UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

Estruturação de uma Folksonomia: Um Estudo em uma Comunidade de Perguntas e Respostas

Paulo Diogo Rodrigues Leão

Orientador

Sean Wolfgand Matsui Siqueira

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2018

Estruturação de uma Folksonomia: Um Estudo em uma Comunidade de Perguntas e Respostas

Paulo Diogo Rodrigues Leão

DISSERTAÇÃO APRESENTADA COMO REQUISITO PARCIAL PARA OBTENÇÃO DO TÍTULO DE MESTRE PELO PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA DA UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO (UNIRIO). APROVADA PELA COMISSÃO EXAMINADORA ABAIXO ASSINADA.

Aprovada por:

Sean Wolfgand Matsui Siqueira, D.Sc. - UNIRIO

Bernardo Pereira Nunes, D.Sc. - UNIRIO

Jairo Francisco de Souza, D.Sc. - UFJF

RIO DE JANEIRO, RJ - BRASIL

SETEMBRO de 2018

Catalogação informatizada pelo(a) autor(a)

Leão, Paulo Diogo Rodrigues
L433 Estruturação de uma Folksonomia: Um Estudo em uma Comunidade de Perguntas e Respostas / Paulo Diogo Rodrigues Leão. -- Rio de Janeiro, 2018.
62 p

Orientador: Sean Wolfgand Matsui Siqueira.
Dissertação (Mestrado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2018.

1. Taxonomia. 2. Folksonomia. 3. Coerência semântica. 4. Algoritmo. 5. Árvore de tags. I. Wolfgand Matsui Siqueira, Sean, orient. II. Título.

A todos os pagadores de impostos do Brasil e todas as bandas de heavy metal do planeta terra.

Agradecimentos

Agradeço ao professor Sean pelo acompanhamento, generosidade e competência intelectual. Foram muitos e-mails trocados e muitas conversas que só abriram a minha mente para a pesquisa cientifica. Através de seus ensinamentos que entendi como conduzir uma pesquisa.

Aos colegas Crystiam Kelle, Davi Duarte, Jerry Medeiros, Jansen Oliveira, Marcelo Tibau e Thiago Procaci pelo apoio que me deram durante essa jornada, seja revisando textos, conversando sobre a pesquisa ou assistindo apresentações.

Aos professores do PPGI que tão bem me prepararam para a realização desta dissertação. Aos funcionários da secretaria do PPGI que sempre muito bem me atenderam quando precisei.

Às centenas de pesquisadores espalhados pelo mundo, cujas ideias, artigos, livros, vídeos, palestras, cursos permitiram a existência do presente trabalho.

Ao Instituto Brasileiro de Geografia e Estatística (IBGE), por me liberar alguns dias para que eu pudesse me dedicar mais ao mestrado.

Leão, Paulo Diogo Rodrigues. **Estruturação de uma Folksonomia: Um Estudo em uma Comunidade de Perguntas e Respostas**. UNIRIO, 2018. 62 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Comunidades de conhecimento são compostas por indivíduos que compartilham os mesmos interesses e voluntariamente trabalham juntos para expandir conhecimento e entendimento de um domínio através de aprendizagem e compartilhamento. Quando um usuário faz uma pergunta em uma comunidade de conhecimento, geralmente ele pode adicionar *tags* para classificar a pergunta. Essas *tags* facilitam aos especialistas encontrar aquela pergunta e então respondê-la. Como são próprios usuários que as criam, um dos problemas dessas *tags* está no fato de que não existe uma hierarquia para auxiliar na busca por temas mais abrangentes. Neste trabalho é proposta uma forma de hierarquizar a folksonomia de uma comunidade online de conhecimento. Para tal estudo, foi utilizado o *dataset* da comunidade de Biologia da Stack Exchange, visto que buscar perguntas referentes a um tema é um problema vivenciado pelos seus usuários. Essa hierarquização proposta pode ajudar os usuários na busca de perguntas, seja por temas específicos, seja por assuntos mais abrangentes. Foi possível mapear 462 (65%) das 703 *tags* da comunidade. Das *tags* que não foram encontradas, apenas 26 (0,03%) foram utilizadas pelo menos 100 vezes na comunidade.

Palavras-chave: Taxonomia; Folksonomia; Hierarquia; Coerência semântica; Algoritmo; Árvore de tags;

ABSTRACT

Knowledge communities are composed of individuals who share the same interests and voluntarily work together to expand knowledge and understanding of a domain through learning and sharing. When a user asks a question in a knowledge community, he or she can usually add tags to rank the question. These tags make it easy for experts to find that question and then answer it. One of the problems with these tags is that as users themselves create them, then there is no hierarchy to aid in the search for broader themes. This study proposes a way of hierarchizing the folk-sonomia of an online knowledge community. In this study, the Stack Exchange's biology community was used, since searching for questions about a subject is a problem experienced by its users. This proposed hierarchy can help users in the search for questions, either by specific themes or by broader issues. It was possible to map 462 (65%) of the 703 tags of the community. From the *tags* that were not found, only 26 (0,03%) were used at least 100 times in the community.

Keywords: Taxonomy; Folksonomy; Hierarchy; Semantic coherence; Algorithm; Tag tree;

Sumário

Lista de Figuras

Lista de Tabelas

1. INTRODUÇÃO

Neste capítulo são apresentados os fundamentos que motivaram a realização deste trabalho, além de comentar a respeito da itinerância do pesquisador, com o objetivo de expor o problema abordado e mostrar as questões de pesquisas desenvolvidas.

1.1 O Tema e sua Importância

Atualmente, somos mais de 4 bilhões de pessoas ao redor do mundo usando a internet[1]. O Google é o portal mundial de Web, assumindo a posição de website com maior número de visitações, chegando à marca de 16,38% de visitas. Em seguida está o Facebook, o segundo mais visitado com praticamente 1/3 do número de visitas do Google: 5,89%[2]. Para buscar informações, os usuários utilizam páginas de busca como o Google e o Bing, o que explica o volume de acessos a essas ferramentas: Google com uma estimativa de 1.800.000.000 de visitantes únicos mensalmente e Bing com uma estimativa de 500.000.000[3]. Alternativamente, fontes especializadas têm sido utilizadas para a busca de informações, que no contexto Web seriam providas por plataformas de buscas sociais, como as próprias redes sociais (por exemplo, o próprio Facebook), fóruns e sites de Perguntas e Respostas (P&R) como Stack Exchange[4], Quora[5] e Yahoo Answers[6].

[1] https://wearesocial.com/blog/2018/01/global-digital-report-2018
[2] https://www.vpnmentor.com/blog/vital-internet-trends/
[3] http://www.ebizmba.com/articles/search-engines
[4] https://stackexchange.com/
[5] https://www.quora.com/
[6] https://answers.yahoo.com/

Essas comunidades são compostas por indivíduos que compartilham os mesmos interesses e voluntariamente trabalham juntos para expandir conhecimento e entendimento de um domínio através de aprendizagem e compartilhamento [29]. Essas fontes especializadas estão de acordo com o que Wang et al.[55] comentam: uma solução de gerenciamento de conhecimento não deve tornar acessível apenas o conteúdo escrito, mas também a interação com profissionais e estudiosos de determinada área que podem executar um papel social, pois as pessoas geralmente preferem o contato com um especialista ao invés de apenas obter a informação textual.

Assim, nessas comunidades de conhecimento, apresentar ferramentas para busca de especialistas é importante para evidenciar usuários que possam contribuir com um conteúdo solicitado. Um especialista pode tanto responder um problema técnico diretamente quanto indicar alguma fonte de conhecimento, na qual seja possível encontrar uma solução [55]. Diferente das organizações tradicionais, nas quais aqueles que tem um conhecimento específico em um tópico são considerados especialistas [37], a definição de especialista em comunidades online de conhecimento é mais abrangente dado que cada usuário pode ter um nível de conhecimento em uma determinada área [1].

Outras funcionalidades interessantes nessas comunidades são a possibilidade de classificar o conteúdo (perguntas e respostas) e, em algumas delas, a utilização de marcadores (*tags*) para classificar as mensagens. Entretanto, esses marcadores geralmente estão em formato de textos e não levam em consideração uma padronização de terminologia, bem como uma estrutura taxonômica [40]. Ainda assim, os marcadores apoiam os usuários na busca pela informação que necessitam, pois fornecem um mecanismo de indexação do conteúdo [38]. Do mesmo modo, também auxiliam os especialistas a encontrarem as perguntas com assuntos que dominam.

Nessas comunidades de P&R, geralmente os marcadores são criados pelos próprios usuários, formando uma folksonomia [40]. Segundo Laniado et al. [25], folksonomias são democráticas, escaláveis, inclusivas e tem um baixo custo. Porém, a inexistência de autoridade e um ponto de vista coerente e único no domínio traz limitações: a falta de hierarquia, de controle de sinônimos, de precisão, a possibilidade de *gaming* [20] [24]. O

gaming acontece quando os usuários começam a criar *tags* diferentes (e, em alguns casos, extravagantes) para cada conteúdo que adicionam na comunidade com o objetivo de promover suas publicações. Esse tipo de ação corrompe a comunidade, uma vez que essa prática é um tipo de *Spam*[7]. Para reduzir essa atividade, uma estratégia possível seria limitar quem pode criar as *tags*. No caso das comunidades do Stack Exchange, por exemplo, apenas usuários com pontuação maior que 300 podem criar novas *tags* e apenas usuários com pontuação maior que 2500 podem criar *tags* que são sinônimos. Isso pode reduzir a criação das *tags* erradas e a probabilidade de *gaming*. Existe, ainda, uma funcionalidade na comunidade, conforme Figura 1.1, em que o usuário aponta uma *tag* "Master" e uma "Synonym" e a própria comunidade vota se de fato se trata de sinônimos. Uma *tag* é aceita como sinônimo de outra caso alcance 4 votos positivos. Por outro lado, caso ela alcance pontuação igual a -2 a sugestão é excluída[6].

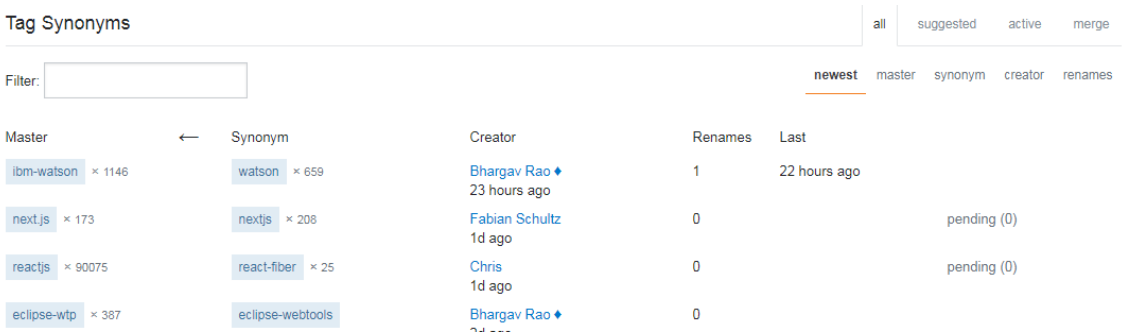

Figura 1.1: Funcionalidade de atribuição de sinônimos da comunidade

Nas comunidades do Stack Exchange, quando uma questão é criada utilizando a *tag* sinônimo, o próprio sistema apresenta a *tag* correta. De acordo com a Figura 1.1, questões criadas utilizando "watson" serão apontadas automaticamente para "ibm-watson".

As comunidades tem uma página para cada *tag* exibindo um resumo, uma definição mais detalhada, conforme Figura 1.2, com os usuários que mais responderam àquela *tag*, as perguntas que estão em alta no momento, algumas *tags* relacionadas, o número de vezes que a *tag* foi utilizada, as *tags* sinônimos, além da possibilidade do usuário visualizar um

[7]Na sua forma mais popular, spam é sinônimo de lixo eletrônico e designa mensagens de correio eletrônico com fins publicitários. O termo, no entanto, pode ser aplicado a mensagens enviadas por outros meios e noutras situações até modestas. Geralmente os spams têm caráter apelativo e na maioria das vezes são incômodos e inconvenientes.

histórico de atualização das definições da tag, sendo também possível visualizar quando a *tag* foi criada e quando foi a última vez que ela foi alterada.

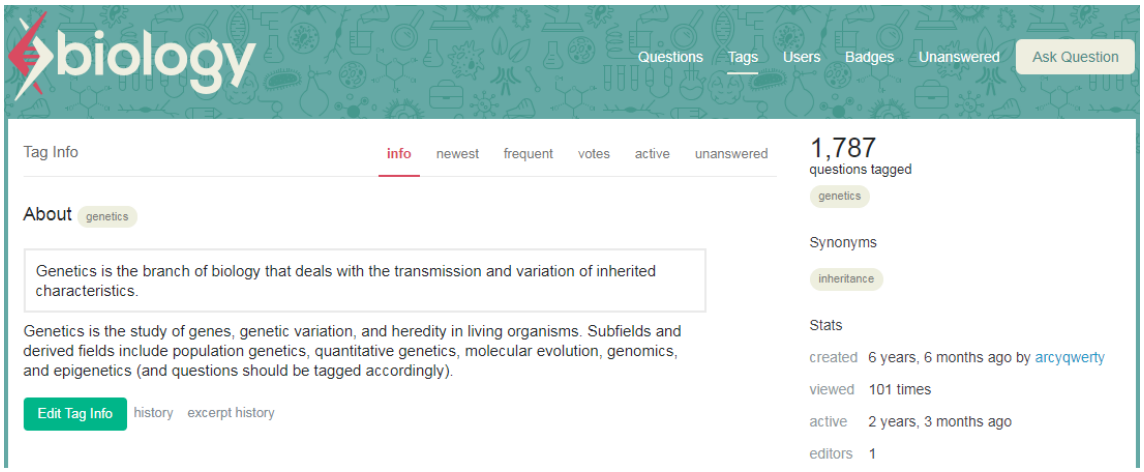

Figura 1.2: Exemplo *tag* comunidade

Conforme a Figura 1.3, na página da *tag* também é possível obter outras informações como: quais são os usuários que mais responderam as perguntas referentes àquela tag, perguntas "quentes" e *tags* relacionadas.

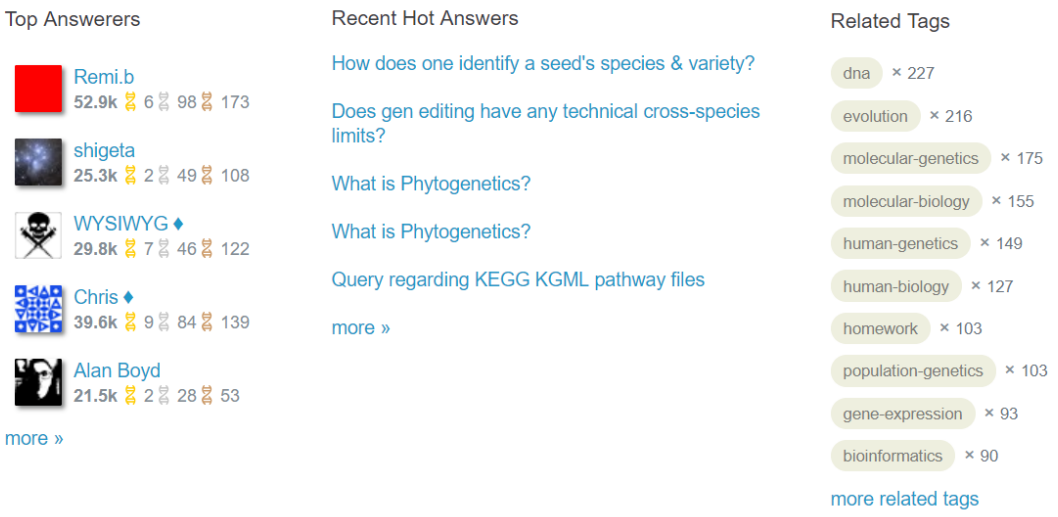

Figura 1.3: Mais informações sobre uma *tag* da comunidade

Enquanto os esquemas de classificação tradicionais, baseados em taxonomias, favorecem a busca e a navegação, folksonomias encorajam outro paradigma de navegação, baseado na procura e *serendipity* [18]. A organização dessas representações baseadas em

folksonomias não é tão simples para apresentação ao usuário, justamente devido à falta de hierarquias. Para amenizar esse problema, as aplicações geralmente utilizam métricas para mostrar algum relacionamento entre as *tags*, um exemplo é a coocorrência. Nos sites do Stack Exchange, geralmente, quando uma *tag* é acessada são exibidas as *tags* relacionadas. Porém, essas soluções não são suficientes, já que se limitam a criar uma lista de *tags* relacionadas, mas sem a definição de um critério para organizá-las como uma hierarquia.

Neste estudo, é proposto um método para hierarquizar a folksonomia de uma comunidade, considerando que dessa forma será possível, por exemplo, verificar quais são os tópicos superiores e inferiores (ou gerais e específicos) de uma *tag*. Essa hierarquização permite também que os usuários façam buscas por um tema mais abrangente, por exemplo, que um usuário da comunidade de biologia, buscando por perguntas sobre hepatite e tuberculose possa simplesmente buscar por doenças infecciosas [54].

O foco desta dissertação encontra-se na estruturação do conteúdo de comunidades online, em particular comunidades de perguntas e resposta (P&R), com a finalidade de gerar um novo conhecimento e novas formas de interação. Espera-se que este conhecimento sirva para que seja possível encontrar perguntas relacionadas a um tópico mais abrangente e que o usuário encontre mais facilmente uma resposta ou um usuário que deseje contribuir com seus conhecimentos encontre mais tópicos que possa responder ou avaliar.

1.2 Itinerância do pesquisador

Ao iniciar o mestrado, ainda não compreendia tudo o que significava pesquisar e também não tinha um plano de pesquisa programado. Na graduação, tive certo contato com professores envolvidos em programas de iniciação cientifica, porém não dispunha de tempo disponível para participar do projeto.

Meu primeiro contato com a pesquisa científica surgiu ao participar do grupo *Semantics and Learning* (SaL) e ali me interessei por comunidades de perguntas e respostas, principalmente por participar da comunidade StackOverflow desde 2010. Sempre apreciei participar de comunidades na internet, nunca muito ativo, mas sempre com algum tipo

de contribuição.

Após ler alguns trabalhos e assistir algumas apresentações de alunos do grupo, comecei a me interessar pelas *tags* das perguntas e pensar em como elas são ferramentas poderosas para que outros usuários possam responder e encontrar perguntas, me interessei principalmente quando um colega de grupo precisou produzir uma árvore com as *tags* das comunidades da StackExchange.

Com isso em mente, comecei a ler sobre o assunto e encontrei inclusive perguntas no site meta do StackExchange com usuários solicitando algo do tipo [8].

1.3 Problema de pesquisa

O uso de *tags*, palavras-chave, categorias etc., para ajudar na pesquisa e navegação entre objetos é encontrado, por exemplo, em publicações científicas, sistemas de classificação de bibliotecas e classificação biológica [12]. As *tags* podem corresponder a estruturas bem definidas, como hierarquias, com um conjunto de categorias mais estreitas ou mais abrangentes que fazem uma estrutura parecida com uma árvore composta por relações hierárquicas. Por outro lado, nos sistemas on-line é comum encontrar sistemas de marcação sem nenhuma estrutura, em que qualquer palavra relevante pode ser utilizada para marcar o item e, portanto, sem uma hierarquia predefinida de categorias e subcategorias. Essa é a realidade de diversas comunidades de perguntas e respostas, como o Stack Exchange [10].

O conjunto emergente de *tags* livres e objetos associados são geralmente referidos como folksonomias, por enfatizar sua natureza colaborativa. Um dos desafios relacionados a folksonomias é extrair uma hierarquia entre as *tags* [45]. Embora a maioria dos sistemas de marcação seja igualitária, ou seja os usuários podem criar *tags* livremente, a forma como os usuários pensam sobre objetos tem alguma hierarquia incorporada, por exemplo, "câncer" é geralmente considerado como um caso especial de "malignant-tumor". Ao

[8] https://meta.stackexchange.com/questions/45438/a-proposal-for-tag-hierarchy-on-stack-exchange-sites

revelar esse tipo de hierarquia a partir de estatísticas de coocorrência de *tags*, podemos ajudar a ampliar ou restringir o escopo da pesquisa no sistema [57][43], dar recomendações sobre perguntas que ainda não foram respondidas ou auxiliar na categorização de perguntas que estão sendo criadas [45].

1.4 Objetivo da pesquisa

O objetivo principal desta pesquisa é definir um conjunto de procedimentos que permitam uma estruturação hierárquica automática e coerente a partir de uma folksonomia de uma comunidade de conhecimento.

1.5 Método

Para avaliar a proposta, experimentos foram realizados com base em dados coletados da comunidade de Biologia do Stack Exchange.

Para avaliar se o processo teve um bom desempenho, foram implementados algoritmos com diferentes abordagens de seleção da palavra correta. Dessa forma, espera-se que caso o algoritmo deste trabalho escolha uma palavra para ser adicionada na hierarquia, ela esteja correta.

1.6 Organização da Dissertação

Esta dissertação está organizada em cinco capítulos, sendo este o primeiro. O capítulo dois é destinado à apresentação de conceitos relacionados às comunidades de conhecimento, folksonomias e WordNet, apresentando uma visão ampla sobre o tema abordado e os trabalhos relacionados a esta pesquisa, mostrando suas similaridades e diferenças.

No capítulo três é desenvolvido o processo de hierarquização da folksonomia e um estudo de caso utilizando a comunidade de Biologia do Stack Exchange. Também são descritos como os dados da comunidade são carregados no banco de dados e quais os

principais problemas da execução nessa base específica.

O capítulo quatro é destinado à apresentação da estratégia de avaliação do trabalho que foi realizada para averiguar se o processo de estruturação traz bons resultados em relação as heurísticas que foram executadas.

Por fim, o capítulo cinco é destinado às conclusões e considerações finais.

2. CONCEITOS FUNDAMENTAIS

2.1 Comunidade de práticas

Com o aumento da necessidade por conhecimento, os usuários o buscam também em fontes externas como fóruns, sites de Perguntas e Respostas (P&R) [11][55][62], ferramentas que possibilitam a formação de comunidades, que são compostas por indivíduos que compartilham os mesmos interesses e de forma voluntária trabalham juntos para expandir conhecimento e entendimento de um domínio através de aprendizagem e compartilhamento [29].

Comunidades de conhecimento são um tipo de comunidade de práticas, especializadas no compartilhamento e busca de conhecimento [8][55][56]. Wang et al. [55] comentam que uma solução de gerenciamento de conhecimento não deve tornar acessível apenas conteúdo escrito, mas também especialistas que podem executar um papel social, pois as pessoas geralmente preferem o contato com um especialista ao invés de apenas o texto.

Comunidades de perguntas e respostas (P&R) são sites em que os usuários podem postar perguntas, responder perguntas de outros usuários e avaliar perguntas e questões de outros usuários. A partir dessas interações, os usuários são ranqueados de acordo com suas contribuições [42], podendo fazer com que a participação deles aumente [41].

As comunidades de conhecimento, geralmente utilizam *tags* para categorizar as perguntas que são postadas. Como essas *tags* são criadas pelo próprio usuário, elas formam uma folksonomia. Conforme a Figura 2.1, nas comunidades do Stack Exchange, quando

um usuário cria uma pergunta, ele deve dar um título, descrever com detalhamento a pergunta, pode adicionar imagens ou links para recursos externos, dentre outras opções e por último, o usuário deve adicionar as *tags*, no mínimo uma e no máximo 5 [22][9].

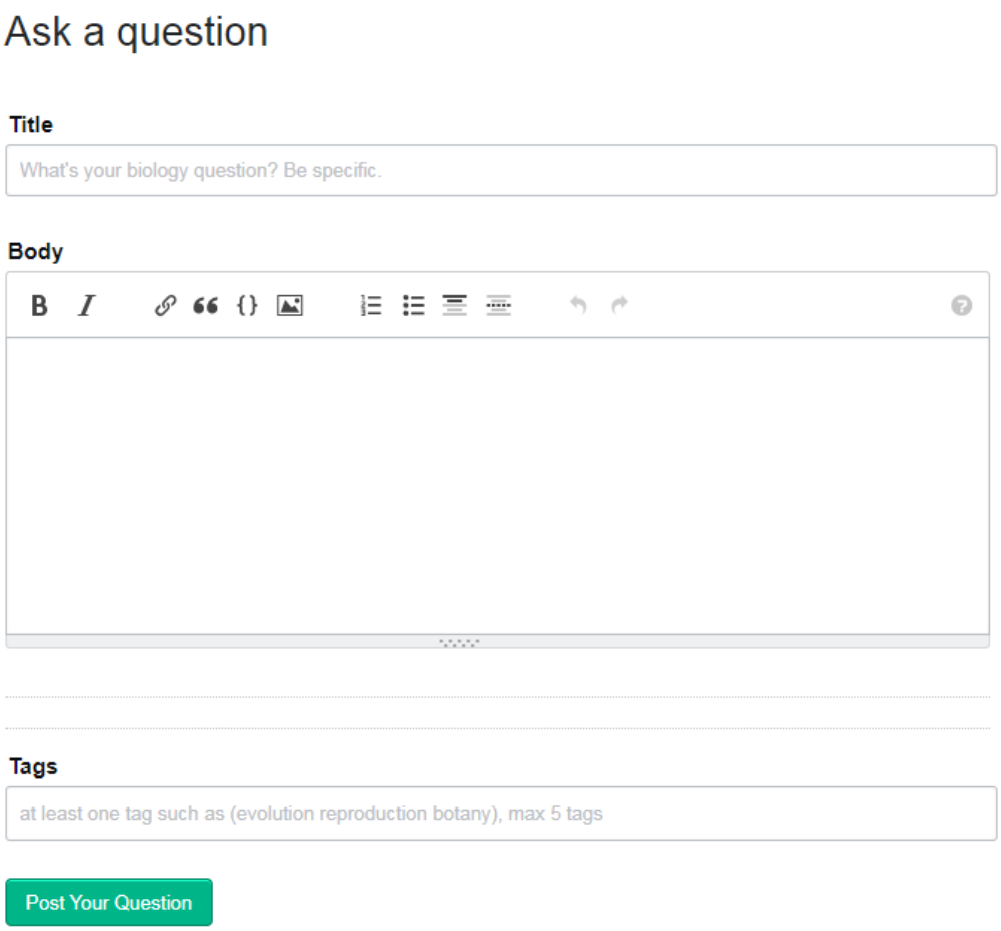

Figura 2.1: Formulário para criação de pergunta

Os usuários são encorajados a responder as próprias perguntas[1], por isso, ao cadastrar uma pergunta, eles podem respondê-la no ato da criação. Conforme a Figura 2.2, quando o usuário selecionar a opção "Answer your own question" será exibida outra caixa de texto para que ele escreva a resposta para a pergunta que será publicada. Outra opção é o usuário marcar a resposta como "Community Wiki", e todos os usuários com score maior que 100 podem efetuar alterações e nenhum deles ganha pontuação quando a resposta receber "upvotes" ou "downvotes"[2]. Essas respostas, segundo o próprio StackExchange, servem para compartilhar a carga de uma resposta.

[1]https://stackoverflow.blog/2011/07/01/its-ok-to-ask-and-answer-your-own-questions/
[2]https://stackoverflow.blog/2011/08/19/the-future-of-community-wiki/

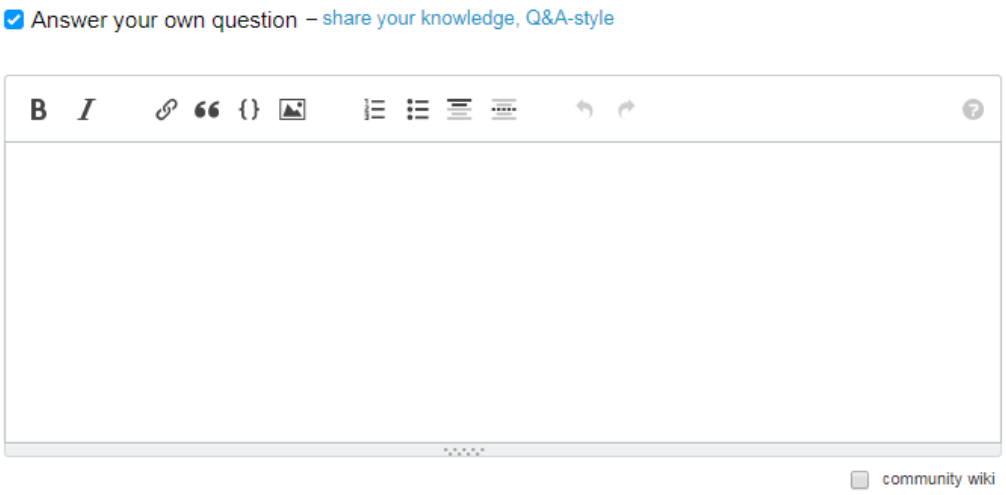

Figura 2.2: Formulário para que o usuário responda a própria pergunta

Quando o usuário decide adicionar *tags* à pergunta, ele pode visualizar um resumo das *tags* disponíveis com o que digita na caixa de texto, conforme a Figura 2.3. Ao digitar o texto "evo" na caixa de texto, por exemplo, das *tags* existentes são exibidas as que contém esse termo. As informações da *tag* podem ser acessadas ao clicar no texto da *tag*. Além disso é exibido quantas vezes a *tag* foi utilizada na comunidade.

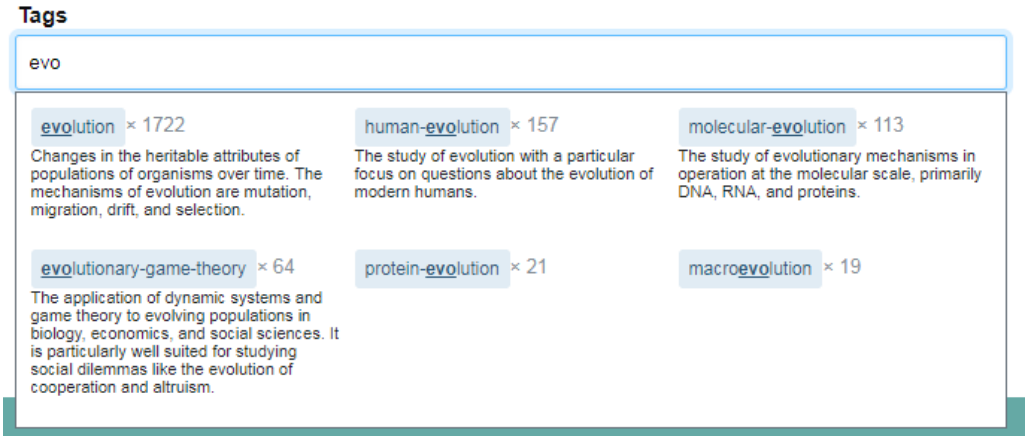

Figura 2.3: Visualização das tags quando usuário vai criar a pergunta

2.2 Taxonomias e Folksonomias

Taxonomias tem um papel considerável em várias áreas [30]. As taxonomias são usadas para verificar a relação entre documentos e ligar bugs e alterações efetuadas [64][59]. No entendimento do programa, taxonomias fornecem uma maneira eficaz de obter a se-

melhanças entre palavras chave dos comentários e identificadores no software [47].

Entretanto segundo [64], a maioria das taxonomias existentes usadas em aplicações são criados manualmente de acordo com a aplicação específica requisitos e seu tamanho não é grande o suficiente. Taxonomias resultantes de abordagens automáticas provavelmente levariam a resultados não tão bons quando aplicados em aplicações mais específicas[64]. Um dos problemas de uma taxonomia é a temporalidade, temas surgem na internet muito rapidamente. Outro problema é a granularidade, em abordagens tradicionais, a construção de uma taxonomia pode ser de domínio aberto, então, alguns termos mais específicos podem não ser encontrados nessas taxonomias [65].

O termo taxonomia é associado com a linguagem documentária, usada para classificar termos em uma estrutura hierárquica. A taxonomia pode ser utilizada para classificar vários tipos de objetos como: seres vivos, coleções de livros, documentos, com o objetivo de possibilitar a identificação, localização e acesso [52].

Os esquemas de classificação que eram usados antes da Web foram construídos a partir de uma visão de informação linear e sem a participação do usuário. Uma vez que a Web é um ambiente ágil, onde as informações mudam constantemente, foi necessário construir caminhos nos quais o usuário esteja no controle do vocabulário do usuário, uma vez que essas informações são importantes para a organização e posterior recuperação das informações que estão transitando pela web.

Um dos motivos que fez a web 2.0 mudar a internet na época foi passar de um mundo em que eram usadas taxonomias burocráticas e rígidas para estruturas mais flexíveis como o uso das *tags*, criadas pelo próprio usuário. Sendo assim, visto que os usuários estavam participando mais da criação e utilizando informações dos serviços, portais, redes sociais etc., se fez necessária uma forma mais flexível para recuperar, organizar e recuperar a informação [50].

A expansão da produção de conhecimento na Web trouxe grandes impactos no papel dos usuários, permitindo que eles participem ativamente da construção, organização e recuperação das informações. Nesse contexto, uma forma de participação na organização

digital de recursos na web é a atribuição de *tags* aos recursos.

A prática de rotular publicamente ou categorizar recursos em um ambiente compartilhado online é denominada "*Social Tagging*" e essa categorização é chamada de folksonomia. Criado por Thomas Vander Wal, em 2004, como resultado da junção dos términos "folk" e "taxonomia", o termo folksonomia "é o produto da livre e pessoal atribuição de *tags*, por um usuário, para um recurso identificado por uma *Uniform Resource Identifier* (URI), para facilitar a sua recuperação. Folksonomia é o resultado de: um sujeito, um conteúdo (recurso com uma URI) e um rótulo" [18].

Quando compartilhadas com outras pessoas ou visualizadas no contexto do que outras pessoas marcaram, essas coleções de identificadores de recursos, *tags* e pessoas começam a ganhar valor por meio de efeitos de rede. A pesquisa de *tags* pode permitir a descoberta de recursos relevantes, e as relações sociais que se desenvolvem entre os usuários tornam-se um meio de descoberta de informações [51].

Entretanto, em comparação com taxonomias, folksonomias tem uma falha inerente, não existem relações hierárquicas explícitas entre as *tags*[46]. É comentado por Kome et al. [23] que existem relações hierárquicas implícitas em folksonomias. Em [5] e [7] mostram que a organização de *tags* em estruturas hierárquicas ajudará os aplicativos de recuperação de informações baseados em *tags*.

Segundo [63], se for feita uma comparação com a organização tradicional de metadados, uma foksonomia apresenta uma melhoria em relação à cooperação. Em taxonomias tradicionais, que são pré-definidas apenas por grupos de especialistas, são limitadas e tornarm-se desatualizadas com frequência. As folksonomias por sua vez resolvem esses problemas transferindo a carga de alguns indivíduos para todos os usuários da web.

A natureza comunitária das folksonomias reduz significativamente o custo da indexação e a torna uma opção viável para a geração de metadados [22]. Entretanto, se for feita uma comparação com sistemas de classificação de bibliotecas, por exemplo, que são construídas e mantidos por especialistas, as folksonomias sofrem de problemas de inconsistência e oferecem uma qualidade de indexação menor[22]. Por não existir um controle

as folksonomias tornam-se propensas a inconsistências causadas por variações ortográficas, sinônimos, acrônimos e hipônimos[22]. Essas inconsistências, por outro lado, podem levar a problemas como "explosão de *tags*", em que um subconjunto de *tags* de uma folksonomia é utilizado para classificar a maioria dos itens em um conjunto e as outras *tags* são utilizadas minimamente [19].

Segundo Kroski[24], as folksonomias tem uma série de benefícios:

São inclusivas. Enquanto taxonomias utilizam um vocabulário controlado que é excludente por natureza, as folksonomias incluem o vocabulário de todos e refletem as necessidades de todos, sem preconceitos culturais, sociais ou políticos. Como as folksonomias incluem visões alternativas junto com as populares, elas apresentam uma oportunidade única para descobrir interesses de "cauda longa"[3]. Quando combinados, esses interesses não tradicionais ou de nicho superam em muito os mais populares.

São atuais. Os sistemas baseados em *tags* oferecem uma fluidez que não é possível em uma taxonomia hierárquica controlada. Os usuários criam *tags* com a mesma rapidez com que criam conteúdo. Essa flexibilidade permite respostas rápidas a mudanças na terminologia e a eventos mundiais. Uma grande taxonomia, pode levar anos para adicionar uma data de morte ao registro de autoridade do autor. Na criação de esquemas de classificação tradicionais, o catalogador deve prever antecipadamente as categorias permanentes. O problema com este modelo é que as situações mudam, por exemplo, os países mudam de nome, a tecnologia se expande e às vezes grupos de pessoas mudam a maneira como se referem a si mesmos.

Facilitam a descoberta. As taxonomias são projetadas para encontrar recursos específicos, enquanto as folksonomias estão predispostas a descobrir recursos desconhecidos e inesperados. Esses sistemas promovem a exploração e o aprendizado à medida que os usuários navegam em tópicos relacionados, *tags* e usuários. Existe um valor legítimo em um sistema de descoberta, pois os usuários têm a oportunidade de

[3] A cauda longa [2] consiste nos interesses da minoria que se encontram no final de uma distribuição estatística, que mapeia os tópicos mais populares.

localizar novos recursos que talvez nunca tenham encontrado através da pesquisa.

Os itens são binários. Em um esquema de classificação tradicional, um vocabulário controlado deve ser feito antecipadamente, no qual um termo de categoria é selecionado, o que inclui todos os termos relacionados. Quando objetos futuros são catalogados, deve ser determinado que eles se encaixam em uma categoria específica ou não. Em uma folksonomia, esses itens podem se encaixar em várias categorias. Por exemplo, uma imagem pode conter *tags* como cachorro, dog, rex, canino. Em uma folksonomia, o esquema é multifacetado.

São democráticas e auto gerenciadas. Todos têm a oportunidade de adicionar algo ao todo. Da mesma forma, esses sistemas são auto-moderados. Por sua natureza, esses sistemas incentivam os usuários, do ponto de vista individual, a escolher *tags* que descrevam itens apropriadamente, o que, por sua vez, os ajuda a lembrá-los no futuro. Da mesma forma, como a marcação é feita em um fórum público, a dinâmica social induz os usuários a escolher *tags* relevantes. Ao marcar um novo item, muitos sistemas oferecem aos usuários uma lista das *tags* mais usadas. A ideia é que os termos mais populares tendem a ser os mais relevantes, assim como um artigo ou livro frequentemente citado é considerado como detentor de mais autoridade no meio acadêmico.

Porém Kroski[24] e Bagheri et al. [3] também citam alguns pontos negativos:

Falta controle de sinônimos. Não existe um vocabulário controlado e, portanto, não existe um termo para descrever um conceito ou entidade. Isso é considerado uma deficiência quando usuários diferentes usam termos diferentes para descrever a mesma coisa, como computador e pc. Não há como regular o uso de plurais. Além disso, muitas comunidades fornecem listas de "termos relacionados" para incentivar o uso de sinônimos "populares".

Não tem precisão semântica. Folksonomias são sistemas de descoberta, sem a poderosa capacidade de pesquisa de uma taxonomia hierárquica. Caracteristicamente, eles terão baixas taxas de precisão.

Não possui hierarquia. Não há relacionamentos pai-filho, categorias e subcategorias. A hierarquia é um traço distintivo das taxonomias tradicionais que são capazes de fornecer uma classificação de entidades mais profunda e mais robusta. Tais sistemas permitem aos usuários uma granularidade mais fina na busca de recursos.

Não retornam todos os resultados de um assunto. Devido à falta de controle de sinônimos, uma pesquisa de folksonomia não vai trazer uma lista completa devido ao uso de *tags* semelhantes. Uma pesquisa por "genetics", por exemplo, não trará recursos que foram marcados com "human-genetics", "molecular-genetics" ou "population-genetics".

Neste trabalho, buscamos combinar a flexibilidade das folksonomias com a hierarquização das taxonomias, promovendo assim as características positivas de ambos.

Segundo [46] ao criar um agrupamento hierárquico de uma folksonomia podem ser reveladas relações de similaridade entre *tags*. A Figura 2.4 (a) mostra um exemplo de *tag cloud* que geralmente é usada, na qual apenas a popularidade de uma *tag* está relacionada. A A Figura 2.4 (b) mostra um exemplo de um agrupamento em *clusters* de *tags*. A pesar de que os *clusters* capturam semelhanças entre as *tags*, problemas ainda permanecem: *clusters* misturam diferentes relações, como sinônimos e hyperônimos; é difícil avaliar a exatidão do agrupamento, ou seja, é difícil dizer se duas *tags* são semelhantes ou não. Segundo [46] as relações direcionadas e fáceis de avaliar entre *tags* são preferidas, como na Figura 2.4 (c).

Deferente de outros serviços como del.icio.us[4] e Flickr[5] onde os usuários criam tags livremente em suas postagens, as folksonomia das comunidades do Stack Exchange são folksonomias onde os usuários adicionam *tags* que já existem, caso o usuário queira criar uma pergunta e uma das *tags* que ele deseja adicionar na pergunta ainda não existam na comunidade, ele deve ter uma pontuação na comunidade referente a criação de *tags* que na comunidade de Biologia é de 300[6], se o usuário já atingiu essa pontuação basta ele

[4] del.icio.us
[5] https://www.flickr.com/
[6] Cada comunidade tem uma pontuação diferente para criação de tags.

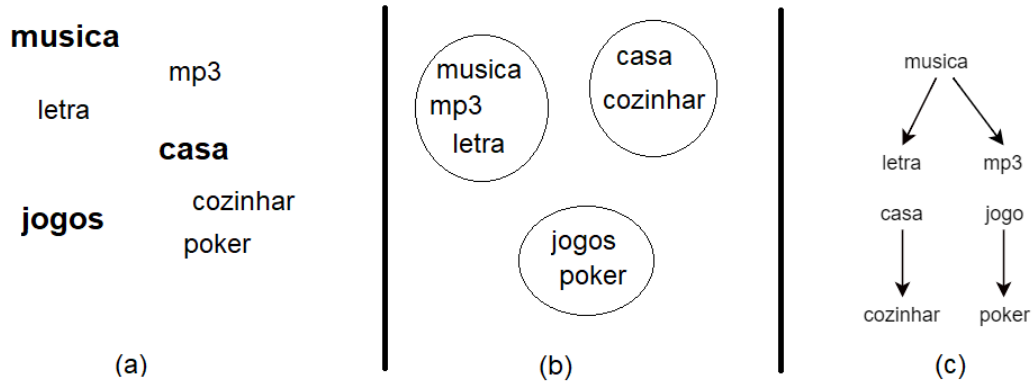

Figura 2.4: Exemplos de (a) nuvem de tags *flat*, (b) *clusters* e (c) relações.

acrescentar a *tag* na pergunta para que ela seja criada na comunidade[7]. Caso o usuário não tenha essa pontuação ele precisará recorrer a moderação da comunidade para que a *tag* seja criada. Em alguns sites do Stack Exchange, *tags* que não são usadas em pelo menos uma pergunta nos últimos 6 meses são removidas automáticamente da comunidade. Os sites recomendam também que os usuários só deve criar novas *tags* quando for constatado pelos usuários que a pergunta criada cobre um novo tópico sobre o qual nenhum outro usuários tenha feito alguma pergunta antes naquele site.

2.3 WordNet

Geralmente as *tags* em uma folksonomia não tem semântica, então, para fornecê-la às *tags* das folksonomias estudadas, foi utilizado o WordNet devido a facilidade de uso e a vasta documentação online.

A WordNet de Princeton [36] [13] constitui a primeira base de dados de conhecimento linguístico em que o significado lexical é representado através de uma rede de relações lexicais e conceituais, sendo o significado de cada unidade lexical derivado da sua posição na rede.

O conceito é a unidade básica nas wordnets. A estruturação das unidades sinônimas é distribuída em quatro categorias lexicais (substantivos, verbos, adjetivos e advérbios) que

[7]https://biology.stackexchange.com/help/privileges/create-tags

formam “conjuntos de sinônimos” (synsets) [34][4]. Cada nó da rede é constituído por um synset, que contém todas as lexicalizações [8] de um conceito. Por exemplo “center” e “middle” estão incluídas no mesmo synset, uma vez que ambas são lexicalizações do conceito “centre” [35] conforme Figura 2.5.

- S: (n) **center**, centre, middle, heart, eye (an area that is approximately central within some larger region) *"it is in the center of town"; "they ran forward into the heart of the struggle"; "they were in the eye of the storm"*

Figura 2.5: Exemplo de synset

Ao navegar por um synset, estão disponíveis diversas opções, dentre elas: *direct hyponym*, *has instance*, *direct hypernym*, *derivationally related form*, *domain region* dentre outros, conforme Figura 2.6.

S: (n) center, **centre**, middle, heart, eye (an area that is approximately central within some larger region) *"it is in the center of town"; "they ran forward into the heart of the struggle"; "they were in the eye of the storm"*

- ***direct hyponym*** / *full hyponym*
  - S: (n) center stage, centre stage (the central area on a theater stage)
  - S: (n) city center, city centre, central city (the central part of a city)
  - S: (n) storm center, storm centre (the central area or place of lowest barometric pressure within a storm)
  - ■ ■ ■
  - S: (n) medical center (the part of a city where medical facilities are centered)
  - S: (n) midfield ((sports) the middle part of a playing field (as in football or lacrosse))
  - S: (n) seat (a center of authority (as a city from which authority is exercised))
  - S: (n) midstream (the middle of a stream)
  - S: (n) City of London, the City (the part of London situated within the ancient boundaries; the commercial and financial center of London)
- *has instance*
- *direct hypernym* / *inherited hypernym* / *sister term*
- *derivationally related form*
- *domain region*

Figura 2.6: Expandindo opções de um synset

As wordnets e os dicionários comuns assemelham-se, pois apresentam glosas, bem como tesauros, uma vez que são organizadas a partir de sinônimos [13]. A diferença é o fato de organizarem suas bases lexicais baseando-se em relações semânticas ou conceituais e não em uma ordem alfabética. As wordnets adotaram essa organização com a intenção de apresentar o léxico em uma organização inspirada no léxico mental, conforme

[8] A lexicalização trata da utilização de um determinado termo pelo léxico de uma língua, como uma formação usual.

teorias psicolinguísticas [36]. Essas bases de dados lexicais podem ser aproveitadas para a construção de sistemas dedicados ao Processamento de Linguagem Natural (PLN) [13].

O WordNet é organizado por relações semânticas, tais como: sinônimo, antônimos, hipônimo, merônimo e relações morfologicas [36]. Uma vez que uma relação semântica é uma relação entre significados e que significados podem ser representados por synsets, é natural pensar em relações semânticas como relações entre synsets [35]. É característico das relações semânticas que elas sejam recíprocas: se há uma relação semântica R entre o significado x e o significado y, então há também uma relação R' entre y e x [36].

A relação mais importante para o WordNet é a similaridade, já que a capacidade de julgar essa relação entre formas de palavras é um pré-requisito para a representação de significados em uma matriz lexical [35]. De acordo com uma definição, duas expressões são sinônimas se a substituição de uma pela outra nunca muda o valor de verdade de uma sentença na qual a substituição é feita [34].

Outra relação é o caso do antônimo, por exemplo, ricos e pobres são antônimos, mas dizer que um indivíduo não é rico não implica que ele deva ser pobre; muitas pessoas não se consideram nem ricas nem pobres [36]. Antônimo, que parece ser uma simples relação simétrica, é na verdade bastante complexo, ainda que as pessoas tenham pouca dificuldade em reconhecer antônimos quando os veem [36].

Ao contrário de sinonímia e antonímia, que são relações lexicais entre formas de palavras, hiponímia e hiperonímia são relações semânticas entre significados de palavras, por exemplo: doença é um hiperônimo de diabetes e diabetes é um hipônimo de doença [35].

Quando uma busca é executada no WordNet, são retornados todos os synsets disponíveis para aquele termo. Conforme a Figura 2.7, quando efetuamos a busca pelo termo "centre", verifica-se que são exibidas diversas possibilidades para este termo.

Devido à sua estrutura rígida, o WordNet é muitas vezes referido como uma ontologia[15] [16]. O WordNet representa uma abordagem que revela as formas sistemáticas pelas quais uma linguagem mapeia conceitos em palavras. O WordNet se concentra no léxico, mas sua estrutura rígida e na representação de palavras e conceitos superiores, sendo com-

Figura 2.7: Todos os synsets para a palavra centre

parada a uma estrutura de conhecimento independente de linguagem. O mapeamento de conceitos em ontologias para synsets mantém essa distinção e possibilita padrões de mapeamento[14].

O conhecimento sintático e semântico expresso pelo WordNet tem sido utilizado como recurso linguístico por diversos campos, como: Aprendizado de Máquina, Processamento de Linguagem Natural e Recuperação de Informação e sendo utilizado em diversos assuntos, dentre eles: Reconhecimento de Entidade, Desambiguação de Sentido de Palavras, *Matching* de Ontologia e Predição de Serviços Web [26].

As relações semânticas no WordNet foram escolhidas pois se aplicam extensivamente em toda a linguagem e são familiares, significando que um usuário não precisa de preparo avançado em lingüística para entendê-las. Cada relação semântica é representada por indicadores entre formas de palavras ou entre synsets. Atualmente, a versão mais recente

do WordNet contém mais de 117.000 synsets, representando relações semânticas entre palavras e sentidos [26].

Conforme [26], o WordNet é um recurso que é amplamente utilizado e adotado em trabalhos qe tratam de linguística computacional. A quantidade de relações semânticas expostas pelo WordNet compõe uma rede semântica densa entre os synsets, e quanto mais relações, mais denso se torna.

2.4 Trabalhos relacionados

Estudos que trabalham com as *tags* de comunidades de conhecimento já têm sido explorados na comunidade científica. Uma forma de buscar por *tags* (marcadores) relacionados é através da identificação de agrupamentos (clusters) ou técnicas de clustering [5]. O algoritmo é baseado na contagem do número de coocorrência de qualquer par de *tags* e o ponto de corte é decidido quando a coocorrência é significativa o suficiente. Este trabalho não faz a hierarquização das *tags* e também não leva em consideração a semântica ao fazer os cluster, mas apenas a coocorrência das *tags*. Entretanto, esta abordagem não tem em vista a semântica das *tags* o que pode levar a uma hierarquia errada.

Outros trabalhos utilizam a coocorrência para executar a estruturação das *tags*, Begelman et al. in [5] aplicaram coocorrência para construir uma relação gráfica de *tags* e, então, executar recursivamente um algoritmo de partição para construir um cluster de *tags* para buscar *tags* relacionadas. Li et al. em [28] propuseram que padrões de coocorrências frequentes de *tags* de usuário podem ser usados para caracterizar e capturar tópicos que são de interesse dos usuários, porque essas *tags* de coocorrência são relacionadas e podem ser agrupadas para representar tópicos. Porém estas abordagens não levam em conta a semântica das *tags*, o que poderia melhorar a hierarquia gerada.

Ao trabalhar com a semântica das *tags*, Lee e Yong [27] descevem um sistema que utiliza a *WordNet* para fazer a desambiguação das *tags* do *Flickr*, indo portanto além da simples identificação de clusters. O usuário busca por uma *tag* e o sistema mostra todos os termos da *WordNet*. Esse estudo não considera as *tags* que aparecem em conjunto para

dispor de um contexto melhor e resultados errados eram exibidos.

Alguns trabalhos utilizam conceitos de ontologias em folksonomias, Mika [33], o modelo bipartido de ontologias é estendido à dimensão social, levando a um modelo de atores, conceitos e instâncias. É demonstrada a aplicação da representação mostrando como a semântica baseada na comunidade emerge do modelo por meio de um processo de transformação de grafo, esse processo pelo qual um grafo existente é alterado para produzir uma variação do grafo. Van Damme et al. [53] argumentam que a interação social manifestada nas folksonomias e em seu uso deve ser explorada para a construção e manutenção de ontologias. Em seguida, esboçam uma abordagem abrangente para derivar ontologias de folksonomias integrando vários recursos e técnicas. É sugerido combinar a análise estatística de folksonomias com dados de uso associados e suas redes sociais implícitas, recursos lexicais online como dicionários, Wordnet, Google e Wikipedia, ontologias e recursos da Web Semântica, mapeamento de ontologia e abordagens de correspondência e funcionalidade que ajudam os atores humanos a alcançar e manter o consenso sobre sugestões de elementos de ontologia resultantes.

Existem abordagens baseadas em enciclopédias, que se concentram na extração de hierarquias conceituais da Wikipedia. WikiTaxonomy [39] a partir do sistema de categorias da Wikipedia constrói uma taxonomia. Contém 105.000 relações com a precisão de 88%. O Kylin Ontology Generator (KOG) [58] usa a Markov Logic Network (MLN) para prever as relações entre as classes de infobox da Wikipedia. Yago [21] interliga as categorias da Wikipédia com os synsets do WordNet. Existem mais de 200.000 classes e 400.000 relações em Yago e a precisão é de aproximadamente 96%. Nossa pesquisa é diferente de abordagens baseadas em enciclopédias pois não existe informações de estrutura entre as tags das comunidades do Stack Exchange.

Existem trabalhos que tratam de problemas que são orientados aos usuários como em Xie et al. [61] que propõem uma abordagem para construir e enriquecer perfis de usuários e recursos com base nas relações fornecidas pela folksonomia, com base nas *tags*, recursos e usuários. Essa abordagem extrai relações semânticas entre *tags* e constrói uma hierarquia entre elas para calcular similaridades de objetos da Internet, de modo a permi-

tir que os provedores de serviços ou de conteúdo da Internet percebam o significado das *tags* como os seres humanos. No trabalho de Takehara et al. [48] é apresentado um novo esquema para recuperar os conteúdos desejados pelos usuários, com tópicos aos quais os usuários estão interessados, de várias plataformas de mídia social. Esse trabalho extrai a estrutura hierárquica de grupos de conteúdo de diferentes plataformas de mídia social e torna os conteúdos recuperáveis, mesmo se os usuários não especificarem plataformas adequadas e não inserirem consultas adequadas. A estrutura hierárquica extraída mostra vários níveis de abstração de grupos de conteúdo e seus relacionamentos hierárquicos, que podem ajudar os usuários a selecionar tópicos relacionados à consulta de entrada. No entanto, essas abordagens utilizam intervenção humana para fazer a hierarquia das *tags*.

Outros trabalhos tentam construir hierarquias a aplicando algoritmos, Chen e Luo [10] apresentam um algoritmo para construir uma hierarquia de *tags*. O trabalho de Tibély et al. [49] executa métodos de extração de hierarquia de *tags* que estão baseados na teoria de redes. Schmitz et al. [44] discutem como a mineração de regras de associação pode ser adotada para analisar e estruturar folksonomias, e como os resultados podem ser usados para aprendizagem de ontologias e suporte à semântica emergente. Conquanto, essas abordagens não consideram a semântica das *tags* para procurar os relacionamentos, o que pode levar a uma hierarquia errada.

Trabalho	**Semântica**	**Coocorrência**	**Automático**	**Mais de um método de desambiguação**
BEGELMAN (2006) [5]		X	X	
CHEN (2017) [10]		X	X	
HOFFART (2013) [21]	X	X		
LEE (2007) [27]	X		X	
LI (2008) [28]		X	X	
MIKA (2005) [33]	X	X		
PONZETTO (2008) [39]	X	X		
SCHMITZ (2006) [44]		X	X	
TAKEHARA (2017) [48]	X	X		
TIBÉLY (2013) [49]		X	X	
VAN DAMME (2007) [53]	X		X	
WU (2008) [58]	X		X	
XIE (2017) [61]	X	X		
Método do trabalho	**X**	**X**	**X**	**X**

Tabela 2.1: Tabela com comparativo dos trabalhos relacionados

Os trabalhos analisados referenciados tratam de assunto similar ao proposto nesta dissertação. O resumo da Tabela 2.1 faz um comparativo relacionando os trabalhos com as características que são importantes para a construção de árvores a partir de *tags*. Este trabalho, entretanto, propõe a hierarquização apoiada pela *WordNet*, utilizando apenas a palavra que define a *tag*, *tags* que aparecem em conjunto em outras perguntas e utilizando mais de um método de desambiguação entre os termos da *WordNet*. Além disso, avançando em relação aos trabalhos citados, a proposta aqui descrita não requer intervenção do usuário, trazendo uma abordagem automática para a hierarquização.

3. ESTRUTURAÇÃO DE UMA FOLKSONOMIA

O objetivo deste capítulo é discorrer sobre o processo de estruturação de uma folksonomia, apresentando seu processo e logo após um estudo de caso que utilizou as *tags* da comunidade de Biologia do Stack Exchange como fonte de dados.

3.1 Processo de estruturação

Conforme a Figura 3.1, para cada elemento da folksonomia são atualizados os itens que tem apenas um elemento na *WordNet*. Por exemplo, a *tag histology* tem apenas uma palavra relacionada, que significa exatamente a definição escrita na comunidade de biologia. Caso não seja encontrado nenhum synset para a palavra, ela é ignorada e o algoritmo segue para a próxima tag.

A segunda etapa executa a desambiguação para encontrar a palavra que mais se encaixa com a *tag*, por exemplo a *tag eyes* apresenta 7 que são substantivos e 1 verbo, sendo este descartado, uma vez que somente os substantivos são necessários. Para fazer a escolha da palavra correspondente foi utilizado o contexto em que a *tag* é utilizada, ou seja, as *tags* que são utilizadas junto com a *tag* em questão, tendo em vista todas as perguntas da comunidade. A *tag* "human-biology", por exemplo, é utilizada em conjunto com as *tags* "microbiology" e "neurology" em algumas perguntas na comunidade conforme Figura 3.2, sendo assim, essas *tags* são o contexto utilizado. Então é feita uma comparação de similaridade entre as palavras, utilizando a fórmula de Wu e Palmer[60], que já é implementada pela *WordNet*.

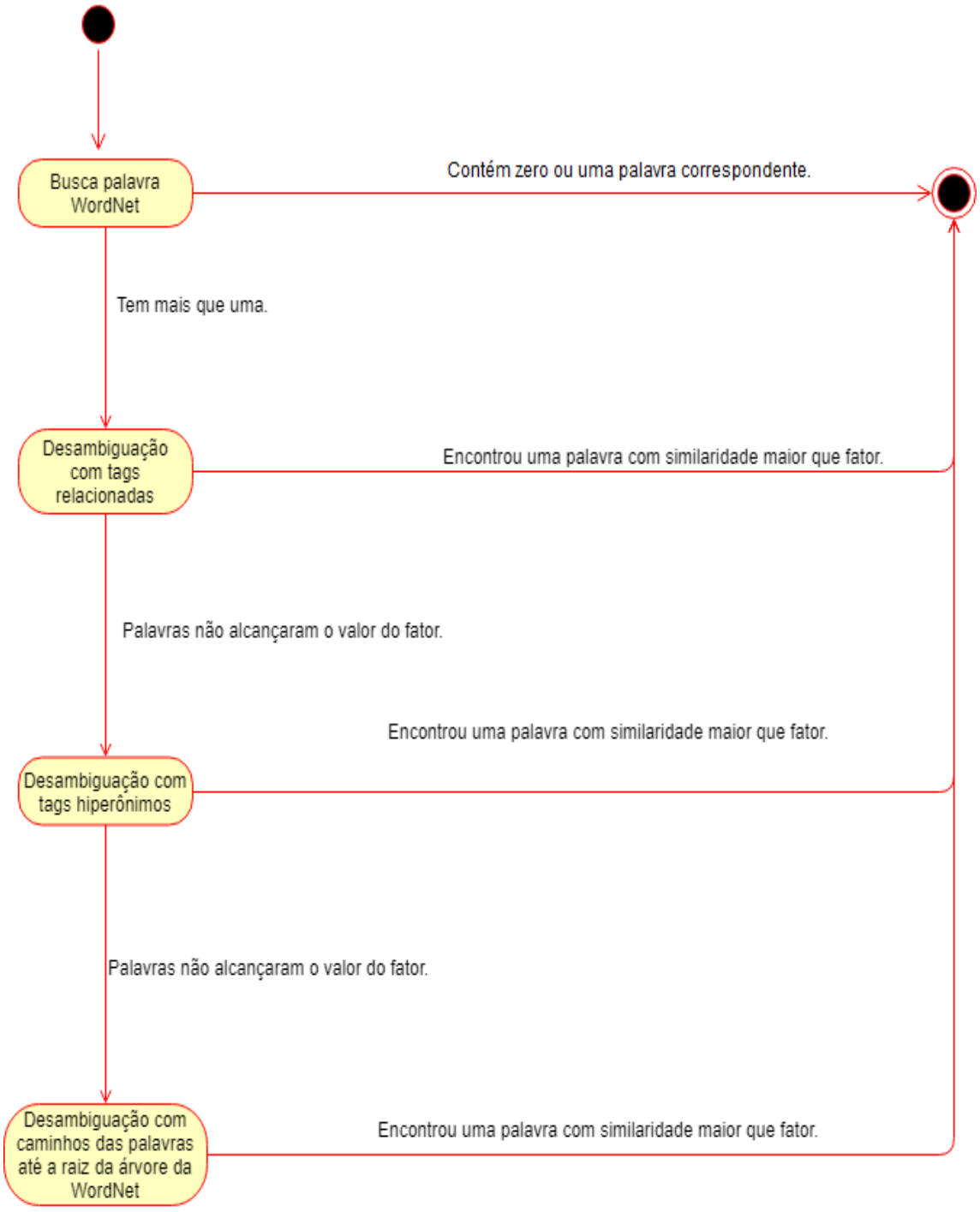

Figura 3.1: Processo

No trabalho [60], o problema abordado é o da tradução automática entre línguas. Se constata que, durante a tradução de verbos do inglês para o chinês ou o contrário, não é fácil obter uma cobertura boa dos significados listando somente duplas traduzidas. Os autores então, sugerem para os verbos uma representação semântica, ou seja, para cada verbo é definido um conjunto de conceitos de domínios conceituais diferentes. Assim, com base nessa representação, pode ser definido uma medida de similaridade. Uma métrica baseada no comprimento do caminho entre os conceitos e a raiz da hierarquia é apresentada em [60].

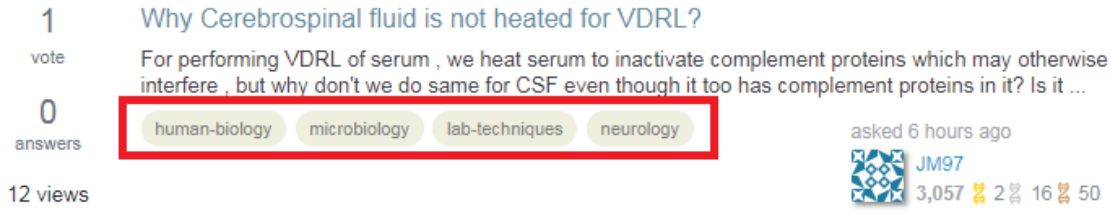

Figura 3.2: Tags que aparecem em conjunto

Para encontrar o valor adequado do fator (no caso da comunidade de biologia, o valor encontrado foi 0.8) foram efetuados testes, nos quais foi gerada uma lista com as palavras e uma verificação manual comparando o significado da palavra com o resumo da *tag* na comunidade em questão. O algoritmo foi executado com um conjunto pequeno de *tags* até que encontrasse apenas *tags* corretas. Para selecionar as *tags* dessa parte do experimento foi utilizado um arquivo CSV com 30 palavras e cada uma era conferida com a página da *tag* na comunidade, isso foi feito até que todas fossem as que estavam na comunidade. Essas palavras foram selecionadas aleatoriamente e conferidas com a definição na comunidade.

Após a segunda etapa, é gerada uma versão preliminar da árvore que contém as *tags* encontradas nos passos um e dois. Essa árvore é constituída a partir das listas de hiperônimos até a raiz da *WordNet*. Para cada *tag* que apresenta uma palavra correta é feito um *merge* a partir dos nós em comum. Conforme a figura 3.3, a *tag* "E" deve ser adicionada à árvore, então é criada uma lista L com todas as palavras até a raiz da árvore da WordNet. A partir do primeiro elemento da lista L, que é o nó raiz da árvore da WordNet, é feita

uma busca nó a nó da lista, até que ela fique vazia. No exemplo da Figura 3.3, dois nós já estavam na árvore e apenas o último nó é adicionado na árvore.

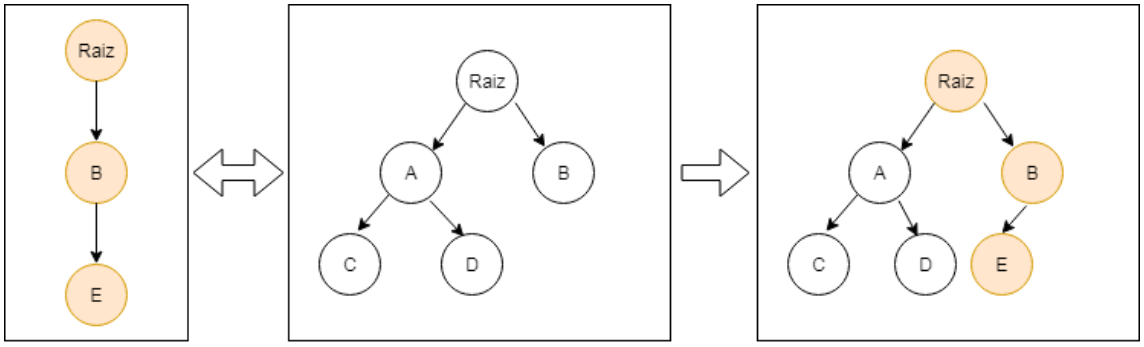

Figura 3.3: Adicionando um nó na árvore

A terceira etapa utiliza a árvore parcial para buscar as *tags* que ainda não foram encontradas. No caso, é feita a verificação da semelhança do caminho da palavra até a raiz da árvore, ou seja, uma lista com os nós da palavra até a raiz da árvore. Para cada palavra é feita uma nova lista e aquela que possuir mais itens que estão na lista da palavra é relacionada, uma vez que, de acordo com os experimentos, as palavras que possuem poucos itens na árvore, provavelmente não tem relação com o conceito que se está buscando. Por exemplo, uma palavra P que tem os nós A, B, C, D e ROOT, em comparação com outros dois nós X que tem os nós X1, X2, C, X3 e ROOT e Y que tem os nós A, B, Y1, Y2 e ROOT, Y tem mais possibilidade de ser a palavra correta que X.

A cada passo é necessário acessar a base para saber quais palavras são contexto da *tag* atual, as palavras que aparecem em conjunto com a *tag* em questão, sendo também necessário obter os synsets na WordNet e caso não seja encontrada somente um elemento, deve ser acionado o desambiguador.

3.2 Estudo de caso

Para verificar o processo de estruturação, foi realizado um estudo de caso utilizando as *tags* da comunidade de biologia da Stack Exchange [1], uma rede de comunidades na qual cada comunidade trata de um assunto específico. Conforme a Tabela 3.1, alguns dos assuntos tratados são:

Assunto	Nº usuários	Nº perguntas	Nº respostas
história geral[2]	≈ 21mil	≈ 8mil	≈ 17mil
biologia[3]	≈ 30mil	≈ 20mil	≈ 23mil
química[4]	≈ 41mil	≈ 26mil	≈ 31mil
matemática[5]	≈ 466mil	≈ 987mil	≈ 1.4milhões
programação[6][7]	≈ 8milhões	≈ 16milhões	≈ 24milhões

Tabela 3.1: Tabela comunidades Stack Exchange

Existem outras comunidades como a Super User [8], em que usuários com conhecimentos de infraestrutura podem fazer perguntas; a Ask Ubuntu[9] na qual usuários fazem perguntas sobre o sistema operacional Ubuntu[10] e várias outras comunidades com objetivos mais específicos como a TeX[11], que discute sobre o LaTeX[12]; a Electrical Engineering[13] que aborda temas relacionados a eletrônica e engenharia elétrica, entre outras.

Cada comunidade do Stack Exchange possui três sites, um para a comunidade em si, um site para que os usuários possam discutir sobre a própria comunidade, chamado de "meta", cujo acesso pode ser feito adicionando a palavra meta na url do site, por exemplo: o site meta para a comunidade Ask Ubuntu é `https://meta.askubuntu.com/` e, ainda, outro site sempre presente é o chat, onde os usuários podem conversar livremente sobre as perguntas e assuntos da comunidade. Nesse caso, basta adicionar a palavra chat na

[1] https://www.stackexchange.com/
[2] https://history.stackexchange.com/
[3] https://biology.stackexchange.com/
[4] https://chemistry.stackexchange.com/
[5] https://math.stackexchange.com/
[6] https://stackoverflow.com/
[7] https://pt.stackoverflow.com/
[8] https://superuser.com/
[9] https://askubuntu.com/
[10] https://www.ubuntu.com/
[11] https://tex.stackexchange.com/
[12] https://www.latex-project.org/
[13] https://electronics.stackexchange.com/

URL da comunidade, ou seja, para a comunidade Ask Ubuntu ficaria `https://chat.askubuntu.com/`.

A comunidade de Biologia do Stack Exchange utilizada como fonte de dados deste estudo, foi criada no dia 14 de dezembro de 2011 e nos primeiros 15 dias de sua criação, 268 usuários foram cadastrados e houve um aumento de aproximadamente 40% por ano. Nos primeiros dias da comunidade, foram criadas 127 perguntas e 188 respostas e houve um crescimento em média de 25% de perguntas e 19% de respostas por ano.

Nessa comunidade existem 703 *tags*, incluindo adjetivos e substantivos. De acordo com Golder e Huberman [17], *tags* subjetivas tendem a ser adjetivos e trazer a opinião do usuário que está usando a *tag* enquanto substantivos descrevem melhor o conhecimento [25]. Assim, como algumas *tags* da comunidade de Biologia do Stack Exchange são adjetivos, não entraram na árvore que foi construída.

Os dados da comunidade foram obtidos a partir dos arquivos disponibilizados pelo próprio *Stack Exchange* no Internet Archive [14][22]. Os arquivos disponibilizados têm um layout padrão, o que facilita a automatização do download e posterior tratamento dos dados. Os arquivos disponibilizados são:

Badges.xml contém as medalhas que os usuários já ganharam;

Comments.xml contém os comentários das perguntas e questões da comunidade;

PostHistory.xml contém um histórico de mudanças dos posts da comunidade;

Posts.xml contém os posts da comunidade;

Tags.xml contém as *tags* da comunidade;

Users.xml contém os usuários da comunidade;

Votes.xml contém os votos dos usuários nas perguntas e respostas da comunidade.

Os arquivos estão no formato XML, o que facilita na hora de efetuar a carga. Os arquivos são estruturados da seguinte forma: o elemento raiz é sempre o objeto do arquivo,

[14] https://archive.org/details/stackexchange

por exemplo, no arquivo Badges.xml, o elemento raiz será <badges> e ele terá uma coleção de elementos <row>.

Foram utilizados *scripts* utilitários disponibilizados no repositório Networks-Learning [15] no *Github* [16] , que geram tabelas no Sistema Gerenciador de Banco de Dados *PostgreSQL*. Os *scripts* geram uma tabela para cada arquivo disponibilizado pelo Stack Exchange e algumas visões (*views*) foram criadas para auxiliar no trabalho, foram elas:

v_post_tags_co_occurrence mapeia quais *tags* já foram utilizadas em conjunto e quantas vezes

v_tags_context com base na view v_post_tags_co_occurrence é feita uma view com as *tags* e a palavra da WordNet que já foi encontrada

Os scripts criam algumas tabelas que não estão presentes nos arquivos que o Stack Exchange disponibiliza:

PostTypes que contém os tipos de post: pergunta, resposta, item etc.

PostTags cria uma tabela com as *tags* por post

CloseAsOffTopicReasonTypes tipos de fechamento de post

FlagTypes tipos de flags que os usuários podem marcar um post

PostHistoryTypes tipos de alterações nos posts

CloseReasonTypes tipos de razões para fechamento de posts

VoteTypes tipos de votos

O primeiro passo foi carregar a base de dados com os dados da comunidade, em seguida foram executados os passos do processo que foi implementado utilizando a linguagem Python. Para acessar os dados foi utilizado o ORM Peewee [17], que é de simples utilização, facilitando o trabalho com SQL e a atualização da base de dados.

[15] https://github.com/Networks-Learning/stackexchange-dump-to-postgres
[16] https://github.com
[17] http://docs.peewee-orm.com/

Para criar a árvore foi utilizada a *WordNet*, que provê uma definição curta, exemplos de uso de cada palavra e também faz um relacionamento entre os termos do grupo.

Para utilizar o WordNet com o Python foi empregado um corpus do NLTK [18]. Esse corpus disponibiliza a última versão da WordNet e implementa uma série de algoritmos que facilitam o trabalho, além de ser código aberto[19].

Uma das dificuldades de mapear *tags* para a *WordNet* são as palavras que não fazem parte do vocabulário. Mesmo após o processo de *stemming*, algumas palavras não são mapeadas. A comunidade tem 703 *tags* e dessas foram mapeadas 462 (65%). Estudando esses dados foi observado que as *tags* mais populares tem grande probabilidade de terem sido mapeadas.

A Figura 3.4 mostra a probabilidade de uma *tag* pertencer à *WordNet* em função de sua popularidade. O eixo X representa as *tags* do *dataset*, em grupos de 75 e ordenadas pela popularidade decrescente. O eixo Y mostra a quantidade de *tags* que pertencem a *WordNet* em cada grupo. As *tags* que não foram encontradas poderiam ser buscadas utilizando-se ontologias de domínio mais específicas.

Foi necessário alterar os delimitadores em palavras compostas como "-" nas *tags* da comunidade. É o caso da *tag* "infectious-disease", que apresentou a necessidade de ser buscada na *WordNet* utilizando "_" ao invés de "-". Algumas palavras em minúsculas na comunidade, como o caso da *tag* "aids", estavam em maiúsculas na *WordNet* e não era possível apenas colocar todas as palavras em maiúsculo ou minúsculo, como é o normal nesses casos, pois não existe uma forma de se buscar uma palavra no WordNet ignorando minúsculas e minúsculas. As *tags* da comunidade já estão em minúsculo e, no caso da comunidade do Stack Exchange, apenas o caractere "-" é utilizado para a separação de palavras, não sendo necessário tratamento em relação a caracteres especiais nessa comunidade.

Os relacionamentos na *WordNet* não são definidos entre palavras, mas sim entre *synsets*, grupos de sinônimos que representam unidades de significado. Por exemplo, a pa-

[18] http://www.nltk.org/
[19] https://github.com/nltk/nltk

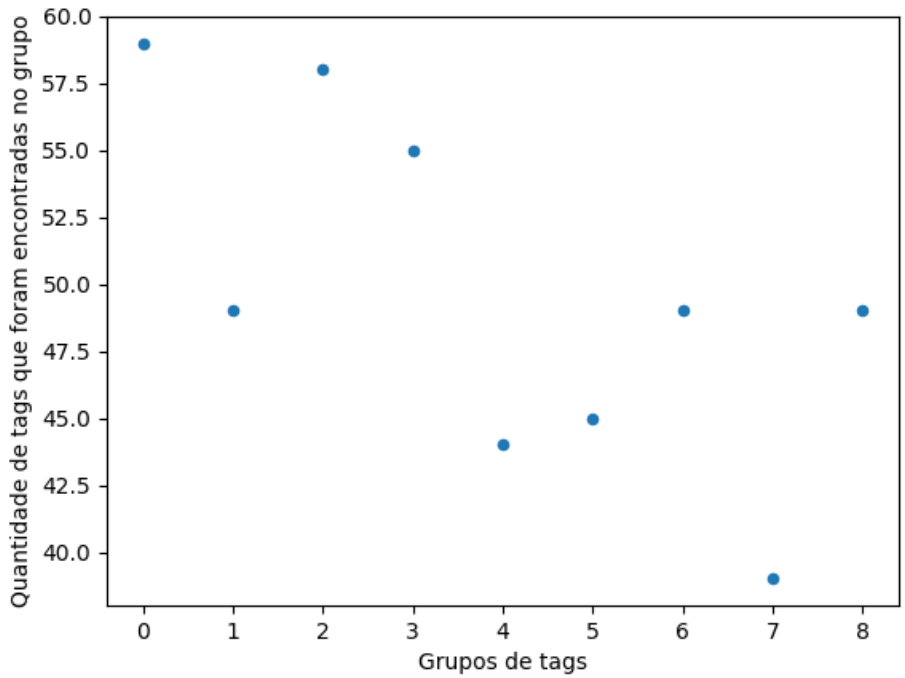

Figura 3.4: A imagem mostra a probabilidade de uma *tag* pertencer a *WordNet*, o eixo y representa a quantidade de palavras do grupo que foram encontradas e o eixo x representa os grupos.

lavra "*eyes*" pertence a 8 *synsets*, sendo 7 substantivos e 1 verbo , como visto na Figura 3.5, o primeiro substantivo é "*the organ of sight*" e o último "*(opinion or judgment) 'in the eyes of the law'; 'I was wrong in her eyes'*".

Para mapear corretamente uma *tag* para a palavra correta na *WordNet*, é necessário executar um processo de desambiguação que utiliza o contexto em que as *tags* são empregadas. As *tags* contexto utilizadas são as *tags* que aparecem junto com a *tag* em outras perguntas, por exemplo, o contexto da *tag* "*eyes*" são as *tags* "*vision*", "*human-eye*", "*light*", "*evolution*", entre outras.

Algumas palavras não foram mapeadas por serem adjetivos, é o caso das *tags*: *autoimmune* e *vestigial*. Por não se tratar de substantivos, nesses casos, a *WordNet* provê um método com as palavras do lugar onde o adjetivo é derivado. Por exemplo, para o *synset* da palavra *vestigial* existe um sinônimo *rudimentary* que é derivado de *rudiment*, dessa forma, se o processo for executado utilizando esse adjetivo, então essa palavra seria adicionada na árvore.

Figura 3.5: A imagem mostra os synsets para a palavra eyes.

A Figura 3.6 é um fragmento da árvore gerada com nós que fazem parte da taxonomia e nós que servem apenas para ser possível gerar a hierarquia. Os nós em preto representam essa última característica (nós para possibilitar a hierarquia), enquanto os nós vermelhos fazem parte da folksonomia. Podemos perceber que na própria folksonomia existem relacionamentos não explorados, como é o caso da *tag* "infectious_disease.n.01" que possui um relacionamento de pai e filho com as *tags* "hepatitis.n.01" e "tuberculosis.n.01". Com essa informação, poderíamos, por exemplo, buscar todas as perguntas da *tag* pai e trazer junto as perguntas das *tags* filhas. Outra informação que poderia ser explorada diz respeito a *tag* "communicable_disease.n.01" que é mais abrangente e traria ainda mais perguntas para o usuário.

No final, é gerada uma árvore com as *tags*. Das 241 *tags* que não foram mapeadas, 158 (65%) são *Multiword expressions* que não possuem um conceito correspondente na

WordNet. As outras são adjetivos ou são termos específicos da área da comunidade, como "zygosity" e "lentivirus", que não apresentam uma palavra relacionada na *WordNet*. Segundo [32], *multi-word expressions* são unidades léxicas que têm mais que uma palavra, com significado idiomático e composicional.

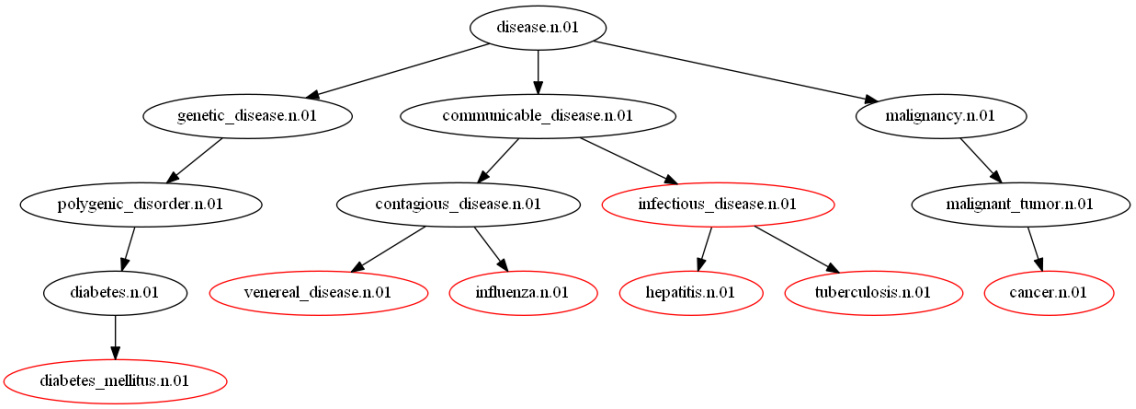

Figura 3.6: Árvore

4. AVALIAÇÃO

A seguir, detalha-se a avaliação do processo de estruturação de folksonomias. Foram implementados, além do estudo de caso, dois algoritmos para a etapa de desambiguação, um selecionando uma palavra aleatoriamente e outro que recupera a palavra mais utilizada.

4.1 Algoritmos

Nesta seção serão descritos alguns algoritmos que foram utilizados para executar a avaliação.

4.1.1 Aleatório

Esta abordagem visa selecionar aleatoriamente entre as palavras candidatas. Deste modo, será possível analisar comparativamente os resultados de uma seleção aleatória e uma seleção com base na abordagem proposta. Por exemplo, a palavra ”eye” tem 6 substantivos possíveis, no caso, o algoritmo escolhe um deles e retorna como palavra correta. Para gerar a posição aleatória, usada para definir a palavra, foi utilizado o método *randint* do módulo *random* que gera um número pseudo-aleatório. O método recebe dois parâmetros e retorna um número N tal que a <= N <= b, onde a e b são os parâmetros do método[1].

[1] https://docs.python.org/3/library/random.html#random.randint

4.1.2 Palavra mais utilizada

Em algumas implementações que adotam o WordNet, os desenvolvedores utilizam a palavra mais utilizada para resolver os problemas. Essa palavra é sempre a primeira do synset. A frequência de uso é determinada pelo número de vezes que a palavra é marcada nos vários textos de concordância semântica. Palavras que não são marcadas semanticamente seguem os sentidos ordenados[2]. Deste modo, será possível analisar comparativamente os resultados de uma seleção com base nesta estratégia e uma seleção com base na abordagem proposta.

4.2 Verificando qual o melhor resultado

Para avaliar os resultados foram selecionados especialistas para verificar qual implementação retorna o melhor resultado. No caso, são exibidas as palavras que cada algoritmo selecionou para um especialista indicando qual a palavra correta para aquele caso.

Para executar a avaliação do estudo foi utilizado o serviço de *crowdsourcing (CS)* CrowdFlower[3], uma plataforma de micro-tarefas que distribui as tarefas ao longo de vários canais de crowdsourcing como o Mechanical Turk e o Crowd Guru.

Um requisitante inicia submetendo um trabalho para o CrowdFlower. Para iniciar um projeto, o sistema exibe uma série de templates conforme Figura 4.1, o usuário seleciona um desses templates e então começa a configurar o projeto da maneira que desejar. Nesse template, o usuário configura um conjunto de microtarefas, também conhecidas como unidades.

Algumas das unidades podem ser unidades de referência, atribuídas com uma resposta correta e servem para avaliar o trabalho feito pelos colaboradores. Se um colaborador dá uma resposta errada a uma unidade de referência, o grau de confiança dele diminui até que suas respostas não sejam contabilizadas nos resultados do trabalho final.

[2]https://wordnet.princeton.edu/documentation/wn1wn
[3]https://www.crowdflower.com

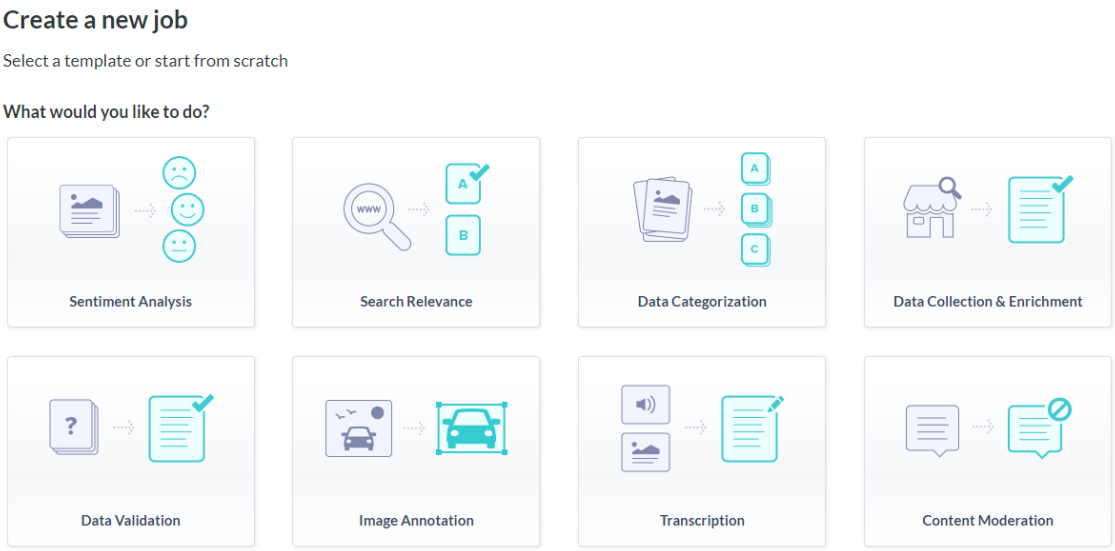

Figura 4.1: Seleção de template CrowdFlower

Além da especificação de parâmetros de configuração comuns como o IP do colaborador e de que país ele é, o CrowdFlower disponibiliza a criação de layouts definidos em linguagem própria o CrowdFlower Markup Language (CML). Essa linguagem abstrai os objetos HTML, permitindo uma interação com os dados que pode ser carregada para o CrowdFlower por meio de um arquivo em um dos seguintes formatos: CSV, TSV, XLS, XLSX, ODS conforme Figura 4.2.

O CrowdFlower disponibiliza um arquivo de exemplo que pode ser utilizado como base para o envio do dataset para o sistema.

Uma vez que o trabalho esteja finalizado, um relatório é gerado para o usuário solicitante e além disso, os artefatos são disponibilizados em arquivos CSV. Os artefatos do CrowdFlower são agregados baseando-se em forma de votação por maioria que pode eliminar respostas conforme as unidades de referência que foram fornecidas.

Um colaborador executa um trabalho, dando respostas sobre um determinado conjuntos de unidades expostas em uma página e se consentido, pode enviar uma quantidade de respostas igual à quantidade de unidades.

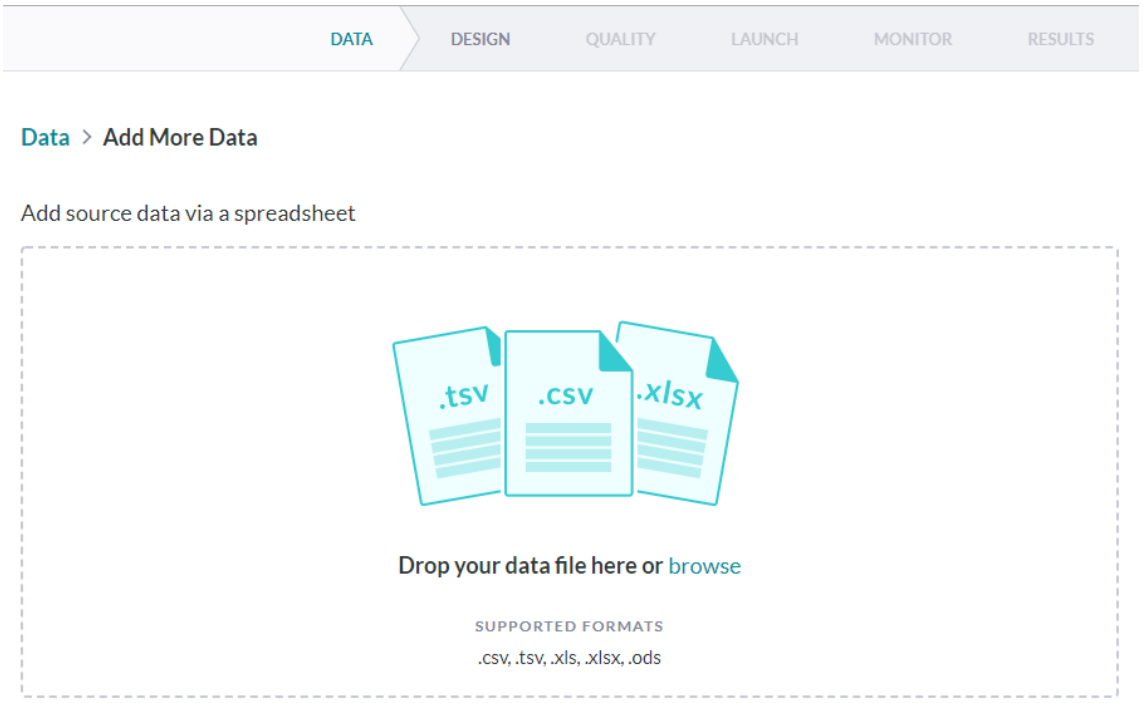

Figura 4.2: Carregar dados para o serviço

4.3 Questionário

Para realizar a avaliação foi gerado um dataset com as palavras que precisam de desambiguação presentes na WordNet. Então, os colaboradores do CrowdFlower selecionaram quais as palavras que se encaixam na definição da *tag*, na comunidade em relação a definição das palavras na WordNet.

Para carregar os dados no CrowdFlower foi necessário utilizar a API disponibilizada em conjunto com um script em Python, o que facilitou o trabalho em relação a carga dos dados e agilizou a criação das perguntas.

Para a elaboração do questionário foi necessário utilizar a sintaxe do próprio serviço, já que as opções eram dinâmicas de acordo com a tag. A Figura 4.3 é a montagem de cada pergunta do questionário que é exibido para os colaboradores. Como não havia um valor para cada opção e utilizar o texto seria insatisfatório para comparar e deixaria um ponto de duplicação de dados, foi utilizado como valor para a resposta das perguntas o índice na lista de opções. Por exemplo, para uma palavra com 3 opções, os valores seriam: 0, 1 e

2. Utilizar o índice da lista facilita no momento de computar os resultados, uma vez que se tem a posição da palavra na WordNet a partir deste índice.

Figura 4.3: Criando questionário dinâmico

Para verificar se o colaborador estava apto a responder ao questionário foi necessário criar algumas perguntas de teste, no próprio CrowdFlower. O próprio sistema indica uma quantidade de perguntas que devem ser adicionadas como teste e, de acordo com o andamento do trabalho, o sistema solicita que sejam criadas mais perguntas. Caso algum colaborador discorde da pergunta de teste, ele pode abrir uma contestação e, se for o caso, a pergunta pode ser corrigida ou a contestação pode ser ignorada.

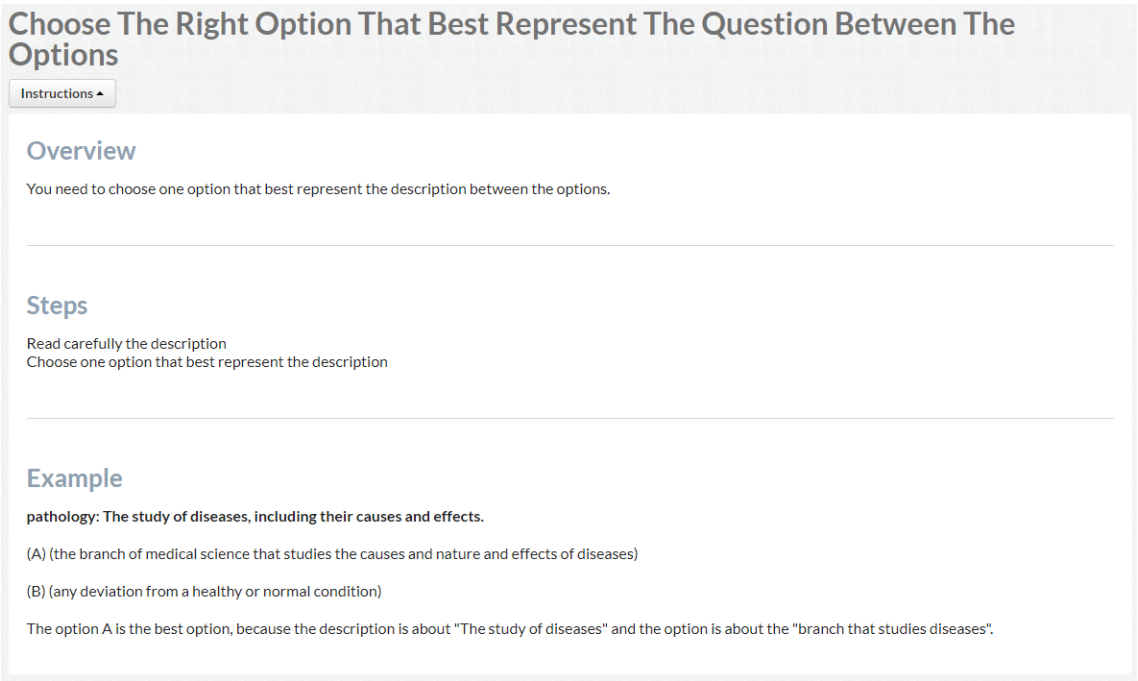

Figura 4.4: Instruções questionário

Para este trabalho, o CrowdFlower solicitou que fossem criadas 9 perguntas de teste e as *tags* são selecionadas aleatoriamente na amostra. Quando uma opção é selecionada

nas perguntas de teste, o pesquisador deve colocar um motivo pelo qual uma opção está sendo selecionada, conforme Figura 4.5.

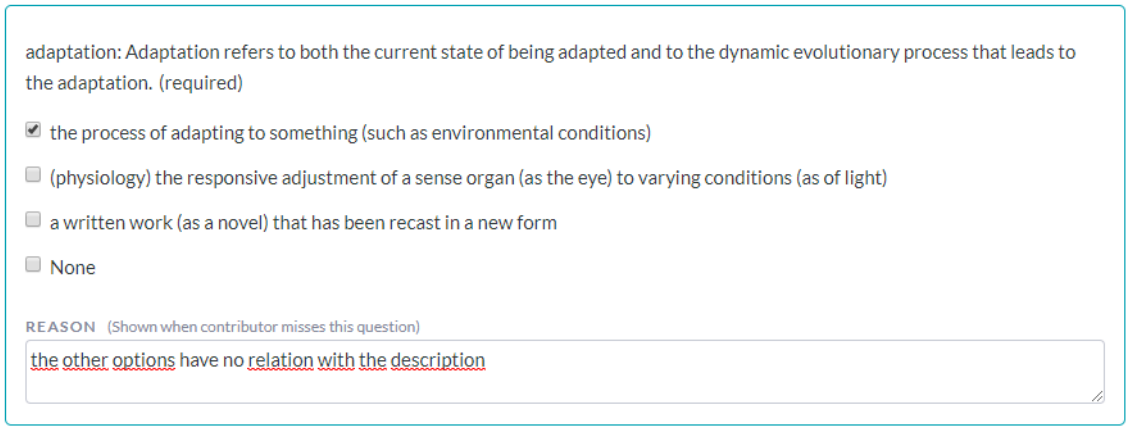

Figura 4.5: Criação pergunta teste

O questionário apresentado aos colaboradores exibia uma área com instruções: uma breve descrição do que o colaborador deve fazer, quais os passos o colaborador deve executar para alcançar o objetivo e um exemplo com o motivo pelo qual uma opção foi selecionada, conforme a Figura 4.4. Quando o colaborador seleciona uma resposta diferente daquela e não concorda com a justificativa, ele pode abrir uma contestação para que aquela pergunta de teste seja revista pelo pesquisador.

Após a exibição das instruções, é exibida uma pergunta com a definição de uma determinada *tag* e as opções disponíveis na WordNet, conforme figura 4.6. Caso o usuário não encontrasse a opção correta entre as opções ele poderia selecionar a opção ”None”.

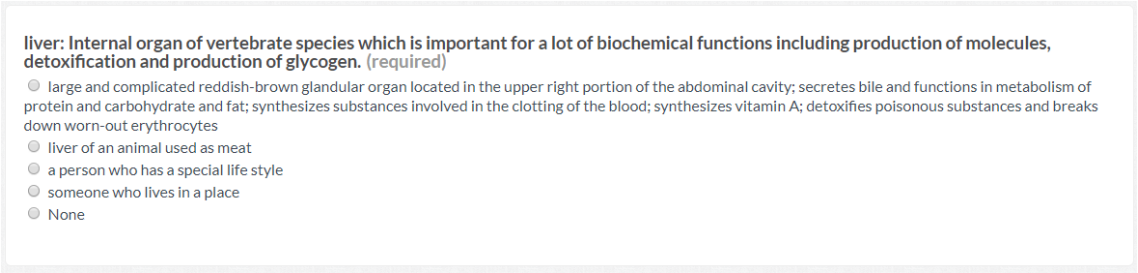

Figura 4.6: Pergunta questionário

4.4 Colaboradores

O CrowdFlower disponibiliza todos os microdados dos colaboradores que responderam ao questionário em formato CSV, sendo assim possível buscar informações sobre a nacionalidade dos colaboradores, sua cidade do colaborador, a partir de qual serviço o colaborador respondeu ao questionário, entre outras informações.

O serviço disponibiliza um arquivo com os dados agregados e com as respostas para cada linha que foi enviada também.

Em relação as perguntas de teste que verificam se um colaborador está apto a responder o questionário, verificou-se que apenas duas dessas perguntas tiveram suas respostas contestadas e em média 18% dos testes falharam. Cada pergunta teste foi testada em média 9 vezes em relação a todos os colaboradores que tentaram responder a pesquisa. As perguntas de teste eliminaram 30 colaboradores.

Os participantes que foram selecionados e enviaram respostas responderam cerca de 22 *tag*. Em média, esses usuários falharam nas perguntas teste uma vez e foram testados nessas perguntas 8 vezes e os colaboradores que a plataforma tem mais confiança foram aqueles que responderam mais perguntas.

4.5 Respostas

Em relação as respostas do questionário, das *tags* que foram enviadas para o serviço, apenas 11 *tags* não foram respondidas pelos colaboradores. Verificando os dados constatou-se que as definições das *tags* não eram muito claras ou não existia uma opção que batia exatamente com a descrição da *tag* na comunidade, das 11 *tags* que não tiveram uma resposta, 7 tiveram nível de confiança 100%, ou seja, os colaboradores não conseguiram escolher uma *tag* na relação exibida.

A distribuição das respostas em relação às opções ficou conforme Figura 4.7. A maior parte das respostas dos colaboradores ficou centralizada nas primeiras opções e verifica-

se também que na maioria das *tags* a primeira opção é a correta, ou seja, a heurística do trabalho pode ser ajustada para dar uma pontuação maior para a primeira opção.

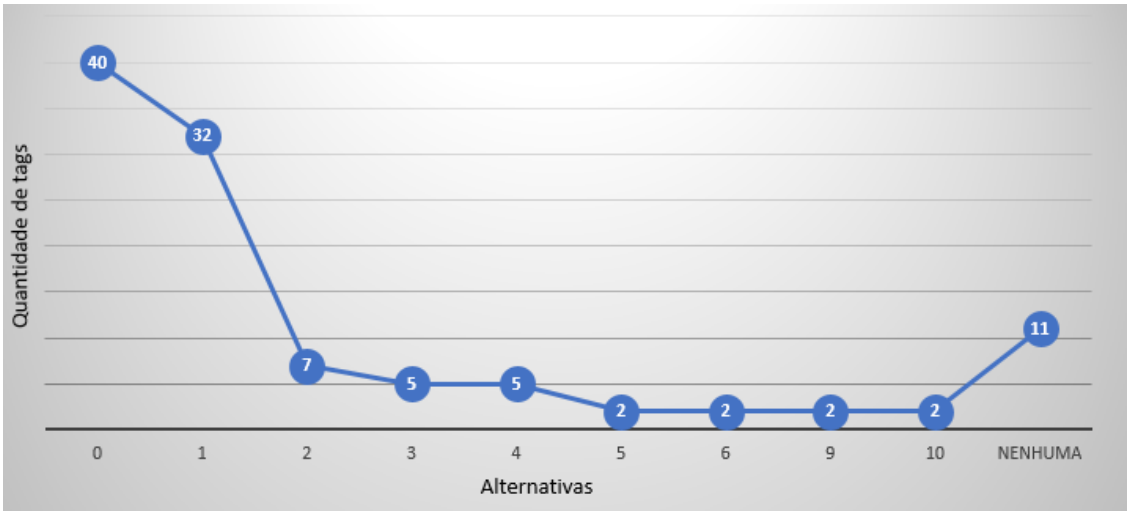

Figura 4.7: Distribuição das respostas em relação as opções

A Figura 4.7 apresenta para cada alternativa a quantidade de *tags* que foram selecionadas. No eixo x estão os índices das alternativas e no eixo y está a quantidade que o referido índice foi selecionado.

Para verificar se os colaboradores estavam de fato lendo a definição e selecionando a resposta que mais refletia a definição da tag, foi executada mais uma rodada de verificação no serviço. Porém, nessa segunda execução da verificação, as opções estavam em uma ordem diferente da exibição da WordNet, ou seja, a opção que era a primeira na WordNet foi colocada em outra posição quando a avaliação estava sendo executada, por exemplo, na quinta posição.

Essa execução foi necessária, pois a maioria das opções selecionadas na primeira execução foi a primeira ou a segunda opções.

De acordo com a Figura 4.8, verifica-se que as duas execuções geraram resultados parecidos, o que deixa evidente que as primeiras opções devem ser observadas com mais cuidado e devem ter uma pontuação maior na comparação com as outras opções.

A Figura 4.8, apresenta para cada alternativa a quantidade de *tags* que foram selecionadas. No eixo x estão os índices das alternativas e o eixo y está a quantidade que o referido

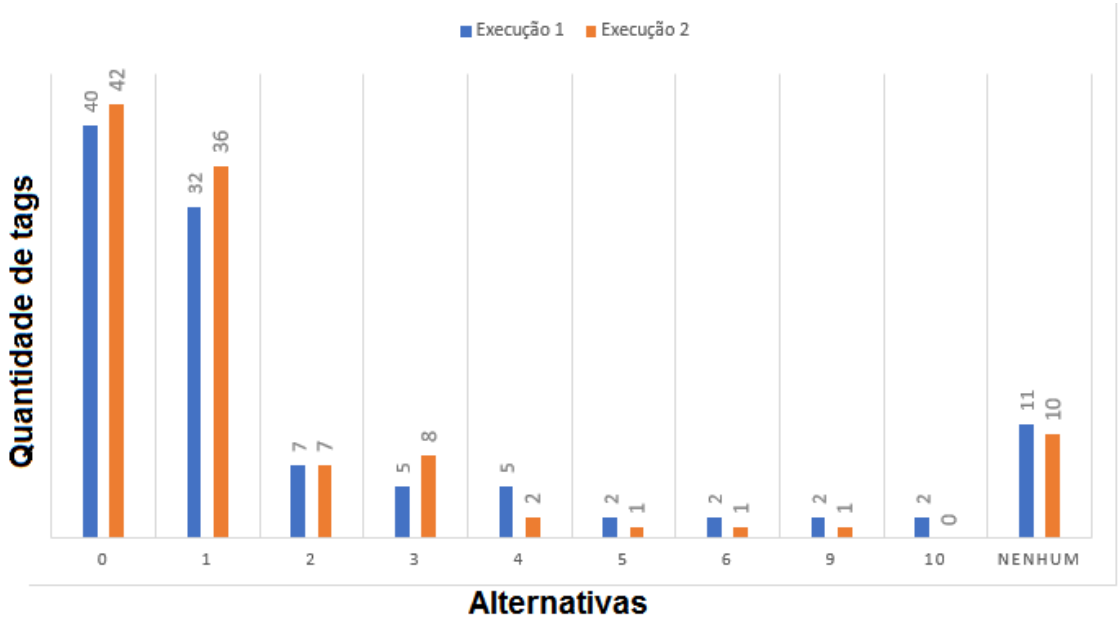

Figura 4.8: Comparação de execuções

índice foi selecionado. A primeira coluna representa a primeira execução da avaliação e a segunda coluna representa a segunda execução da avaliação.

4.6 Comparação entre as heurísticas

Conforme a Figura 4.9, verifica-se que o método deste trabalho conseguiu recuperar 52% de *tags* corretas em relação as *tags* enviadas para verificação no serviço Crowd-Flower, foram enviadas 100 *tags* para o serviço, sendo que 91 foram avaliadas pelos colaboradores e as outras foram utilizadas como pergunta de teste . A heurística que seleciona as opções aleatoriamente conseguiu 36% devido ao fato de a maioria das palavras terem 2 ou 3 opções.

A Figura 4.9, apresenta o percentual de acerto para cada heurística que foi implementada neste trabalho. O eixo x apresenta as heurísticas implementadas e o eixo y apresenta o percentual de acerto nas heurísticas.

Algumas *tags* ficaram com alternativas diferentes das selecionadas nas heurísticas, por exemplo, a *tag* “alcohol” possui descrição conforme Figura 4.10 e as opções para seleção eram: (A) “a liquor or brew containing alcohol as the active agent” e (B) “any of a series

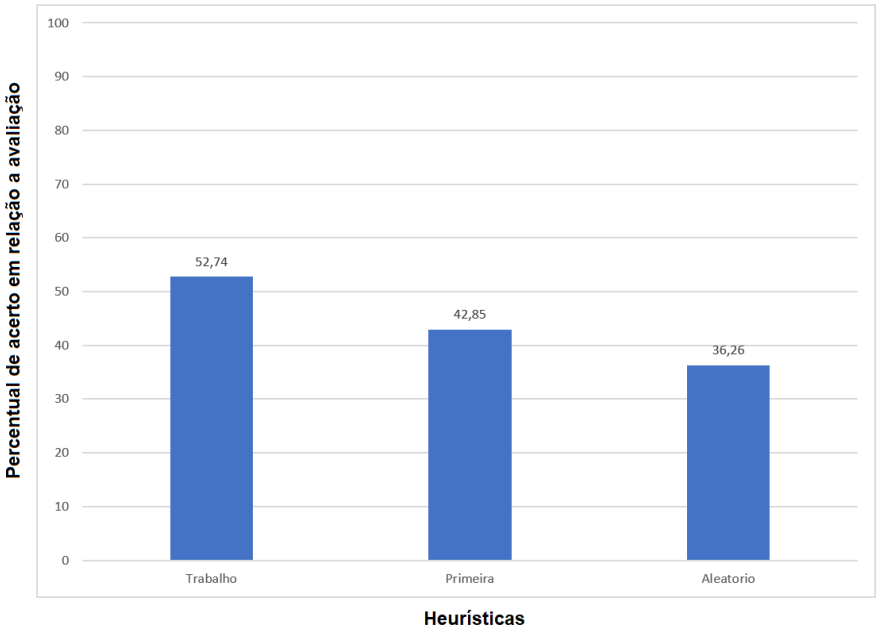

Figura 4.9: Comparação heurísticas

of volatile hydroxyl compounds that are made from hydrocarbons by distillation". Como a descrição não era muito clara, os colaboradores não souberam indicar qual a melhor alternativa. A primeira e a segunda tiveram 31% de seleções, ou seja, nesse caso não houve concordância em relação a melhor alternativa. Isso se deve principalmente a descrição pobre e sem recursos adicionais para que os avaliadores selecionassem a opção que melhor se encaixa neste contexto.

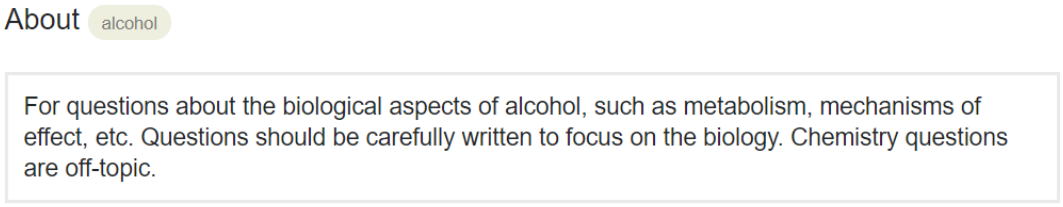

Figura 4.10: Descrição tag *alcohol*

Podemos verificar que a heurística que selecionava a primeira opção acertou e método do trabalho errou em 19 *tags* diferentes. Ao efetuarmos a verificação dos itens para melhor entender o resultado, verificou-se que dentre os itens existiam 6 onde havia opções ambíguas, por exemplo, para a *tag* "reproduction" a definição da comunidade conforme Figura 4.11, existem duas opções que poderiam ser selecionadas: "(the process of generating offspring)" e "(the sexual activity of conceiving and bearing offspring)", o método do

trabalho selecionou a segunda opção, porém pode-se perceber que ambas as opções são válidas para esta descrição.

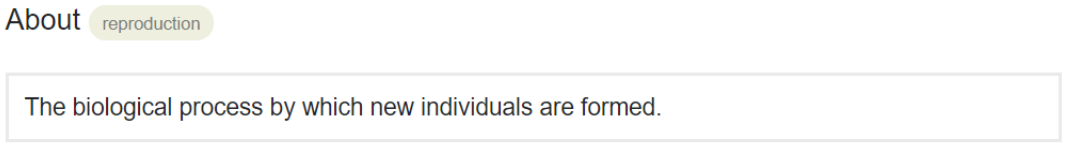

Figura 4.11: Descrição tag *reproduction*

Outro exemplo é a *tag* "radiation" que tem a descrição conforme Figura 4.11, o método do trabalho selecionou o synset "((medicine) the treatment of disease (especially cancer) by exposure to a radioactive substance)", porém quando verificamos as *tags* que aparecem em conjunto podemos encontrar a *tag cancer*, ou seja, com base nas próprias perguntas podemos verificar que muitas perguntas que discutem a doença são relacionadas a essa *tag* e não apenas uma "Ionizing electromagnetic waves".

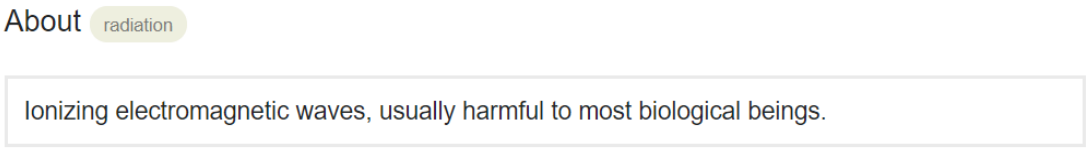

Figura 4.12: Descrição tag *radiation*

Citando caso outro exemplo, a *tag* "communication", que tem a descrição conforme Figura 4.13 e a primeira opção era "(the activity of communicating; the activity of conveying information)" e o trabalho selecionou "(a connection allowing access between persons or places)", desta forma podemos perceber que o método do trabalho errou em alguns casos, porém, nos casos onde existia opções ambíguas foram selecionadas opções que tinham mais relação com a descrição da comunidade.

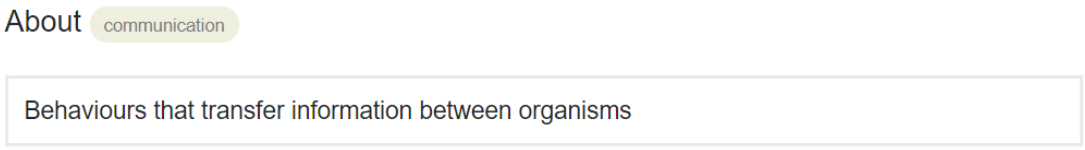

Figura 4.13: Descrição *tag communication*

Em 28 casos o método do trabalho selecionou as opções corretas e não eram a primeira opção nos synsets do WordNet, podemos verificar que quando não era a primeira opção, o

método encontrou termos que eram termos quase que exclusivamente do domínio exclusivo de Biologia, por exemplo, a *tag* "translation" tem a descrição conforme Figura 4.13, o primeiro synset tem a seguinte descrição "(a written communication in a second language having the same meaning as the written communication in a first language)" e o trabalho selecionou "((genetics) the process whereby genetic information coded in messenger RNA directs the formation of a specific protein at a ribosome in the cytoplasm)".

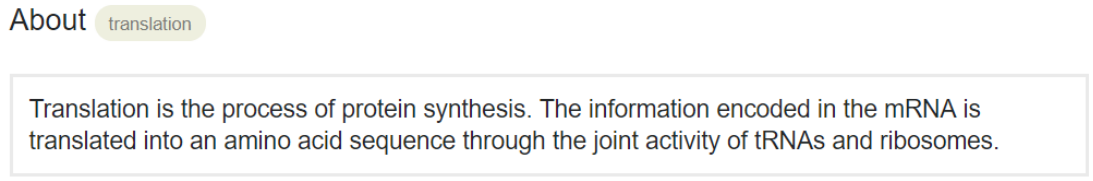

Figura 4.14: Descrição *tag translation*

Outro exemplo é a *tag* "transformation" que tem a descrição conforme Figura 4.15, o primeiro synset tem a seguinte descrição "(a qualitative change)" o trabalho selecionou "((genetics) modification of a cell or bacterium by the uptake and incorporation of exogenous DNA)".

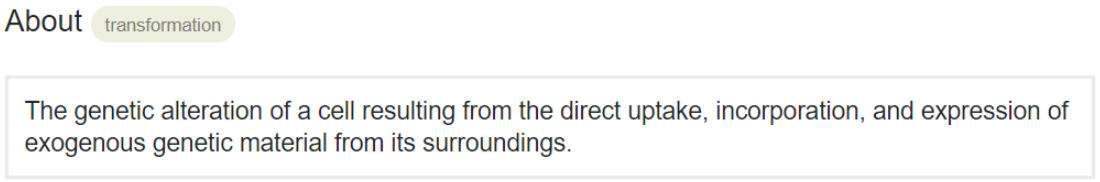

Figura 4.15: Descrição tag *transformation*

Em 24 casos o método do trabalho errou e a primeira opção não era a opção correta. Analisando os dados podemos perceber que em 7 casos, a descrição da *tag*, não era clara o suficiente, por exemplo, para a *tag* "medicine" a descrição da *tag* é conforme a Figura 4.16, na avaliação foi selecionada a opção "(the learned profession that is mastered by graduate training in a medical school and that is devoted to preventing or alleviating or curing diseases and injuries)" e o método do trabalho selecionou a opção "(the branches of medical science that deal with nonsurgical techniques)". Verificando as *tags* que aparecem em conjunto com a *tag* "medicine", foram encontradas as seguintes *tags*: "human-biology", "pharmacology", "pathology", "biochemistry", ou seja, a segunda opção melhor se encaixa com a descrição da comunidade, uma vez que, trata-se do ramo da ciência e não da

profissão.

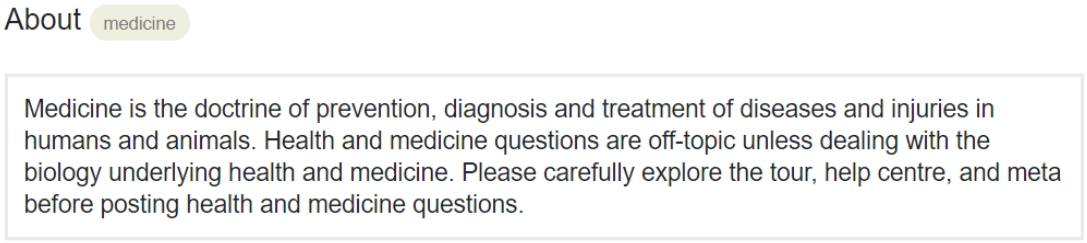

Figura 4.16: Descrição *tag medicine*

Outro exemplo é a *tag* "flowers" que tem a definição conforme a Figura 4.17 e os avaliadores corretamente selecionaram o synset "(reproductive organ of angiosperm plants especially one having showy or colorful parts)" e para este caso o método do trabalho selecionou "(a plant cultivated for its blooms or blossoms)", porém as seguintes *tags* são utilizadas em conjunto com esta *tag*: "botany", "species-identification", "plant-anatomy", "plant-physiology", "flowering", podemos perceber que ambas as opções válidas para esta descrição, uma vez que ambas tratam dos assuntos relacionados.

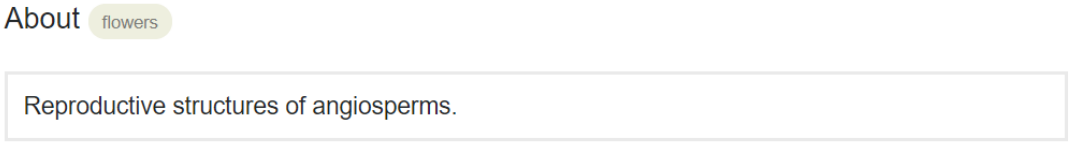

Figura 4.17: Descrição *tag flowers*

Ainda em relação as *tags* que o trabalho não selecionou o synset corretamente em relação a avaliação, podemos perceber que na *tag* "surgery", novamente a descrição não ajudou o avaliador a selecionar a opção corretamente, a descrição para a *tag* é conforme a Figura 4.18 e os avaliadores selecionaram a opção "(a medical procedure involving an incision with instruments; performed to repair damage or arrest disease in a living body)", o trabalho neste caso selecionou a opção "(the branch of medical science that treats disease or injury by operative procedures)", verificando mais atentamente esta *tag*, verificou-se que a opção do trabalho estava correta, pois o segundo texto da *tag* dizia o seguinte: "Surgery is the subspecialty of medicine, performed by surgeons, that deals with the operative management of pathologic conditions in humans and other animals, or to operations performed for experimental purposes.", ou seja, de fato, trata-se do ramo de medicina que

trata de cirurgias e não da prática de operar apenas.

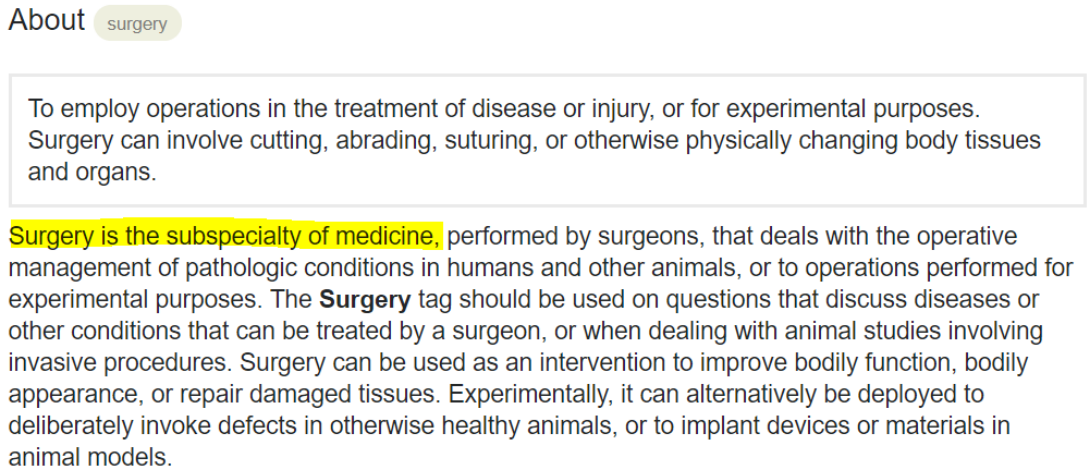

Figura 4.18: Descrição tag *surgery*

Outro fato que foi verificado que 5 das 24 *tags* foram utilizadas menos de 20 vezes na comunidade, por exemplo, a *tag* "review" foi utilizada apenas duas vezes, ou seja, não houve contexto suficiente para que o método do trabalho selecionar a tag corretamente. O método do trabalho selecionou o synset "(a subsequent examination of a patient for the purpose of monitoring earlier treatment)", porém a descrição da *tag* na comunidade é conforme a Figura 4.19 e as *tags* que foram utilizadas em conjunto foram apenas as *tags* "physiology" e "peer-review-journal", ou seja, o método do trabalho não teve contexto suficiente para ter uma decisão correta. Outro fato importante é que a *tag* "peer-review-journal" não está presente no WordNet e a *tag* "physiology" faz parte do WordNet, por isso, o trabalho selecionou a opção incorreta.

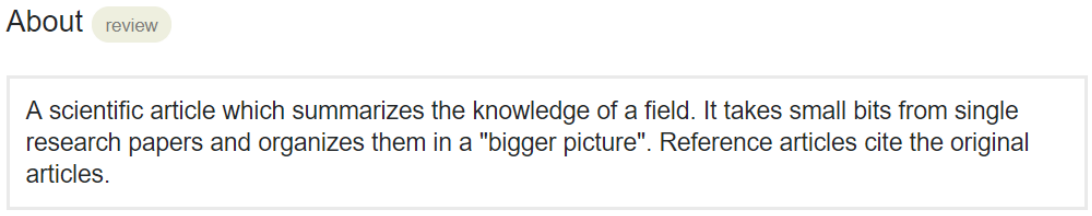

Figura 4.19: Descrição *tag review*

Nessa comparação verifica-se que pode ser feito um trabalho para melhorar as definições das *tags* no que diz respeito a sua definição, promovendo uma descrição mais clara para quais tipos de perguntas a *tag* deve ser usada pelos usuários. Existem *tags* como a

memory que tem a descrição conforme a Figura 4.20, que não tem uma descrição tão clara então pode haver confusão na sua utilização.

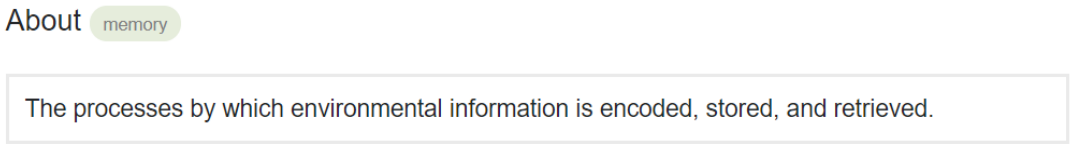

Figura 4.20: Descrição *tag memory*

A heurística do trabalho foi bem-sucedida nas *tags* com maior popularidade, as *tags* certas pela heurística do trabalho foram usadas em média 157 vezes enquanto as *tags* da heurística que selecionava a primeira opção foram usadas em média 98 vezes. Podemos constatar que quando as *tags* são específicas da área e ela tenha sido usada com *tags* que estão presentes no WordNet, faz com que o método do trabalho encontre a opção correta para uma determinada *tag*.

Para avaliar o resultado dos experimentos foram utilizadas métricas de recuperação de informação incluindo *Precision*, *Recall* e *F-measure*[31].

O método *Precision* mede o número de vezes que um *synset* foi selecionado corretamente pelo método do trabalho dividido pelo número de vezes que o *synset* foi selecionado no total. É necessário que o método do trabalho não selecione incorretamente os *synsets* para que o método tenha uma maximização no seu resultado. A principal desvantagem desta métrica é que ela não leva em conta os *synsets* que deveriam ter sido selecionados em uma *tag* específica, mas foi designada para outra.

O *Recall* mede o número de vezes que um *synset* foi selecionado para uma *tag* pelos avaliadores, porém o método do trabalho não selecionou o *synset* correto. A desvantagem dessa métrica é que, se o método do trabalho não selecionasse nenhuma *tag* incorretamente, o *Recall* seria máximo, muito embora não fosse eficiente.

Além do *Precision* e do *Recall*, usamos o *f-measure*, que corresponde à média harmônica entre *Precision* e *Recall*. Com essa informação, podemos dizer o desempenho do classificador apenas com um indicador. Como o *f-measure* é uma média, uma visão mais precisa da eficiência do classificador do que apenas o *Precision* e *Recall*.

Um resumo dos resultados dos experimentos é apresentado na Tabela 4.1, sendo possível verificar que a heurística do trabalho apresenta bons resultados em todas as medidas calculadas em relação as outras heurísticas.

Heurística	**Precision**	**Recall**	**f-measure**
Aleatorio	0.30	0.34	0.31
Primeira	0.50	0.55	0.52
Trabalho	**0.52**	**0.57**	**0.53**

Tabela 4.1: Resumo dos resultados

5. CONCLUSÃO

Neste capítulo serão apresentadas as conclusões e as principais contribuições desta dissertação, também serão sugeridos trabalhos futuros para a continuação da pesquisa realizada.

5.1 Comentários finais

A proposta desta dissertação foi apresentar uma forma de hierarquizar uma folksonomia utilizando apenas a própria *tag* e seu contexto, tendo como base a comunidade de biologia do Stack Exchange. Das 703 *tags* da comunidade foram encontradas 462 (65%) palavras correspondentes na *WordNet* e as *tags* que faltaram são *Multiword expressions* ou não apresentam palavra correspondente na *WordNet*. Das *tags* que não foram encontradas apenas 26 (0,03%) foram utilizadas pelo menos 100 vezes na comunidade.

Nos resultados, verificou-se que as *tags* selecionadas pela heurística do trabalho foram bem-sucedidas, uma vez que grande parte das *tags* avaliadas por colaboradores no serviço CrowdFlower foram as mesmas selecionadas pela heurística construída neste trabalho.

Verificou-se que a maior parte das respostas dos colaboradores ficou centralizada nas primeiras opções, sugerindo que as primeiras opções podem ter algum tipo de peso diferente na escolha das opções na etapa de desambiguação. Na segunda execução da avaliação, o mesmo comportamento foi verificado, reforçando a necessidade de mudar os pesos das primeiras opções quando a etapa de desambiguação for executada.

Quando são feitas comparações entre os métodos, verifica-se que o método construído neste trabalho é superior a selecionar a primeira opção, uma vez que, conseguiu recuperar mais respostas certas.

Diferentemente de outras propostas da literatura, nesta dissertação é apresentado um método para estruturação de uma folksonomia que considera a semântica da *tag* juntamente com as *tags* que aparecem em conjunto. Esse método ajudará na busca de perguntas de um nível superior, na sugestão de *tags* quando usuários criarem novas perguntas, sinalização de existência de *tags*, sinalização de *tags* que são sinônimos, entre outras.

5.2 Contribuições

A principal contribuição deste trabalho consiste na heurística apresentada para construir a hierarquia das *tags* da folksonomia da comunidade. Através desse método foi criada a hierarquia da comunidade de Biologia do Stack Exchange, levando em consideração a semântica da *tag* e a coocorrência, o que outros trabalhos da literatura não consideram.

Outras contribuições técnicas da pesquisa são:

- Implementação de scripts em Python para a extração de dados das comunidades;
- Implementação de algoritmos em Python para realizar as análises do trabalho;
- Levantamento bibliográfico de diversos trabalhos já realizados sobre o tema;
- Elaboração do questionário para avaliação do trabalho no CrowdFlower; e
- Construção de base de dados para carregar os dados da comunidade de Biologia.

5.3 Limitações

Este trabalho limitou-se a análise da comunidade de biologia do StackExchange e através dos resultados dessa comunidade foram derivadas as conclusões apresentadas. Entretanto, é sabido que a análise em outras comunidades é necessária para uma generalização

dos resultados encontrados.

Além disso, essa dissertação limitou-se somente a estruturação da hierarquia da folksonomia utilizando a WordNet como base de dados semântica. Seria interessante utilizar a dbPedia em conjunto com uma ontologia específica da área para encontrar mais termos e as expressões com mais de uma palavra que não estão presentes na WordNet.

5.4 Trabalhos futuros

Uma sugestão para um possível trabalho futuro seria adicionar uma forma mais flexível para a verificação de similaridade entre as palavras utilizando ontologias ou outros métodos de cálculo de similaridade. Existem outros métodos que verificam a similaridade entre duas palavras na *WordNet*, por exemplo, o método *jcn_similarity* que retorna uma nota de quão similar duas palavras são, baseando-se no ancestral mais específico entre as duas palavras. Outro método que poderia ser utilizado é o *lch_similarity* que retorna uma nota baseando-se no caminho mais curto que conecta as palavras e a profundidade máxima na taxonomia em que a palavra ocorre. Nos testes foi utilizado o método *lch_similarity*, porém o método *wup_similarity* demonstrou um resultado melhor na maioria dos casos, ainda que o método *lch_similarity* possa obter um resultado mais favorável em outras comunidades e outras folksonomias. Outra forma de verificar a similaridade seria buscar em ontologias de domínio específico por comunidade.

Outro trabalho futuro seria construir a árvore utilizando entidades do *DBpedia* ao invés da *WordNet*, o que poderia flexibilizar mais a construção da árvore e tornar possível o mapeamento das *tags* que não foram encontradas. Existe uma ontologia que mapeia a *WordNet*, podendo facilitar numa etapa de transformação dos nós da árvore de nós da *WordNet* para entidades na *DBpedia*. Isso poderia auxiliar na comparação entre as perguntas da comunidade, uma vez que poderia ser realizada uma comparação entre as entidades da *thread* da pergunta e efetuar uma comparação entre as *tags* adicionadas pelo usuário e as *tags* na hierarquia. Esse recurso pode facilitar na busca por especialistas, uma vez que as *tags* poderiam ser enriquecidas com as entidades relacionadas e algum tipo de compa-

ração poderia ser feita com os usuários que mais respondem àquelas *tags* e as entidades que esses usuários têm.

Outra ideia seria construir cada etapa como um bloco que pode ser usado em qualquer ordem e parametrizado e assim, promovendo uma facilidade quando o método for utilizado para construir a hierarquia de outras comunidades, aspectos como o fator, quantos nós pais devem ser buscados, entre outras opções poderia ser parâmetros no bloco e ele ser utilizado da maneira que o usuário quiser.

Outro trabalho seria a sugestão de recursos para melhorar as definições das *tags*, já que algumas importantes, como ”memory”, são muito genéricas e não definem bem em quais tipos de pergunta ela deveria ser usada.

Com a hierarquização das *tags* da comunidade seria possível prover um método para que seja calculado o nível de entropia de um usuário utilizando os níveis encontrados, tendo em vista que uma vez que a hierarquia é montada, seria possível montar os níveis de entropia para um usuário utilizando as perguntas que ele respondeu ou criou.

Ainda, outro trabalho futuro possível seria o cálculo automatizado do fator para a comunidade, o que facilitaria na aplicação do método em outras comunidades.

Referências Bibliográficas

[1] ACKERMAN, M. S., PIPEK, V., WULF, V., *Sharing expertise: Beyond knowledge management*. MIT press, 2003.

[2] ANDERSON, C. "The long tail", *Wired magazine* v. 12, n. 10, pp. 170–177, 2004.

[3] BAGHERI, E., ENSAN, F. "Semantic tagging and linking of software engineering social content", *Automated Software Engineering* v. 23, n. 2, pp. 147–190, 2016.

[4] BARTUSIAK, R., AUGUSTYNIAK, Ł., KAJDANOWICZ, T., et al. "Wordnet2vec: Corpora agnostic word vectorization method", *Neurocomputing*, , 2017.

[5] BEGELMAN, G., KELLER, P., SMADJA, F., et al. "Automated tag clustering: Improving search and exploration in the tag space". In: *collaborative web tagging workshop at WWW2006, Edinburgh, Scotland*, pp. 15–33, 2006.

[6] BEYER, S., PINZGER, M. "Synonym suggestion for tags on stack overflow". In: *Proceedings of the 2015 IEEE 23rd International Conference on Program Comprehension*, pp. 94–103, 2015.

[7] BROOKS, C. H., MONTANEZ, N. "Improved annotation of the blogosphere via autotagging and hierarchical clustering". In: *Proceedings of the 15th international conference on World Wide Web*, pp. 625–632, 2006.

[8] BROWN, J. S., DUGUID, P. "Knowledge and organization: A social-practice perspective", *Organization science* v. 12, n. 2, pp. 198–213, 2001.

[9] CAI, X., ZHU, J., SHEN, B., et al. “Greta: Graph-based tag assignment for github repositories”. In: *Computer Software and Applications Conference (COMPSAC), 2016 IEEE 40th Annual*, pp. 63–72, 2016.

[10] CHEN, C., LUO, P. “Enhancing navigability: An algorithm for constructing tag trees”, *Journal of Data and Information Science* v. 2, n. 2, pp. 56–75, 2017.

[11] CONSTANT, D., SPROULL, L., KIESLER, S. “The kindness of strangers: The usefulness of electronic weak ties for technical advice”, *Organization science* v. 7, n. 2, pp. 119–135, 1996.

[12] DONG, H., WANG, W., COENEN, F. “Rules for inducing hierarchies from social tagging data”. In: *iConference*, pp. 345–355, 2018.

[13] FELLBAUM, C. “A semantic network of english verbs”, *WordNet: An electronic lexical database* v. 3pp. 153–178, 1998.

[14] FELLBAUM, C. “Wordnet”. In: , pp. 231–243, Springer, 2010.

[15] GANGEMI, A., GUARINO, N., MASOLO, C., et al. “Sweetening ontologies with dolce”. In: *International Conference on Knowledge Engineering and Knowledge Management*, pp. 166–181, 2002.

[16] GANGEMI, A., GUARINO, N., OLTRAMARI, A., et al. “Cleaning-up wordnet’s top-level”. In: *Proceedings of the 1st International WordNet Conference*, pp. 21–25, 2002.

[17] GOLDER, S. A., HUBERMAN, B. A. “The structure of collaborative tagging systems”, *Journal of Information Science*, pp. 198–208, 2006.

[18] GOLDER, S. A., HUBERMAN, B. A. “Usage patterns of collaborative tagging systems”, *Journal of information science* v. 32, n. 2, pp. 198–208, 2006.

[19] GUY, M., TONKIN, E., OTHERS. “Tidying up tags”, *D-lib Magazine* v. 12, n. 1, pp. 1082–9873, 2006.

[20] HALPIN, H., ROBU, V., SHEPARD, H. "The dynamics and semantics of collaborative tagging". In: *Proceedings of the 1st semantic authoring and annotation workshop (SAAW'06)*, 2006.

[21] HOFFART, J., SUCHANEK, F. M., BERBERICH, K., et al. "Yago2: A spatially and temporally enhanced knowledge base from wikipedia", *Artificial Intelligence* v. 194pp. 28–61, 2013.

[22] JOORABCHI, A., ENGLISH, M., MAHDI, A. E. "Automatic mapping of user tags to wikipedia concepts: The case of a q&a website–stackoverflow", *Journal of Information Science* v. 41, n. 5, pp. 570–583, 2015.

[23] KOME, S. H. *Hierarchical subject relationships in folksonomies*. Dissertação de M.Sc., School of Information and Library Science, 2005.

[24] KROSKI, E. "The hive mind: Folksonomies and user-based tagging", *InfoTangle Blog, December* v. 9, 2005.

[25] LANIADO, D., EYNARD, D., COLOMBETTI, M., et al. "Using wordnet to turn a folksonomy into a hierarchy of concepts". In: *Semantic Web Application and Perspectives-Fourth Italian Semantic Web Workshop*, pp. 192–201, 2007.

[26] LEÃO, F. B. C., OTHERS. *Expanding the semantic knowledge of wordnet through semantic types and ufo*. Dissertação de M.Sc., UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO, 2014.

[27] LEE, S.-S., YONG, H.-S. "Tagplus: A retrieval system using synonym tag in folksonomy", *Journal of Digital Contents Society* v. 8, n. 3, pp. 255–262, 2007.

[28] LI, X., GUO, L., ZHAO, Y. E. "Tag-based social interest discovery". In: *Proceedings of the 17th international conference on World Wide Web*, pp. 675–684, 2008.

[29] LIN, H., FAN, W., ZHANG, Z. "Uncovering critical success factors for web-based knowledge communities". In: *Proceedings of the 16th Biennial Conference of the International Telecommunications Society, Beijing, China*, 2006.

[30] LIU, X., SONG, Y., LIU, S., et al. "Automatic taxonomy construction from keywords". In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1433–1441, 2012.

[31] MAKHOUL, J., KUBALA, F., SCHWARTZ, R., et al. "Performance measures for information extraction". In: *Proceedings of DARPA broadcast news workshop*, pp. 249–252, 1999.

[32] MASINI, F. "Multi-word expressions between syntax and the lexicon: the case of italian verb-particle constructions", *SKY Journal of Linguistics* v. 18, n. 2005, pp. 145–173, 2005.

[33] MIKA, P. "Ontologies are us: A unified model of social networks and semantics". In: *International semantic web conference*, pp. 522–536, 2005.

[34] MILLER, G., *WordNet: An electronic lexical database*. MIT press, 1998.

[35] MILLER, G. A. "Wordnet: a lexical database for english", *Communications of the ACM* v. 38, n. 11, pp. 39–41, 1995.

[36] MILLER, G. A., BECKWITH, R., FELLBAUM, C., et al. "Introduction to wordnet: An on-line lexical database", *International journal of lexicography* v. 3, n. 4, pp. 235–244, 1990.

[37] MORRIS, P. A. "Combining expert judgments: A bayesian approach", *Management Science* v. 23, n. 7, pp. 679–693, 1977.

[38] PETERS, I., STOCK, W. G. "Folksonomy and information retrieval", *Proceedings of the American Society for Information Science and Technology* v. 44, n. 1, pp. 1–28, 2007.

[39] PONZETTO, S. P., STRUBE, M. "Wikitaxonomy: A large scale knowledge resource.". In: *ECAI*, pp. 751–752, 2008.

[40] QUINTARELLI, E. "Folksonomies: power to the people. june 2005". In: *ISKO Italy-UniMIB meeting*, 2005.

[41] RABAN, D., HARPER, F. “Motivations for answering questions online”, *New media and innovative technologies* v. 73, 2008.

[42] ROSENBAUM, H., SHACHAF, P. “A structuration approach to online communities of practice: The case of q&a communities”, *Journal of the American Society for Information Science and Technology* v. 61, n. 9, pp. 1933–1944, 2010.

[43] SADIKOV, E., MADHAVAN, J., WANG, L., et al. “Clustering query refinements by user intent”. In: *Proceedings of the 19th international conference on World wide web*, pp. 841–850, 2010.

[44] SCHMITZ, C., HOTHO, A., JÄSCHKE, R., et al. “Mining association rules in folksonomies”. In: *Data Science and Classification: Proc. of the 10th IFCS Conf.*, pp. 261–270, Berlin, Heidelberg, 2006.

[45] SHEPITSEN, A., GEMMELL, J., MOBASHER, B., et al. “Personalized recommendation in social tagging systems using hierarchical clustering”. In: *Proceedings of the 2008 ACM conference on Recommender systems*, pp. 259–266, 2008.

[46] SI, X., LIU, Z., SUN, M. “Explore the structure of social tags by subsumption relations”. In: *Proceedings of the 23rd International Conference on Computational Linguistics*, pp. 1011–1019, 2010.

[47] SRIDHARA, G., HILL, E., POLLOCK, L., et al. “Identifying word relations in software: A comparative study of semantic similarity tools”. In: *Program Comprehension, 2008. ICPC 2008. The 16th IEEE International Conference on*, pp. 123–132, 2008.

[48] TAKEHARA, D., HARAKAWA, R., OGAWA, T., et al. “Extracting hierarchical structure of content groups from different social media platforms using multiple social metadata”, *Multimedia Tools and Applications* v. 76, n. 19, pp. 20249–20272, 2017.

[49] TIBÉLY, G., POLLNER, P., VICSEK, T., et al. “Extracting tag hierarchies”, *PloS one* v. 8, n. 12, p. e84133, 2013.

[50] TOMAÉL, M. I., VALENTIN, M. "Critérios de qualidade para avaliar fontes de informação na internet", *Avaliação de fontes de informação na Internet. Londrina: Eduel*, pp. 19–40, 2004.

[51] TRANT, J. "Studying social tagging and folksonomy: A review and framework.", *Journal of Digital Information* v. 10, n. 1, pp. 1–44, 2009.

[52] UNIVERSIDADE DE BRASÍLIA (UNB), *Relatório técnico contendo à Taxonomia desenvolvida para os órgãos da Justiça Militar*. Universidade de Brasília (UnB), 2016.

[53] VAN DAMME, C., HEPP, M., SIORPAES, K. "Folksontology: An integrated approach for turning folksonomies into ontologies", *Bridging the Gap between Semantic Web and Web* v. 2, n. 2, pp. 57–70, 2007.

[54] WAL, T. V. Folksonomy coinage and definition. http://www.vanderwal.net/folksonomy.html, 2007. Acessado 16/05/2018.

[55] WANG, G. A., JIAO, J., ABRAHAMS, A. S., et al. "Expertrank: A topic-aware expert finding algorithm for online knowledge communities", *Decision Support Systems* v. 54, n. 3, pp. 1442–1451, 2013.

[56] WASKO, M. M., FARAJ, S. "Why should i share? examining social capital and knowledge contribution in electronic networks of practice", *MIS quarterly*, pp. 35–57, 2005.

[57] WHITE, R. W., BENNETT, P. N., DUMAIS, S. T. "Predicting short-term interests using activity-based search context". In: *Proceedings of the 19th ACM international conference on Information and knowledge management*, pp. 1009–1018, 2010.

[58] WU, F., WELD, D. S. "Automatically refining the wikipedia infobox ontology". In: *Proceedings of the 17th international conference on World Wide Web*, pp. 635–644, 2008.

[59] WU, R., ZHANG, H., KIM, S., et al. "Relink: recovering links between bugs and changes". In: *Proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on Foundations of software engineering*, pp. 15–25, 2011.

[60] WU, Z., PALMER, M. "Verbs semantics and lexical selection". In: *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pp. 133–138, 1994.

[61] XIE, M., LU, J., CHEN, G., et al. "Folksonomy-based internet object profiling and relation extracting". In: *GLOBECOM 2017-2017 IEEE Global Communications Conference*, pp. 1–6, 2017.

[62] ZHANG, W., WATTS, S. "Knowledge adoption in online communities of practice", *Systemes d'Information et Management* v. 9, n. 1, p. 81, 2004.

[63] ZHOU, M., BAO, S., WU, X., et al. "An unsupervised model for exploring hierarchical semantics from social annotations". In: *The Semantic Web*, pp. 680–693, 2007.

[64] ZHU, J., SHEN, B., CAI, X., et al. "Building a large-scale software programming taxonomy from stackoverflow.". In: *SEKE*, pp. 391–396, 2015.

[65] ZHU, J., WANG, H., SHEN, B. "Software. zhishi. schema: A software programming taxonomy derived from stackoverflow.". In: *International Semantic Web Conference (Posters & Demos)*, 2015.