeras aplicações surgiram, entre elas, encontram-se os sistemas colaborativos. Uma classe de serviços colaborativos pode ser representada por meio de *sites* de compartilhamento de receitas gastronômicas, que permite a troca de experiências gastronômicas, seja por meio de outras receitas de um mesmo prato, ou mediante aos comentários realizados, estabelecendo-se assim, um ambiente rico em novas experiências culinárias, com diversificadas receitas, cada uma com suas particularidades.

Apesar do crescimento dos *sites* de compartilhamento de receitas existentes atualmente, visualiza-se alguns pontos em que estes serviços poderiam ser melhores, como por exemplo: a maneira fechada com que apresentam as receitas, não permitindo adaptações, bem como oferecer receitas de baixa qualidade ao usuário, ou mesmo apresentar geralmente as receitas mais populares. Assim, percebendo estas características e a necessidade de ofertar um serviço que as satisfaça, este trabalho tem como hipótese, que efetuar um estudo sobre os *sites* de receitas gastronômicas em busca de descobrir conhecimentos sobre pratos pode ser útil ao usuário, tendo em vista que ele pode obter receitas que estejam mais alinhadas às suas necessidades ou preferências, bem como ter maior autonomia no processo de cozinhar. Além disso, isto configura-se como um passo inicial para a *Web* Semântica.

A metodologia desenvolvida para a descoberta de conhecimentos em receitas gastronômicas utiliza-se de dados coletados de *sites* de compartilhamento de receitas gastronômicas. Utilizou-se informações sobre os ingredientes, quantidades e unidades de medida, instruções de preparo, e informações acerca da popularidade das receitas, como

número de curtidas, tweets, recomendações, avaliação das receitas, entre outras. Para efetuar a separação das informações: ingredientes, quantidades e unidades de medida, foi desenvolvida uma heurística baseada nos padrões identificados nas receitas. Uma heurística também foi desenvolvida para efetuar a associação entre uma receita e possíveis pratos associados, como parte do processo de descoberta de conhecimento sobre pratos. Para possibilitar a identificação dos ingredientes frequentes em um prato, utilizou-se da tarefa de mineração de dados: associação de regras de associação. De forma a facilitar o processo de consulta, criou-se um índice invertido de ingredientes em relação a pratos.

As fontes de dados consideradas na construção da metodologia foram: Tudo Gostoso, Cybercook, Receitas.com, Edu Guedes e Dieta e Receitas. No entanto, cabe salientar que, embora tenham sido utilizadas estas cinco fontes de dados, a metodologia pode ser considerada escalável quando os dados de receitas (sentenças de ingredientes e instruções de preparo) possuírem estrutura similar ao que se vê nas fontes trabalhadas. Apesar de ter utilizado-se de todas as fontes de dados na fase de treino, verifica-se que outras fontes com estruturas similares podem ser processadas pela metodologia proposta.

O objetivo principal do trabalho foi descobrir conhecimento sobre pratos gastronômicos, utilizando-se de informações de receitas e seus ingredientes, quantidades e unidades de medida, bem como suas formas de preparo, além de outras informações associadas aos pratos. Com a metodologia desenvolvida, pôde-se chegar a esse objetivo, uma vez que tal metodologia foi corroborada por resultados satisfatórios obtidos através das experimentações realizadas em etapas importantes da metodologia, atingindo um dos objetivos específicos, referente a avaliação dos principais métodos da metodologia. Outro ponto relevante nesse contexto refere-se ao estudo de caracterização efetuado sobre os dados coletados das diferentes fontes de dados, foi possível realizar diversas análises sobre os dados coletados e sobre o conhecimento descoberto.

Com o desenvolvimento da metodologia, possibilitou-se ainda atingir mais um dos objetivos específicos, que foi projetar um serviço de busca, seleção e visualização de receitas gastronômicas utilizando-se do conhecimento descoberto. As projeções de busca, seleção e de visualização das receitas foram divididas em três etapas. A primeira permite que o usuário busque por receitas de um dado prato, levando-se em consideração o conhecimento descoberto sobre os pratos, como os ingredientes mais frequentes (principais). A segunda maneira permite que o usuário busque uma receita associada a um determinado prato, entretanto, mais do que isso, permite que ele manipule os ingredientes presentes nas receitas, utilizando-se do conhecimento coletivo. Finalmente, a terceira opção oferece ao usuário a possibilidade de encontrar receitas mediante aos ingredien-

tes que o usuário possui em casa ou deseja colocar na receita. Para isso, inicialmente escolhe-se uma categoria e um prato associado à categoria e em seguida seleciona-se os ingredientes desejados.

Vale salientar que foi possível também criar uma taxonomia de ingredientes usados em pratos da gastronomia brasileira, outro objetivo específico alcançado, uma vez que com o desenvolvimento da heurística para identificar ingrediente, quantidade e unidade de medida em uma sentença de ingrediente, e em seguida, por meio da identificação do ingrediente principal, estabeleceu-se uma base de dados de ingredientes rica da culinária brasileira, tendo em vista que foram utilizadas aproximadamente 238 mil receitas gastronômicas coletadas de diversificadas fontes de dados de receitas. Pretende-se em breve, disponibilizar esta base tornando-a pública.

Este trabalho utilizou-se de alguns *softwares* para sua realização, bem como alguns algoritmos foram desenvolvidos para implementar a metodologia proposta. Foram utilizados o *Crawler4j* para coleta das receitas, o *R Project* para a utilização do algoritmo Eclat, o serviço de busca do Google para efetuar as correções ortográficas nos nomes das receitas, o *Solr* para construção do índice invertido, e o EasyFit para a geração das funções PDF nas análises dos dados. A parte das tarefas listadas anteriormente, todos os outros componentes foram desenvolvidos excepcionalmente para esta dissertação. Foram desenvolvidas duas heurísticas: uma para a identificação de ingredientes e suas respectivas quantidades e unidades de medida, e outra para a identificação dos pratos e seus níveis. Foram desenvolvidos ainda, um algoritmo para efetuar a identificação de candidatos válidos de unidades de medida, um algoritmo para efetuar a validação das receitas e um algoritmo para identificar os ingredientes principais. Vale salientar que todos os algoritmos propostos e desenvolvidos neste trabalho apresentaram resultados satisfatórios, inclusive, os principais tiveram sua eficácia avaliada.

Cabe salientar que apesar deste trabalho possuir algumas etapas realizadas manualmente, estas etapas tiveram um amplo esforço e assim foi possível prover uma base de dados de qualidade para que se desse o desenvolvimento da metodologia. Assim, em uma possível expansão no uso de fontes de dados, o trabalho manual é praticamente desnecessário, já que a base de dados existente atualmente é capaz de executar todas as etapas da metodologia com novos dados de entrada.

Como trabalhos futuros, há várias sugestões que são apresentadas a seguir:

- Desenvolver um aplicativo para permitir que o usuário efetue consultas, utilizando-

se de todas as maneiras possíveis para estabelecer uma consulta, permitindo filtros de acordo com o modo de preparo, ou mesmo pelas categorias dos pratos.

- Estender o uso de fontes de dados, coletando dados de outras fontes, abrangendo assim o número de receitas e consequentemente ofertando mais possibilidades ao usuário. A recoleta também é importante para que novas receitas das fontes já coletadas possam ser inseridas.

- Analisar a possibilidade de criar outra maneira de classificação de pratos, utilizando-se de alguma técnica de aprendizado de máquina, com o intuito de estabelecer uma comparação entre a heurística utilizada na metodologia, visando avaliar a forma mais eficiente a fim de melhorar a metodologia.

- Identificar uma base de dados de calorias dos ingredientes e propor receitas que sejam caloricamente mais saudáveis, ou mesmo a substituição da forma de preparo ou do ingrediente para opções mais saudáveis.

- Pesquisar sobre ingredientes que contenham lactose/glúten e identificá-los, com o objetivo de viabilizar receitas que não se utilizem de lactose/glúten, visando atender de forma especial necessidades de usuários com restrições alimentares, considerando a existência de um aplicativo no futuro.

- Desenvolver uma forma de classificação das categorias dos pratos, de forma a estabelecer uma associação mais precisa de receitas às suas possíveis categorias.

- Efetuar um trabalho com os comentários coletados das receitas, a fim de obter conhecimento sobre estes dados que constitui uma importante plataforma de conhecimento coletivo.

Referências Bibliográficas

Agrawal, R., Imieliński, T. and Swami, A. (1993). Mining association rules between sets of items in large databases, *ACM SIGMOD Record*, Vol. 22, ACM, pp. 207–216.

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

Badra, F., Bendaoud, R., Bentebibel, R., Champin, P.-A., Cojan, J., Cordier, A., Després, S., Jean-Daubias, S., Lieber, J., Meilender, T. et al. (2008). Taaable: Text mining, ontology engineering, and hierarchical classification for textual case-based cooking, *9th European Conference on Case-Based Reasoning-ECCBR 2008, Workshop Proceedings*, pp. 219–228.

Badra, F., Cojan, J., Cordier, A., Lieber, J., Meilender, T., Mille, A., Molli, P., Nauer, E., Napoli, A., Skaf-Molli, H. et al. (2009). Knowledge acquisition and discovery for the textual case-based cooking system wikitaaable, *8th International Conference on Case-Based Reasoning-ICCBR 2009, Workshop Proceedings*, pp. 249–258.

Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011). *Modern Information Retrieval - the concepts and technology behind search, Second edition*, Pearson Education Ltd., Harlow, England.

Belda, F. R. and Gamonar, F. D. O. (2014). Proposta de uma rede social como ambiente de convergência com programas de gastronomia e culinária na tv, *Ciência & Desenvolvimento-Revista Eletrônica da FAINOR* **7**(2).

Blansché, A., Cojan, J., Dufour-Lussier, V., Lieber, J., Molli, P., Nauer, E., Skaf-Molli, H. and Toussaint, Y. (2010). Taaable 3: Adaptation of ingredient quantities and of textual preparations, *18th Int. Conf. on Case-Based Reasoning Workshop Procs*, Citeseer, pp. 189–198.

Bridge, D. and Larkin, H. (2014). Creating new sandwiches from old, *Workshop Programme of the 22nd International Conference on Case-Based Reasoning,(this volume)*.

Camilo, C. O. and Silva, J. C. d. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas, *Goiânia: Universidade Federal de Goiás* .

Cojan, J., Dufour-Lussier, V., Gaillard, E., Lieber, J., Nauer, E. and Toussaint, Y. (2011). Taaable 4: knowledge extraction for improving case retrieval and recipe adaptation, *Proceedings of the Computer Cooking Contest* pp. 197–206.

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

Freyne, J. and Berkovsky, S. (2010). Intelligent food planning: personalized recipe recommendation, *Proceedings of the 15th international conference on Intelligent user interfaces*, ACM, pp. 321–324.

Gamonar, F. D. O. and Brasil, F. R. B. (2015). Da culinária de papel às mídias sociais de nicho: Planejando o desenvolvimento de um ambiente colaborativo para a publicação de receitas e dicas culinárias, *Razón y Palabra* **19**(89).

Geleijnse, G., Nachtigall, P., van Kaam, P. and Wijgergangs, L. (2011). A personalized recipe advice system to promote healthful choices, *Proceedings of the 16th international conference on Intelligent user interfaces*, ACM, pp. 437–438.

Gospodnetic, O. and Hatcher, E. (2005). *Lucene*, Manning.

Grainger, T., Potter, T. and Seeley, Y. (2014). *Solr in action*, Manning.

Hahsler, M., Grün, B. and Hornik, K. (2005). A computational environment for mining association rules and frequent item sets.

Hatcher, E., Gospodnetic, O. and McCandless, M. (2004). Lucene in action.

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

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

Kuć, R. (2013). *Apache Solr 4 Cookbook*, Packt Publishing Ltd.

Larkin, H. and Bridge, D. (2014). Subs and sandwiches: Replacing one ingredient by another, *Workshop Programme of the 22nd International Conference on Case-Based Reasoning,(this volume)*.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals, *Soviet physics doklady*, Vol. 10, pp. 707–710.

Lopes, I. L. (2004). Novos paradigmas para avaliação da qualidade da informação em saúde recuperada na web, *Ciência da Informação* **33**(1): 81–90.

Manning, C. D., Raghavan, P. and Schütze, H. (2008). *Introduction to information retrieval*, Vol. 1, Cambridge university press Cambridge.

Mino, Y. and Kobayashi, I. (2009). Recipe recommendation for a diet considering a user's schedule and the balance of nourishment, *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, Vol. 3, IEEE, pp. 383–387.

Monteiro, L. (2001). A internet como meio de comunicação: possibilidades e limitações, *XXVI Congresso Brasileiro de Ciências da Comunicação, Campo Grande/MS, Setembro de*.

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

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

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

Sichieri, R., Coitinho, D. C., Monteiro, J. B. and Coutinho, W. F. (2000). Recomendações de alimentação e nutrição saudável para a população brasileira, *Arquivos Brasileiros de Endocrinologia & Metabologia* **44**(3): 227–232.

Sigurbjörnsson, B. and Van Zwol, R. (2008). Flickr tag recommendation based on collective knowledge, *Proceedings of the 17th international conference on World Wide Web*, ACM, pp. 327–336.

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

Svensson, M., Höök, K. and Cöster, R. (2005). Designing and evaluating kalas: A social navigation system for food recipes, *ACM Transactions on Computer-Human Interaction (TOCHI)* **12**(3): 374–400.

Teng, C.-Y., Lin, Y.-R. and Adamic, L. A. (2012). Recipe recommendation using ingredient networks, *Proceedings of the 4th Annual ACM Web Science Conference*, ACM, pp. 298–307.

Trevisiol, M., Chiarandini, L. and Baeza-Yates, R. (2014). Buon appetito: recommending personalized menus, *Proceedings of the 25th ACM conference on Hypertext and social media*, ACM, pp. 327–329.

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

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

Wagner, J., Geleijnse, G. and van Halteren, A. (2011). Guidance and support for healthy food preparation in an augmented kitchen, *Proceedings of the 2011 Workshop on Context-awareness in Retrieval and Recommendation*, ACM, pp. 47–50.

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

Zaki, M. J., Parthasarathy, S., Ogihara, M., Li, W. et al. (1997). New algorithms for fast discovery of association rules., *KDD*, Vol. 97, pp. 283–286.

Zhang, Q., Hu, R., Mac Namee, B. and Delany, S. J. (2008). Back to the future: Knowledge light case base cookery, *Conference papers*, p. 15.

Zobel, J., Moffat, A. and Ramamohanarao, K. (1998). Inverted files versus signature files for text indexing, *ACM Transactions on Database Systems (TODS)* **23**(4): 453–490.

Apêndice A - Exemplos de telas e requisitos para a aplicação futura

Com o intuito de elucidar algumas das funcionalidades pretendidas na aplicação em que se deseja criar, como motivação para esta pesquisa, as Figuras 1, 2 e 3 apresentam algumas telas que compõem o primeiro protótipo da aplicação. Cada uma das figuras apresentam duas telas. A Figura 1 ilustra na primeira tela a possibilidade de o usuário escolher entre as opções de busca. Pode-se visualizar que há duas opções: busca por um prato e busca por ingredientes. No entanto, o usuário pode entrar tanto com o nome de um prato quanto com uma lista de ingredientes, ou mesmo preencher ambos os campos. A segunda tela ilustra o preenchimento do campo de busca por meio de um prato.

A Figura 2 apresenta ao usuário em sua primeira tela os ingredientes principais para o preparo do prato escolhido. Verifica-se que se apresenta na tela a informação de porcentagem de receitas do prato desejado que possuem os ingredientes na cor verde. Apresenta-se também a quantidade total de receitas para o prato. Em relação às cores, verifica-se que há ingredientes em três tonalidades de cor: verde, amarelo e vermelho. O intuito dessa diferenciação de cores consiste em mostrar ao usuário que ingredientes na cor verde são comuns para esse prato além de representar um ingrediente que ele tenha escolhido explicitamente. A cor amarela mostra que os ingredientes podem estar presentes nas receitas do prato escolhido, ou seja, ingredientes que são recomendados. Por fim, a cor vermelha explicita que os ingredientes não são comumente utilizados no preparo do prato escolhido, mesmo tendo receitas que usaram esse ingrediente para o prato pesquisado. O usuário tem a possibilidade de remover ou adicionar um ingrediente em sua busca.

A segunda tela apresentada na Figura 2 mostra que um ingrediente que estava na cor verde pode ter sua cor alterada, o que significa que ele está removendo o ingrediente da lista desejada (passando para vermelho) ou dando a opção do ingrediente poder estar

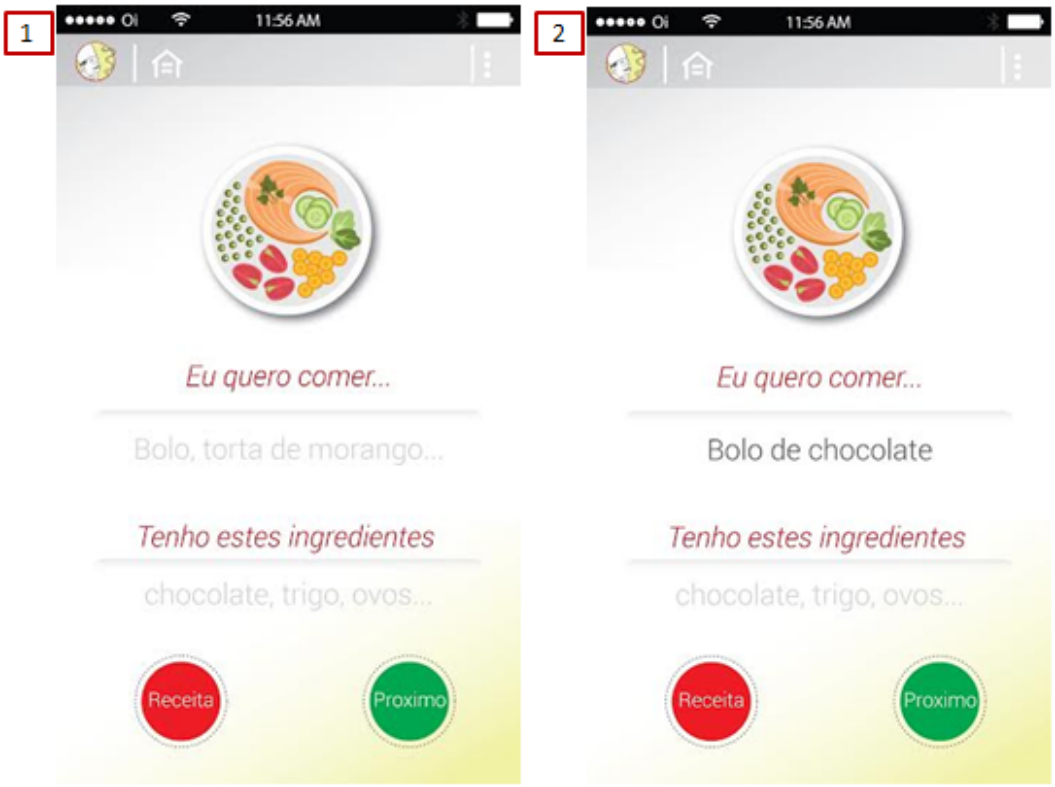


Figura 1: Primeiro protótipo do aplicativo - Telas de busca.

na receita, mas não necessariamente estar (passando para amarelo). Verifica-se também que as informações acerca da quantidade de receitas de um dado prato e a porcentagem variam de acordo com as adaptações realizadas. Ressalta-se que neste protótipo os dados sobre as quantidades e porcentagens são ilustrativos.

Finalmente, as duas telas apresentadas na Figura 3 ilustram como se dá o procedimento de troca de cor. Para isso, basta apenas um clique sobre o ingrediente para que ele ofereça a opção de outra cor, e com mais um clique o usuário habilita a cor escolhida para o ingrediente determinado.

Ressalta-se ainda que as telas oferecem a opção do usuário pedir logo uma receita, sem adaptá-la, e permite também que ele possa fazer as adaptações conforme suas necessidades ou preferências. As Figuras 1, 2 e 3 não apresentam todas as possíveis telas. Uma tela faltante nas figuras é a que permite que o usuário escolha uma receita por meio do modo de preparo. Outra tela que não está presente nas figuras é a tela onde são apresentadas as informações sobre a receita escolhida para um dado prato. Há uma série de funcionalidades que podem estar presentes na aplicação que não estão ilustradas

Figura 2: Primeiro protótipo do aplicativo - Telas que apresentam os ingredientes.

nas figuras.

A seguir são apresentados alguns requisitos funcionais do sistema a fim de permitir um maior entendimento sobre o mesmo.

- RF1. O sistema deve oferecer ao usuário a possibilidade de buscar receitas de um determinado prato.
- RF2. O sistema deve oferecer ao usuário a possibilidade de buscar receitas mediante a escolha de um prato de seu gosto bem como escolha dos ingredientes que devem estar presentes nas receitas.
- RF3. O sistema deve oferecer ao usuário a possibilidade de entrar apenas com ingredientes que o mesmo possui em casa, ou ingredientes que sejam de sua preferência culinária, e logo recomendar pratos que contenham tais ingredientes.
- RF4. O sistema deve permitir que o usuário escolha receitas mediante a forma de preparo das receitas (assado, cozido, frito, refogado e cru).

Figura 3: Primeiro protótipo do aplicativo - Telas que mostram a possibilidade de troca de cor.

- RF5. O sistema deve fornecer ao usuário informações que elucidam a porcentagem de receitas que possuem os ingredientes desejados, de acordo com a combinação desejada de ingrediente.

- RF6. O sistema deve permitir que o usuário manipule os ingredientes que devem estar presentes ou ausentes em uma receita de maneira fácil.

- RF7. O sistema deve oferecer um *ranking* de receitas como resultado ao usuário, após a busca realizada.

- RF8. O sistema deve oferecer ao usuário receitas de todas as fontes de dados presentes (Cybercook, Dieta e Receitas, Edu Guedes, Receitas.com e Tudo Gostoso), quando isto for possível.

- RF9. O sistema deve ser capaz de recomendar receitas que contenham os ingredientes similares aos desejados pelo usuário, quando não houver nenhuma receita com todos os ingredientes desejados.

- RF10. O sistema deve oferecer ao usuário a possibilidade de buscar por uma receita por meio das várias categorias de receitas existentes.
- RF11. O sistema deve oferecer ao usuário a possibilidade de logo após escolher um prato ou os ingredientes, receber o *ranking* de receitas associadas a busca feita, ou ainda permitir que ele faça adaptações antes de solicitá-las.
- RF12. O sistema deve oferecer ao usuário a opção de refazer uma busca quando a anterior não possuir resultados.
- RF13. O sistema deve apresentar ao usuário receitas de um prato levando em consideração o conhecimento coletivo acerca do prato buscado pelo usuário.