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

Estudo de Relações Semânticas entre Categorias Textuais de Perguntas e Respostas em Comunidades Q&A

Davi Faisca Duarte

Orientador

Dr. Sean Wolfgand Matsui Siqueira

Co-orientador

Dr. João Luís Tavares da Silva

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2018

Estudo de Relações Semânticas entre Categorias Textuais de Perguntas e Respostas em Comunidades Q&A

DAVI FAISCA DUARTE

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:

Dr. Sean Wolfgand Matsui Siqueira — UNIRIO

Dr. João Luís Tavares da Silva — UniFTEC

Dr. Bernardo Pereira Nunes — UNIRIO

Dr. Jairo Francisco de Souza — UFJF

RIO DE JANEIRO, RJ - BRASIL
SETEMBRO DE 2018.

Duarte, Davi Faisca
B118 Estudo de Relações Semânticas entre Categorias Textuais de Perguntas e Respostas em Comunidades Q&A / Davi Faisca Duarte, 2018.
xiii, 187f.

Orientador: Sean Wolfgand Matsui Siqueira.
Coorientador: João Luiz Tavares da Silva.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

1. Sistemas de Informação - Comunidades Online - Web Semântica
I. Siqueira, Sean Wolfgand Matsui. II Silva, João Luiz Tavares.
III. Universidade Federal do Estado do Rio de Janeiro (2003-).
Centro de Ciências Exatas e Tecnologia. Curso de Mestrado em Informática. Título.

CDD - 004.678

DUARTE, DAVI FAISCA **Estudo de Relações Semânticas entre Categorias Textuais de Perguntas e Respostas em Comunidades Q&A**. UNIRIO, 2018. 58 páginas. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

Plataformas online de Perguntas e Respostas (*Community Question Answering*) atuam como importante papel na educação, formadas por indivíduos com interesses comuns, mas com diferentes níveis de experiência, propiciam que os usuários aprendam colaborativamente através de perguntas e respostas espontâneas e específicas. Estas comunidades são populares e, no entanto, muitas questões permanecem sem resposta, ocasionando um problema de falta de *feedback* para o usuário, podendo gerar desmotivação e desinteresse, ou mesmo disponibilizando respostas erradas. O presente trabalho investiga alternativas para minimizar tais problemas através da proposta de uma abordagem de suporte para futuras metodologias automáticas de recomendação de respostas relacionadas. O principal objetivo deste trabalho é investigar se as similaridades semânticas entre categorias textuais de perguntas e respostas é suficiente para estabelecer relações semânticas entre eles e, assim, identificar relações entre melhores respostas ou textos relacionados ao tema da pergunta. Para isto, este trabalho propõe uma arquitetura genérica que integra três ferramentas de extração de categorias, o *DBpedia Spotlight*, *Open Calais* e *Tag The Web*, para calcular um *score* de similaridade semântica de categorias representando as perguntas e as respostas relacionadas e perguntas similares com outras repostas para medir a similaridade entre perguntas com respostas distintas. Os experimentos demonstraram que é possível estabelecer estas relações a partir da extração de categorias e utilizá-los como suporte a futuras recomendações de respostas para questões em aberto.

Palavras-chave: Web Semântica, Comunidades Q&A, Conectividade Semântica, Linked Data, Stack Exchange.

ABSTRACT

Online Question Answering platforms play an important role in education, formed by individuals with common interests but with different levels of experience, enabling users to learn collaboratively through spontaneous and specific questions and answers. These communities are popular but many questions remain unanswered, causing a problem of lack of feedback for the user, which can generate demotivation and disinterest, or even making the wrong answers available. In this research we propose to investigate alternatives to minimize such problems through the proposal of a support approach for future automatic methodologies of recommendation of related answers. The main objective of this research is to investigate if the semantic similarities between textual categories of questions and answers are enough to establish semantic relationships between them and, thus, to identify relationships between better answers or texts related to the theme of the question. For this, this work proposes a generic architecture that integrates three categories extraction tools, the *DBpedia Spotlight*, the *Open Calais* and the *Tag The Web*, to calculate a score semantic similarity of categories representing the questions and related answers and similar questions with other answers to measure the similarity between questions with different answers. The experiments demonstrated that it is possible to establish these relations from the extraction of categories and to use them as support for future recommendations of answers to open questions.

Keywords: Semantic Web, Q&A Communities, Semantic Connectivity, Linked Data, Stack Exchange.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

Neste capítulo serão apresentados os elementos que motivaram a realização deste trabalho. Desta forma, o capítulo tem como objetivo apresentar o problema a ser abordado, levantar a hipótese a ser investigada e mostrar os objetivos que serão exploradas no desenvolvimento desta dissertação.

1.1 Contextualização e Motivação

A rede mundial de computadores tem um papel importante na sociedade contemporânea, proporcionando inovações na maneira de se comunicar e trocar conhecimento [49]. A internet constitui-se em uma grande fonte de dados, pois naturalmente encontra-se uma gama de informações sendo ofertada em diferentes padrões e postada por um número considerável de pessoas que possuem propósitos diversos. O rápido e fácil acesso a essas informações é o que torna a Web tão fascinante para os indivíduos, consequência do seu forte avanço [4]. No caso de um eventual problema, existem dois paradigmas que podem ser utilizados para representar o processo de busca por uma solução na Web: o Paradigma da Biblioteca e o Paradigma da Vila [20].

O Paradigma da Biblioteca percebe a internet como uma enciclopédia, onde a procura por uma resposta rápida para o seu problema permanece em localizar a "página correta", que possua o conhecimento desejado [50]. Como uma forma de ajuda, pode-se manusear Ferramentas de Busca. Assim sendo, esses recursos são disponibilizados por organizações como Google[1], o Yahoo[2] e Microsoft[3]. Até o presente momento é a tática mais disseminada para se adquirir informação na internet na época atual [1].

No Paradigma de Vila, uma pessoa com uma questão desconhecida, precisa identificar

[1] https://www.google.com/
[2] https://www.yahoo.com/
[3] https://www.bing.com/

alguém que consiga lhe auxiliar, normalmente um dos participantes mais antigos e sábios da Vila. Na internet, o Paradigma da Vila ocorre como uma consulta social, que representa uma postagem de uma pergunta em alguma comunidade online e armazena respostas dos usuários [50]. A forma de partilhar suas dificuldades em um contexto de cooperação apareceu nos fóruns online e Sites de Perguntas e Respostas, porém foi disseminada também para outras Redes Sociais, como por exemplo o Facebook [15] e Twitter [14].

A Consulta Social é um esforço de modificar as relações sociais na busca de informação, de forma prática e oportuna, atingindo o benefício do "conhecimento da multidão" [28]. Partilhar perguntas na web é uma tática muito eficiente quando se trata de questões que demandam certo grau de individualização na resposta, visto que se aceita que os seus contatos disponham de dados individualizados a seu respeito [25,28]. Por outro lado, essa particularização nas respostas pode ser improvável de ser obtida por outras formas de busca e fontes de informação.

Outra forma encontrada para consulta social é o termo *Social Query* [5, 28, 44], que da mesma forma representa o procedimento de compartilhamento de perguntas em uma comunidade online e o recebimento de respostas. Huberman [21] e Mui e Whoriskey [29] declaram que cenários que possibilitam a construção de comunidades online com um número grande de pessoas, como por exemplo, Twitter e o Facebook, são espaços férteis e eficazes para localizar conhecimento por meio da utilização da *Social Query*. Dessa forma, constata-se que a existência de uma quantidade considerável de participantes nessas comunidades aumentaria a probabilidade de conseguir o conhecimento ou a resposta adequada.

Estes sites padrões de consulta não são os modos mais eficientes para procurar informações, visto que suas respostas muitas vezes não são direcionadas ao que se realmente esperava [33]. De acordo com Horowitz [20], determinadas questões são bem mais elucidadas por usuários que já vivenciaram tal acontecimento, por exemplo, em conteúdos específicos, solicitações de referências e sugestões. Uma explicação para isto, encontra-se no fato destes sites oferecerem um bom resultado nas buscas quando essas são realizadas num cenário conhecido e sem grandes alterações. Por outro lado, esses sites não possuem um desempenho eficaz em condições em que é necessário uma fundamentação ou algo específico. Ao se realizar uma consulta na web da palavra "manga", surgem resultados sobre a fruta ou parte de uma camisa [33]. Por causa disso, usam-se de forma mais eficiente as comunidades online de perguntas e respostas como *Stackoverflow*, *Quora* e *Yahoo! Answers*, por exemplo. Em relação a esses sites padrões de consulta, nelas os participantes perguntam e respondem de maneira espontânea. Ainda assim, há indivíduos que optam por fazer perguntas apenas a um grupo restrito de pessoas em lugar de compartilhar suas

dúvidas com pessoas estranhas em comunidades de perguntas e respostas [28].

Uma comunidade online é definida como um grupo de pessoas que interage em um ambiente virtual. Essas pessoas possuem um objetivo, utilizam o suporte de alguma tecnologia e são regidos por normas e regras [34]. Um tipo particular de comunidade online são as comunidades de Perguntas e Respostas, ou *Question and Answer (Q&A) Communities* (ou *Community Question Answering*), que são formadas por indivíduos com um interesse comum, porém com diferentes níveis de experiência

Estas plataformas online de Q&A atuam como importante papel na educação, principalmente no contexto de cenários de aprendizagem informal. Participantes de comunidade de Q&A colocam suas questões e obtém respostas, com *feedback* e sugestões de outros usuários em um curto espaço de tempo depois de publicar a sua questão a ser respondida. Esse sistema atua como um ambiente dinâmico que promove uma aprendizagem social e colaborativa, o que contribui para o crescente número de usuários para plataformas como o *Stack Exchange, Reddit, Quora e Yahoo Answering*, entre outros. Esses sites são populares e, no entanto, muitas questões permanecem sem resposta, ocasionando um problema de falta de *feedback* para o usuário, o que pode gerar desmotivação para uso da plataforma, e uma metodologia automática poderia recomendar respostas relacionadas.

1.2 Problematização

Quando um indivíduo se depara com um problema que não consegue resolver sozinho, ao invés de procurar sozinho por uma solução na Web, utilizando alguma ferramenta de busca tradicional, como o Google ou Bing, ele opta por tornar seu problema público, em alguma comunidade online, fazendo assim uma pergunta [20]. As perguntas são publicadas por participantes que apresentam um problema ou que possam ter interesses comuns, provocando uma discussão, enquanto outros participantes tentam ajudar, baseados em princípios como altruísmo, ganho de reputação, reciprocidade ou mesmo benefícios de aprendizagem direta [23, 24]. Aqueles que têm uma posição de prestígio na comunidade ou uma alta reputação são considerados especialistas e geralmente podem fornecer as respostas mais efetivas e favoritas.

Há diversos sites ou plataformas que permitem a organização de comunidades Q&A, como o Stack Exchange, Reddit, Quora e Yahoo Answering. Esses sites promovem um ambiente dinâmico que contribui para o crescente número de usuários. Assim, permitem que o usuário se aproveite do "conhecimento da multidão" [48], ao possibilitar que cada pergunta receba múltiplas respostas. Assim, ao receber diversas respostas para uma

pergunta, a resposta mais recorrente tenderá a ser tão boa quanto a resposta de um especialista.

No entanto, o problema de compartilhar perguntas em comunidades online é que não há garantias de que se receberá ajuda, nem de quanto tempo irá se passar até que alguém decida responder [45]. Em comunidades Q&A, o fluxo intenso de novas perguntas também pode dificultar que uma pergunta seja visualizada por alguém apto e consequentemente respondida [46].

No Stack Exchange, por exemplo, as perguntas normalmente são respondidas em um período curto de tempo. Porém, acompanhando o crescimento do site, já é perceptível que o percentual de perguntas que são ignoradas ou não resolvidas continua crescendo [47]. Alguns autores apontam alguns problemas nas Comunidades de Perguntas e Respostas, como: (i) a maior parte das perguntas não são respondidas em até dois dias [56]; (ii) aproximadamente 23% das perguntas não recebem uma resposta ou recebem respostas que não são classificadas uma boa resposta [42]; (iii) quando perguntas são feitas de forma repetida, os especialistas ficam menos inclinados a responder e os novos usuários são desencorajados a participar da comunidade quando as suas perguntas não são respondidas [42].

Em comunidades Q&A, normalmente, muitas perguntas não são respondidas e uma metodologia automática poderia recomendar respostas relacionadas. Mas para isto é preciso estabelecer relações semânticas adequadas entre domínios relevantes, relacionando os termos entre perguntas e respostas.

Este trabalho se propõe a identificar relações semânticas entre categorias textuais das perguntas e respostas em comunidades Q&A. Os resultados obtidos poderão ser utilizados como suporte para futuras recomendações de respostas para questões não respondidas em comunidades Q&A, sem a intervenção de especialistas.

1.3 Hipótese

Para solucionar o problema enunciado anteriormente, foi elaborada a seguinte hipótese que esta pesquisa investigará:

A hipótese é que SE a partir da similaridade semântica de categorias textuais entre perguntas e respostas for possível estabelecer relações semânticas entre eles, ENTÃO podemos relacionar melhores respostas ou textos relacionados ao tema da pergunta.

1.4 Objetivos

O *objetivo geral* deste trabalho é realizar um estudo a fim de identificar relações semânticas entre categorias textuais das perguntas e respostas em comunidades Q&A, através de um sistema/critério de avaliação de similaridade semântica entre as informações postadas de um domínio específico em comunidades Q&A. Os resultados obtidos poderão ser utilizados como suporte em futuras recomendações de respostas para questões não respondidas nestas comunidades Q&A, sem a intervenção de especialistas.

Para atingir o objetivo geral, os seguintes *objetivos específicos* foram traçados: (1) realizar uma pesquisa bibliográfica, a fim de fazer um levantamento das principais abordagens no contexto do problema; (2) desenvolver uma arquitetura genérica de modo que possa ser aplicada a diversos contexto a cerca do tema; (3) avaliar o uso de categorias extraídas a partir da análise textual das perguntas e respostas para representação semântica; (4) comparar três ferramentas de extração de categorias; (5) avaliar o uso de uma métrica de similaridade semântica para estabelecer relações semânticas e (6) avaliar a relação semântica entre perguntas e respostas de comunidades Q&A.

1.5 Método

Com o objetivo de avaliar a proposta do trabalho, foram realizados experimentos (análises quantitativas) dos resultados obtidos. Contudo, para que a realização dos experimentos fosse possível, primeiramente foram coletados dados de três comunidades online de perguntas e respostas existentes na Web.

As três comunidades cujos dados foram coletados, cada uma abrange um tema específico, no qual perguntas, respostas e usuários estão sujeitos a um processo de premiação de reputação. Oferecem um mecanismo que permitem aos usuários avaliar as respostas por meio de uma pontuação e também categorizar as discussões nela.

Depois de coletados os dados e extraídas as informações pertinentes para o estudo, os textos das perguntas e respostas das comunidades foram submetidos a três ferramentas de extração de categorias com o objetivo de extrair termos das perguntas e respostas. Uma vez feito, foram feitas algumas análises com as categorias extraídas dos textos.

Posteriormente, os termos extraídos das perguntas e respostas foram utilizados para realizar o cálculo de similaridade semântica, obtendo assim o score semântico entre a pergunta e respostas.

Por fim, para avaliar o método proposto foi criado um cenário que consiste em comparar os resultados gerados pela abordagem proposta com o resultado advindo da votação dos usuários da comunidade ao avaliar a resposta, utilizando as perguntas e apenas suas respostas.

1.6 Organização da Dissertação

O presente trabalho está organizado da seguinte forma: o Capítulo 2 apresenta uma breve conceitualização de comunidades online, comunidades Q&A, web semântica e técnicas de análise semântica nestes contextos; no Capítulo 3, são apresentados alguns dos trabalhos relacionados; o Capítulo 4 trata da solução proposta; no Capítulo 5, apresenta-se alguns experimentos e análise de resultados preliminares, enquanto no Capítulo 6 é finalizado com as conclusões e perspectivas futuras.

2. Conceitos Fundamentais

Neste capítulo, são apresentados os fundamentos teóricos relacionados ao tema da pesquisa. Na Seção 2.1, está relatado um breve histórico sobre a evolução das comunidades online, desde os primeiros quadros de aviso, até a era da Web social. Na Seção 2.2, é descrito o funcionamento dos Sites de Perguntas e Respostas e uma das estratégias utilizadas por essas comunidades para facilitar a colaboração. Na Seção 2.3, discorre sobre Web Semântica.

2.1 Comunidades online

O conceito de comunidades online surgiu na década 90. Howard Rheingold [37] foi o primeiro autor a difundir o conceito de comunidade online. Ele define a comunidade online como uma agregação cultural formada pelo encontro sistemático de um grupo de pessoas no ciberespaço. Este tipo de comunidade é caracterizada pela co-atuação de seus participantes, os quais compartilham valores, interesses, metas e posturas de apoio mútuo, por meio de interações no universo online.

Com o passar dos anos, com os avanços tecnológicos e com a globalização, surgem novas formas de comunicação e transmissão cultural, como por exemplo: o computador, os satélites, a Internet e o e-mail, etc. Com a popularização do computador pessoal e o advento da Internet e suas ferramentas de comunicação, facilitou-se as interações online entre as pessoas, que muitas vezes encontram-se distantes umas das outras. Nos anos 80, o computador pessoal foi substituído pelo computador coletivo, interligado por meio de um sistema de rede, assim se estabelece a era da comunicação digital [30]. Estas novas tecnologias facilitaram a constituição de grupos de indivíduos ligados por vínculos não formalizados, os quais tinham características comuns, formando as comunidades online. Estes grupos de pessoas interconectadas, utilizando-se do computador e da Internet como ferramentas de comunicação e interação, constituíram as primeiras comunidades online

[30].

Segundo Vasilescu [51], comunidade online é conceituada como um espaço comum onde as pessoas se relacionam por meio de uma conexão estabelecida no ambiente web, para trocar ideias, debater temas e solicitar ajuda de outras pessoas. Dessa forma, estabelecem uma reunião de pessoas com objetivos em comum ou de natureza diversa com a finalidade de proporcionar discussões acerca desses assuntos.

De acordo com Murray Turrof, o uso do computador poderia proporcionar a um conjunto de pessoas diversos tipos de experiências, privilegiando a propagação da inteligência coletiva. Idealizador do sistema de intercâmbio de informação eletrônica, ele acreditava que um grupo bem preparado conseguiria resultados melhores de inteligência do que os demais membros [18] apud [36]. Consequentemente, por meio de interconexão surgiria uma nova forma de atividade coletiva, focada na multiplicação e permuta de informações, saber e interesses [19] apud [13].

Uma comunidade online é definida como um grupo de pessoas que interagem em um ambiente virtual. Essas pessoas possuem um objetivo, utilizam o suporte de alguma tecnologia e são regidos por normas e regras [34]. Apesar de estarem dispersas geograficamente, essas pessoas conseguem se conectar por meio do uso da tecnologia, essas relações são potencializadas quando há uma busca comum de informações. Observa-se uma grande troca de experiências e conhecimentos, consequentemente, mas não de forma obrigatória, o uso dessas comunidades com foco acadêmico.

De toda forma, encontramos diversos usos para essas comunidades, pois nelas participam diferentes indivíduos, então podemos pensar que atualmente existe uma vasta quantidade de participantes e ao mesmo tempo de espécies de comunidades. Por exemplo, cita-se comunidades para um propósito geral (por exemplo, os usuários dos grupos do Facebook[1]), comunidades para profissionais do mercado de trabalho (por exemplo, os usuários do LinkedIn[2]), outras são destinadas ao compartilhamento de projetos de software (como o Github[3] e o Bitbucket[4]), algumas são usadas para compartilhar imagens ou fotografias (Instagram[5] e Flickr[6]). Outras comunidades têm como objetivo o compartilhamento de conhecimentos através da construção de conteúdos (como a Wikipédia[7]) ou

[1] https://www.facebook.com/
[2] https://www.linkedin.com/
[3] https://github.com/
[4] https://bitbucket.org/
[5] https://www.instagram.com/
[6] https://www.flickr.com/
[7] https://www.wikipedia.org/

através de perguntas e respostas (Yahoo! Answers[8], Quora[9] e Stack Exchange[10]).

O surgimento das comunidades online não está ligado a um intuito econômico, mas na interação dos membros delas, o importante é que podemos verificar a criação de relações mútuas que são estabelecidas e formadas [52]. Muitas pessoas preferem recorrer a esse modelo de busca ao modelo tradicional de site de conteúdo.

O foco deste trabalho é nas comunidades online, principalmente as comunidades de perguntas e respostas e a importância delas na disseminação e compartilhamento de conhecimentos.

Um tipo particular de comunidade online são as comunidades de Perguntas e Respostas (*Question and Answer Communities - Q&A* ou *Community Question Answering*), que são formadas por indivíduos com um interesse comum, porém com diferentes níveis de experiência. Nas comunidades Q&A, quando um indivíduo se depara com um problema particular, ao invés de procurar sozinho por uma solução na Web, utilizando ferramentas de busca tradicionais, ele opta por tornar seu problema público, em alguma comunidade online, fazendo assim uma pergunta [20]. As perguntas são publicadas por participantes que apresentam um problema ou que possam ter interesses comuns, provocando uma discussão, enquanto outros participantes tentam ajudar, baseados em princípios como altruísmo, ganho de reputação, reciprocidade ou mesmo benefícios de aprendizagem direta [23, 24]. Aqueles que têm uma posição de prestígio na comunidade ou uma alta reputação são considerados especialistas e geralmente podem fornecer as respostas mais efetivas e favoritas.

2.2 Sites de Perguntas e Respostas

No que diz respeito aos sites de perguntas e respostas, são comunidades online que possibilitam as pessoas terem acesso a informações publicando perguntas e compartilhando conhecimento pelo fato de estarem respondendo perguntas elaboradas por outros usuários do serviço. Muitas vezes as perguntas são rapidamente respondidas porque há um grande número de membros nessas comunidades. Esses sites facilitam aos usuários postarem perguntas acerca de diferentes temas, por outro lado, encontram-se, também, sites aplicados a debates de assuntos específicos [47]..

O sistema de buscas tradicional nem sempre é a melhor opção para responder pergun-

[8] https://answers.yahoo.com/
[9] https://www.quora.com/
[10] https://stackexchange.com/

tas mais subjetivas, específicas e, por que não, pessoais. Para isso, existem outros sites em que outras pessoas respondem e compartilham ideias e experiências. Por exemplo, podemos citar o *Quora*, é o mais utilizado no universo Web. Fundado em 2009, o site traz perguntas que os próprios usuários fazem e respondem, além de sugerir edições em perguntas e respostas já compartilhadas.

O funcionamento de um site de perguntas e respostas se baseia em uma pergunta que abre uma trilha (*thread*), organizando os debates em função da pergunta postada, ou seja, um usuário posta uma pergunta, em seguida, em um curto espaço de tempo outros usuários postam respostas ou comentários relativos à pergunta. Além disso, categorias podem ser atribuída as *threads* indicando o tema do debate (por exemplo: categoria Python, categoria banco de dados, categoria redes etc.), o que facilita a identificação do tema da questão e sua indexação para filtros na busca de questões relacionadas. Esses site também oferecem a possibilidade de avaliação da questão, da resposta e também do usuário, que podem ser avaliados por outros usuários baseado em suas perguntas ou respostas postadas.

Para se beneficiar dos recursos oferecidos pelo Quora por exemplo, basta que o usuário se cadastre. O site tem um feed com as melhores respostas, além de permitir que você navegue por categorias como Séries, Tecnologia, Música e Esportes, bem como é possível votar nas melhores respostas.

Em relação ao Quora, encontra-se perguntas bem abrangentes, desde temas genéricos, como perguntas sobre séries, a questionamentos específicos ligados à informática, a título de exemplo, como se utilizar o torrents. Há também perguntas de cunho mais pessoal, citando assim qual música a pessoa colocaria no dia do seu casamento e assim uma sucessão de perguntas direcionadas ao esclarecimento de dúvidas ou apenas para adquirir conhecimento.

Dentro desse contexto podemos encontrar determinadas comunidades que oferecem certos incentivos para aqueles com maiores índices de acesso ao site ou por terem respondido mais perguntas, mas normalmente o que acontece é uma contribuição voluntária entre os membros das comunidades [39]. Contudo, o foco principal destas comunidades Q&A é a divulgação, de forma interativa, das respostas mais adequadas às perguntas postadas recentemente, no menor intervalo de tempo possível [47].

Dessa forma, estas comunidades Q&A apresentam-se como uma opção ao modo padrão de busca na internet e possibilitam que pessoas recebam informações de outras pessoas diretamente [57]. Geralmente os indivíduos utilizam estas comunidades: (i) quando necessitam de um conhecimento específico e consequentemente de uma resposta rápida

vinda de alguém que já tenha vivenciado algo semelhante. (ii) Por não encontrar nenhum outro lugar na web que ofereça a resposta. (iii) ou para interagir com outros usuários [2].

2.3 Web Semântica

O trabalho aqui apresentado visa explorar a semântica do conteúdo de comunidades Q&A, partindo de tecnologias da Web Semântica. O termo Web Semântica refere-se a rede de informações que dá significado aos conteúdos na Internet. Para Berners-Lee, Hendler e Lassila [7], a Web Semântica pode ser vista como "uma extensão da Web atual, onde a informação tem um significado bem definido, permitindo que computadores e pessoas trabalhem em cooperação". Ou seja, além do conteúdo em si, que pode ser interpretado pelo ser humano, existem mais informações que permitem que os conteúdos sejam reconhecidos e manipulados de maneira mais significativa pelos próprios sistemas. Tais informações permitem que os computadores executem tarefas mais sofisticadas e retornem resultados mais precisos e contextualizados para o usuário [38].

Tim Berners Lee, no ano 2000, propôs um modelo de camadas para a Web do futuro, conforme exemplificado na Figura 2.1. A ideia dessa estrutura é reutilizar uma parte da estrutura já existente na Web, criando gradativamente novas camadas sobre as já existentes. Descrevendo os recursos, tecnologias e as linguagens para a Web Semântica o modelo de camadas de Tim Berners Lee apresenta um modelo para a Web ser desenvolvida [6].

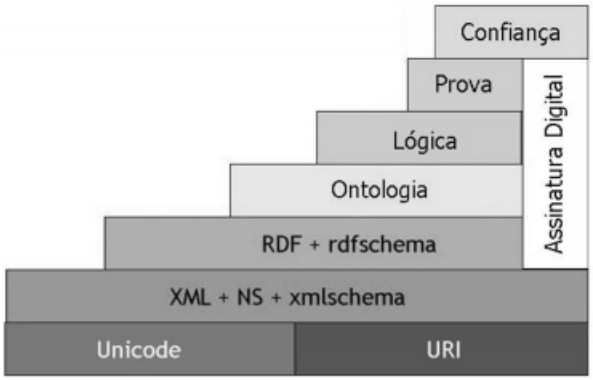

Figura 2.1: Modelo de camadas proposto por Tim Berners Lee para a Web Semântica

Como parte do desenvolvimento da Web Semântica, surgiu o conceito de Dados Conectados (Linked Data). Dados Conectados refere-se ao uso da Web para conectar dados relacionados que não estavam previamente relacionados, apresentando um conjunto de boas práticas para publicar e conectar conjuntos de dados de forma estruturada na Web, com o objetivo de criar uma "Web de Dados" [8]. Essas práticas se baseam em tecnologias Web, como HTTP (Hypertext Transfer Protocol) e URI (Uniform Resource Identifier),

com o objetivo de possibilitar a leitura dos dados conectados, de forma automática, por agentes de software.

Uma das principais características da Web Semântica está na capacidade de vincular recursos da Web a categorias (ou seja, em descrever semanticamente os recursos disponíveis na Web). Categorias associadas a um conjunto de documentos são importantes para ajudar a identificar tópicos em publicações, melhorando assim a sua categorização e indexação, auxiliando em tarefas como busca, recuperação de informações e recomendações. Torna-se cada vez mais comum a tarefa de identificar os tópicos associados a documentos, seja de forma manual ou automática. Diversas abordagens são utilizadas para recuperar de forma automática categorias de um documento, as mais recentes tendem a utilizar técnicas de modo a dar mais semântica, utilizando categorias extraídas da Wikipédia, ao invés de um sistema de classificação tradicional criado por especialistas de domínio.

O conteúdo criado na Wikipédia é extraído e estruturado em um banco de dados, DBpedia[11], de modo a tornar esta informação disponível na Web. *O DBpedia Spotlight*[12] é uma ferramenta disponível para automaticamente anotar menções a recursos da DBpedia em um texto, permitindo assim conectar fontes de informação não estruturadas. O *DBpedia Spotlight* utiliza como base as categorias da Wikipédia para executar a extração de informações a partir de texto, entre elas o reconhecimento de entidades mencionadas (*Named Entity Recognition - NER*) e a resolução de nomes.

Outra ferramenta que possibilita o processamento de texto e a extração de entidades é o *Open Calais*[13], que utiliza ontologias, algoritmos de processamento de linguagem natural (NLP) e aprendizagem automática, treinado por especialistas. O *Open Calais* possibilita identificar pessoas, lugares, companhias, fatos e eventos, a partir de um texto não estruturado. Para isto, conta com as autoridades curadas mantidas por membros da equipe de dados da Thomson Reuters e também aproveita o gerenciamento de identidade fornecido pelos especialistas da Thomson Reuters.

Complementarmente, o *Tag The Web*[14] é uma ferramenta que utiliza a estrutura taxonômica da Wikipédia, baseado na geração de uma impressão digital (*fingerprint*) através da relação semântica entre os nós do grafo de categorias da Wikipédia [26]. O objetivo é associar um conjunto de tópicos a partir das categorias da Wikipédia para um dado recurso web. O uso de Web Semântica, mais especificamente de Dados Conectados, possibilita conectar recursos Web de maneira que os dados estejam estruturados. A conexão

[11] https://wiki.dbpedia.org/
[12] https://www.dbpedia-spotlight.org/
[13] https://www.opencalais.com/
[14] http://www.tagtheweb.com.br/

dos dados pode ser realizada através de categorias extraídas de recursos Web, utilizando uma métrica de conectividade em entidades dadas em um conjunto de dados de referência.

O *Explicit Semantic Analysis* (ESA) [16] é um método de análise semântica explícita, este método compara a similaridade semântica entre dois textos curtos. Ele utiliza técnicas de aprendizagem de máquina para construir um interpretador semântico que mapeia fragmentos de texto em linguagem natural em uma sequência ponderada de conceitos extraídos da Wikipédia ordenados pela sua relevância. Desta forma os textos não são comparados diretamente, mas como uma coleção representada por vetores ponderados de conceitos. A coleção indexada (frequentemente artigos da Wikipédia) pode ser pré-processada e armazenada, de modo que usar o ESA não constitui uma atividade muito custosa em comparação com outros métodos que demandam um alto processamento [11].

O *Semantic Connectivity Scores* (SCS) [31] é uma medida baseada em co-ocorrência e Web Semântica para descobrir relacionamentos entre entidades. O SCS apresenta um score de conectividade semântica baseado no índice de Katz [22], tendo como principal característica o uso de propriedades transversais, que descrevem relações não hierárquicas entre entidades para indicar uma forma de conectividade independente de sua similaridade. Um modelo de grafo não direcionado é utilizado de modo a reduzir a complexidade computacional. O score semântico é baseado no número de caminhos e distâncias (comprimento de um caminho) entre entidades, o SCS considera apenas caminhos com um comprimento máximo (t = 4). De acordo com Nunes et al. [31], apresenta um desempenho superior do que o ESA para mensurar o parentesco entre entidades por meio da estrutura de grafos de categorias da Wikipédia.

3. Trabalhos Relacionados

Este capítulo apresenta os principais trabalhos relacionadas ao tema da pesquisa desta dissertação. A presente revisão da literatura levou em consideração o enquadramento dos trabalhos em quatro grandes temáticas: o compartilhamento de perguntas online, a identificação de questões semelhantes, a resposta de questões usando respostas anteriores e o contexto de sistemas de perguntas e respostas. O objetivo desta revisão concentra-se na análise das principais técnicas de extração das categorias e o cálculo da similaridade semântica entre as categorias. Ao final, conclui-se com uma apreciação das principais contribuições da pesquisa em relação aos trabalhos levantados, através de um quadro comparativa com as pesquisas correlatas.

3.1 O Compartilhamento de Perguntas Online

Atualmente, o compartilhamento de problemas em comunidades online é uma prática recorrente. Após a publicação de uma pergunta em uma comunidade, é esperado uma resposta rapidamente. Entretanto no paradigma da Consulta Social, não há garantias de quanto tempo levará para a pergunta ser respondida, ou se a pergunta será respondida, em determinado momento.

No Stack Exchange, por exemplo, as perguntas normalmente são respondidas em um período curto de tempo. Porém, acompanhando o crescimento do site, já é perceptível que o percentual de perguntas que são ignoradas ou não resolvidas continua crescendo [47]. Alguns autores apontam alguns problemas nas Comunidades de Perguntas e Respostas, como: (i) a maior parte das perguntas não são respondidas em até dois dias [56]; (ii) aproximadamente 23% das perguntas não recebem uma resposta ou recebem respostas que não são classificadas uma boa resposta [42]; (iii) quando perguntas são feitas de forma repetida, os especialistas ficam menos inclinados a responder e os novos usuários são desencorajados a participar da comunidade quando as suas perguntas não são respondidas

[42].

3.2 Identificar Questões Semelhantes

Uma estratégia para responder perguntas não respondidas é buscar por perguntas semelhantes. Alguns estudos apresentaram soluções para identificar perguntas semelhantes em Comunidades de Perguntas e Respostas ou posts em fóruns de discussão.

Em [54] e [53] é proposto uma solução para identificar perguntas duplicadas em comunidades de perguntas e respostas sobre programação utilizando *deep learning* e técnicas de recuperação de informação. Apresenta um estudo sobre Comunidades de Programação Q&A (PCQA), onde posts em comunidade de programação apresentam trechos de código, o que dificulta uma linguística diferente da linguagem natural. A solução apresenta uma metodologia para detecção de perguntas duplicadas utilizando 3 abordagens: (i) vetor de similaridade, ao qual representa a pergunta em um vetor de alta dimensionalidade realizando a medida de similaridade entre o par de questões baseado no cosseno do vetor utilizando vetor *tf-idf (term frequency–inverse document frequency*[1]); (ii) similaridade tópica, calculada usando um modelo de tópico para extrair temas de textos curtos e calculando a similaridade de distribuições tópicas entre o par utilizando uma implementação de *Latent Dirichlet Allocation* (LDA); (iii) score de associação, que utiliza de frequência de co-ocorrência e alinhamento de palavras utilizando treinamento de um rede perceptron para o cálculo do score. Por fim realiza a detecção de perguntas duplicadas utilizando alguns modelos para classificação binária: decision tree, K-nearest neighbours (K-NN), linear SVM, logistic regression, random forest e naive Bayes. Experimentos realizados com o Stack Overflow, uma comunidade de programação, demonstraram que o método apresenta um bom desempenho, em alguns casos, mais de 30% de melhoria comparada com benchmarks do estado da arte.

Em [10] e [11], é apresentado um modelo de recuperação aprimorada de categorias que utilizam categorias para recuperação de perguntas semelhantes em Comunidades de Perguntas e Respostas. Quanto mais uma categoria está relacionada a uma consulta, mais provável é que a categoria contenha questões relevantes a consulta. O modelo classifica uma questão de um banco de questões, baseada em uma interpolação de dois escores de relevância: um é um escore de relevância global entre a consulta e categoria, e o outro é um escore de relevância local entre a consulta e a questão. A solução utiliza 5 modelos de representação para as questões: VSM (Vector Space Model), Okapi (Okapi BM25 Mo-

[1]Métrica estocástica para calcular a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico.

del), LM (Language Model), TR (Translation Model) e TRLM (Translation-Based Language Model). Foram realizados testes em algumas comunidades do Yahoo! Answers, os resultados demonstram que uma combinação de dois modelo, o VSM+TRLM, foi a que obteve a melhor performance.

Já em [56], é proposto o uso dos conceitos da Wikipédia e processamento de linguagem natural para identificar perguntas semelhantes em uma Comunidade de Perguntas e Respostas. Os métodos tradicionais medem a semelhança com base na representação de *bag-of-words* (BOWs) em uma representação em um vetor *tf-idf*. Entretanto a representação BOWs não captura dependências entre palavras relacionadas, nem manipula sinônimos ou palavras polissêmicas. Os autores propõem uma maneira de construir um *thesaurus* conceitual baseado nas relações semânticas extraídas do conhecimento da Wikipédia. É realizado um mapeamento das questões para conceitos da Wikipédia utilizando o algoritmo *Forward Maximum Matching* para encontrar conceitos candidatos. O modelo proposto mede a similaridade entre as questões com hiperônimos, sinônimos e conceitos associativos derivados da Wikipédia. Experimentos, conduzidos em comunidades do Yahoo! Answers mostraram que, com a ajuda do *thesaurus* da Wikipédia, o desempenho da recuperação de perguntas é superior em comparação com os métodos tradicionais.

Em [9] é proposto identificar perguntas semanticamente equivalentes em fóruns utilizando uma rede neural convolucional (CNN – *Convolutional Neural Network*). A CNN proposta, primeiro transforma as palavras em vetores de palavras (*word embeddings*), usando uma grande coleção de dados não rotulados, e então aplicados a CNN para construir representações de vetores distribuídos para os pares de questões. Por fim, pontua as perguntas usando uma métrica de similaridade, que é realizada na última camada da CNN utilizando similaridade do cosseno. O método proposto foi avaliado utilizando a comunidade Ask Ubuntu, uma comunidade Q&A do Stack Exchange para usuários de Ubuntu e desenvolvedores e comparado o resultado da detecção de perguntas semelhantes do CNN com outro método de machine learning, Máquina de Vetores de Suporte (*SVM – Support Vector Machine*). Os resultados mostram que o CNN supera a SVM por uma margem significativa.

3.3 Responder Questões com Respostas Anteriores

Outros estudos, além de identificar perguntas semelhantes, apresentam respostas para perguntas duplicadas, baseadas nas respostas anteriores em Comunidades de Perguntas e Respostas.

Em [41], é investigada a validade de utilizar respostas anteriores para responder novas questões em uma Comunidade de Perguntas e Respostas. Para isso, são utilizadas técnicas estatísticas, de processamento de linguagem natural e de similaridade entre textos. É apresentada uma abordagem em dois estágios, primeiro classificando as questões candidatas e depois extraindo as respostas plausíveis das melhores questões. Na classificação das questões candidatas, o espaço de busca é limitado para perguntas passadas com respostas que foram marcadas como melhor resposta pelos próprios usuários, e foram classificadas com pelo menos três estrelas. Após é realizado um ranqueamento utilizando a medida de similaridade de cosseno em um espaço vetorial com pesos *tf-idf*. Na segunda etapa, as respostas dos melhores candidatos são extraídas para responder a questão, constituindo uma representação em tríade contendo: a nova questão, a melhor questão do passado e a melhor resposta da questão. Foram extraídos 95 características (por exemplo, comprimento do texto, número de pontos de interrogação, contagem de palavras finais e etc.) usando várias considerações de predição de linguagem léxica, linguagem natural e desempenho de consulta, para representar esta tríade. Foram realizados experimentos com o Yahoo! Answers, utilizando 4 modelos como classificadores: *Random Forest, Logistic Regression, SVM* e *Naive Bayes*. Os resultados mostraram que o sistema respondeu com precisão entre 65,9% e 86,4% das perguntas, dependendo da categoria.

Em [40], é proposta uma arquitetura, utilizando matriz de similaridade contendo informação lexical e sequencial, para responder a novas questões feitas em Comunidades de Perguntas e Respostas baseadas em perguntas e respostas anteriores. Nesta arquitetura, perguntas e respostas foram representadas como uma lista ordenada de vetores de palavras usando o modelo de linguagem neural. A interação complexa entre questões e respostas foi modelada em uma matriz bidimensional (*S-matrix*). Por último uma CNN foi treinada para fornecer probabilidade de resposta adequada. Experimentos foram realizados utilizando o maior site chinês de comunidades Q&A, o Baidu Zhidao. A principal contribuição deste trabalho demonstrou que a abordagem por meio de técnicas de *deep learning* é uma alternativa aos métodos populares para este problema.

Em [43], os autores apresentam um sistema para responder questões não respondidas em Comunidades de Perguntas e Respostas, utilizando palavras-chave semânticas em combinação com técnicas tradicionais de busca textual. Além disso também recomenda usuários especialistas que possam responder às perguntas. Utiliza tecnologias de Web Semântica e Dados Conectados, integrando datasets de duas bases de dados, estruturando através de uma ontologia para vinculá-los à nuvem de dados conectados (*Linked Data Cloud*). Utiliza sistemas de extração de entidades (*Wikipedia-Miner* e *Open Calais*) para nomear as entidades e fazer a anotação dos dados com palavras-chave. Posteriormente

realizar consultas (*SPARQL*) sobre respostas para perguntas já respondidas. Deste modo, não consideram um cálculo de similaridade semântica, mas apenas os mesmos termos encontrados nas consultas. Experimentos foram realizados utilizando Comunidades de Perguntas e Respostas do StackOverflow e do Reddit, e os resultados mostram que o sistema proposto é bastante eficiente para encontrar as respostas certas.

3.4 Contexto de Sistemas de Perguntas e Respostas

Alguns estudos estão no contexto de Sistemas de Perguntas e Respostas (*Question Answering System*) que utilizam uma equipe de especialistas e também pessoas voluntárias, assim como textos relacionados para responder às questões.

Em [27], os autores apresentam um estudo sobre adoção de um modelo semântico para *Question Answering System*, que utiliza técnicas de Processamento de Linguagem Natural , Recuperação da Informação e aprendizagem de máquina para analisar, recuperar e classificar conteúdo, auxiliando o processo de responder perguntas de forma automática. Utilizam várias combinações como *Latent Semantic Analysis* (LSA), *Non-negative Matrix Factorization* (NNMF), *Random Index* e *Explicit Semantic Analysis* (ESA) para calcular a similaridade semântica dos termos. O método foi avaliado utilizando coleções de documentos de texto não estruturados, de bases do ResPubliQA 2010, que compreendem documentos sobre a legislação da União Europeia e as transcrições do Parlamento Europeu. Os resultados mostram uma melhoria na comparação com outros sistemas do estado da arte demonstrando como é promissor o uso de modelos semânticos para o campo de Q&A.

Em [3] também é proposto um *Question Answering System* para responder perguntas automaticamente usando uma abordagem semântica, através de um sistema híbrido que faz uso de diversas abordagens incluindo o método de Análise Semântica Explícita (ESA). Realiza o processo através de 4 passos: (i) Classificação da Questão, onde é identificado o tipo de resposta necessária para a pergunta; (ii) Processamento da Questão, neste passo é compreendido a semântica da questão, obtendo um conjunto de conceitos extraídos a partir da Wikipédia utilizando o método ESA para calcular a similaridade; (iii) Documentos Relevantes, com base nos conceitos definidos da Wikipédia e métodos para obter a frequência dos documentos mais relevantes; (iv) Extração da Resposta, neste último passo dependendo do tipo de pergunta, a informação relevante é extraída para responder a questão através de um *ranking* para a relevância e frequência das respostas. Experimentos foram realizados utilizando a base de dados de Q&A da Conferência Text REtrieval

(TREC-QA 2004). Os resultados mostram que com este método foi possível identificar respostas corretas para perguntas, com destaque para a utilização do método ESA com vantagem no desempenho para responder questões comparado a outros métodos.

Em [32] é apresentada uma proposta, baseada em inteligência artificial, para identificar perguntas duplicadas em *Question Answering System* no momento que o usuário tenta incluir uma nova pergunta e automaticamente indicar respostas com base em respostas anteriores. Apresenta um modelo de treinamento a partir das questões, para realizar um score de questões duplicadas e o conceito de "clusters duplicados", que fornece uma estrutura semi-automatizada para identificação de conteúdo duplicado. No modelo de score de questões duplicadas, utiliza similaridade de cosseno com pesos *tf-idf* combinado a um modelo estatístico computado com *Latent Dirichlet Allocation* (LDA). Os clusters duplicados são representados em uma grafo não direcionado, ao qual cada par duplicado e questão duplicada identificados com o modelo é constituído a aresta do grafo e o vértice do grafo, respectivamente. Foram realizados experimentos com o *AnswerXchange*, uma popular social Question Answering System que apoias usuários que trabalham com impostos federais e estaduais dos EUA. Para o score das questões duplicadas foram utilizados classificadores binário linear (*logistic regression*) e não linear (*random forest*), comparados a similaridade de cosseno, os resultados mostram que estes dois modelos atingem um desempenho que é consistente com os objetivos deste estudo exploratório.

3.5 Contribuições da Pesquisa

Diversas abordagens podem ser utilizadas para reduzir a quantidade de perguntas sem respostas, alguns estudos focam em identificar perguntas semelhantes, outros utilizam de respostas anteriores para responder novas questões em Comunidades de Perguntas e Respostas, e por último nos Sistemas de Perguntas e Respostas usam-se especialistas em informação para responder às questões.

Por outro lado, Comunidades de Perguntas e Respostas diferem dos Sistemas de Perguntas e Respostas automática, já que as fontes de conhecimento advém de seus próprios usuários que agregam valor enquanto participantes das interações sociais entre eles. Como conclui [17], a qualidade do conhecimento que os usuários podem encontrar em Comunidades de Perguntas e Respostas pode mesmo exceder a qualidade do conhecimento recuperado a partir de especialistas em informação. Isto decorre da formação de pequenos grupos de especialistas dos domínios destas comunidades.

Mesmo assim, a quantidade de questões sem resposta é crescente nestas comunida-

des, assim como a preocupação com a qualidade das respostas [12, 32]. A abordagem do presente trabalho utiliza categorias extraídas de perguntas e respostas e também a partir de base de respostas de perguntas já respondidas, assim como no trabalho em [Singh 2014]. A grande maioria dos trabalhos anteriores utilizam, principalmente, técnicas de NLP e técnicas estatísticas para extração das características textuais das perguntas e/ou respostas, como no trabalho em [Shtok, 2012]. Considerando que poucos trabalhos utilizam propriamente as categorias extraídas com base no senso comum da *folksonomia* da Wikipédia e que a riqueza do conhecimento comum da wikipédia pode resultar em uma melhora na qualidade semântica das categorias relacionadas, esta dissertação se propõe a investigar este uso.

Em [26] é proposto um método genérico para classificação baseado no senso comum expresso pela estrutura taxonômica da Wikipédia. O trabalho utiliza o Tag The Web e classifica o conteúdo de comunidades Q&A do Stack Exchange, demonstrando que a categorização das comunidades conforme o Tag The Web é adequada. Entretanto, não é apresentada a aplicação do Tag The Web no contexto de identificação de correspondências entre perguntas e respostas nestas comunidades.

Com relação a medir a similaridade entre perguntas e respostas, a grande maioria utiliza técnicas de aprendizagem de máquina e realiza cálculo de similaridade com técnicas complexas, que demandam um treinamento anterior com a base de dados. Abordagens mais comuns são de representação dos dados em vetores multidimensionais combinados a técnicas como *tf-idf* e similaridade de cosseno, como no trabalho em [Zhang, 2017], como métodos padrões para ranquear o resultado das buscas. O presente trabalho utiliza o método de similaridade semântica *Semantic Connectivity Score* (SCS) [31], que é semelhante ao *Explicit Semantic Analysis* (ESA), utilizado em [Aroussi 2016]. Conforme descrito na Seção 2, o SCS apresenta um resultado bem superior ao ESA. Complementarmente utiliza a identificação de *fingerprints* do trabalho em [26] com objetivo de investigar a possibilidade em estabelecer relações semânticas entre perguntas e respostas, a fim de recomendar respostas possíveis para perguntas não respondidas.

3.5.1 Análise Comparativa com as Pesquisas Correlatas

Na Tabela 3.1, é apresentado uma análise resumida dos trabalhos relacionados, incluindo a referência para o trabalho, a fonte de dados utilizado, a técnica utilizada para extração das categorias e como foi realizada a similaridade semântica.

Tabela 3.1: Trabalhos Relacionados

Referência	Fonte de Dados	Extração Categorias	Similaridade Semântica
Zhang (2017) [53, 54]	Stack Exchange	NLP e LDA	vetor tf-idf, K-NN, decision tree, linear SVM, logistic regression, random forest e naive Bayes
Cao (2012) [10, 11]	Yahoo! Answers	NLP	VSM, Okapi BM25 Model, LM (Language Model), TR (Translation Model) e TRLM (Translation-Based Language Model)
Zhou (2013) [55, 56]	Yahoo! Answers	NLP	Hiperônimos, sinônimos e conceitos associativos derivados da Wikipédia.
Bogdanova (2015) [9]	Stack Exchange	NLP e word embeddings	Convolutional Neural Network (CNN)
Shtok (2012) [41]	Yahoo! Answers	NLP e técnicas estatísticas	Random forest, Logistic regression, SVM e Naive Bayes
Shen (2015) [40]	Baidu Zhidao	NLP, informações lexical e sequencial	Convolutional Neural Network (CNN)
Singh (2016) [43]	Stack Exchange e Reddit	Wikipedia-Miner e Open Calais	SPARQL e RDF
Molino (2012) [27]	ResPubliQA 2010	NLP e WordNet	Latent Semantic Analysis, Non-negative Matrix Factorization, Random Index e ESA
Aroussi (2016) [3]	TREC-QA 2004	ESA e Wikipédia	ESA
Podgorny (2018) [32]	AnswerXchange	NLP e LDA	Logistic regression e Random forest
Medeiros (2018) [26]	Stack Exchange	DBpedia Spotlight e Tag The Web	Apenas entre comunidades e as categorias obtidas
Esta dissertação	Stack Exchange	DBpedia Spotlight, Open Calais e Tag The Web	SCS

Os trabalhos aqui analisados tratam de temática similar à da abordagem proposta nesta dissertação. O resumo comparativo na Tabela 3.1 relaciona os trabalhos de acordo com as características consideradas importantes para identificação de relações semânticas entre as perguntas e respostas de comunidades Q&A. Uma observação pertinente é que a maioria dos trabalhos é bastante recente (tem menos de três anos). Conforme pode ser observado, a grande maioria dos trabalhos relacionados consideram a análise em dados reais de comunidades Q&A; apenas um trabalho compara abordagens de categorizações das entidades da Wikipédia com outras categorizações (ex.: Thomson Reuters) e todos consideram mecanismos menos eficazes de identificação de similaridade semântica (como o ESA).

4. Solução Proposta

Este capítulo tem como objetivo apresentar a arquitetura utilizada para a solução do problema de pesquisa e detalhando os passos das etapas realizadas. No decorrer do capítulo, serão apresentadas como foi o realizada a extração dos dados das comunidades, o armazenamento, o processamento dos dados e as ferramentas utilizadas em todo o processo. Em síntese, o objetivo deste capítulo é mostrar como foi conduzido o desenvolvimento da solução proposta para testar a hipótese desta dissertação.

4.1 Arquitetura

O presente trabalho se propõe a identificar relações semânticas entre categorias textuais das perguntas e respostas em comunidades Q&A. Normalmente, muitas perguntas não são respondidas e uma metodologia automática poderia recomendar respostas relacionadas. Mas para isso é preciso identificar relações semânticas adequadas entre domínios relevantes, relacionando perguntas e respostas.

A solução examina as categorias obtidas a partir de perguntas e respostas de comunidades Q&A, realizando, em seguida, um cálculo de uma métrica (*score* semântico) que indica a relação semântica entre as perguntas e respostas utilizando e comparando três métodos de categorização. As figuras 4.1 e 4.2 ilustram graficamente as etapas do processo, respectivamente a arquitetura conceitual e a arquitetura implementada.

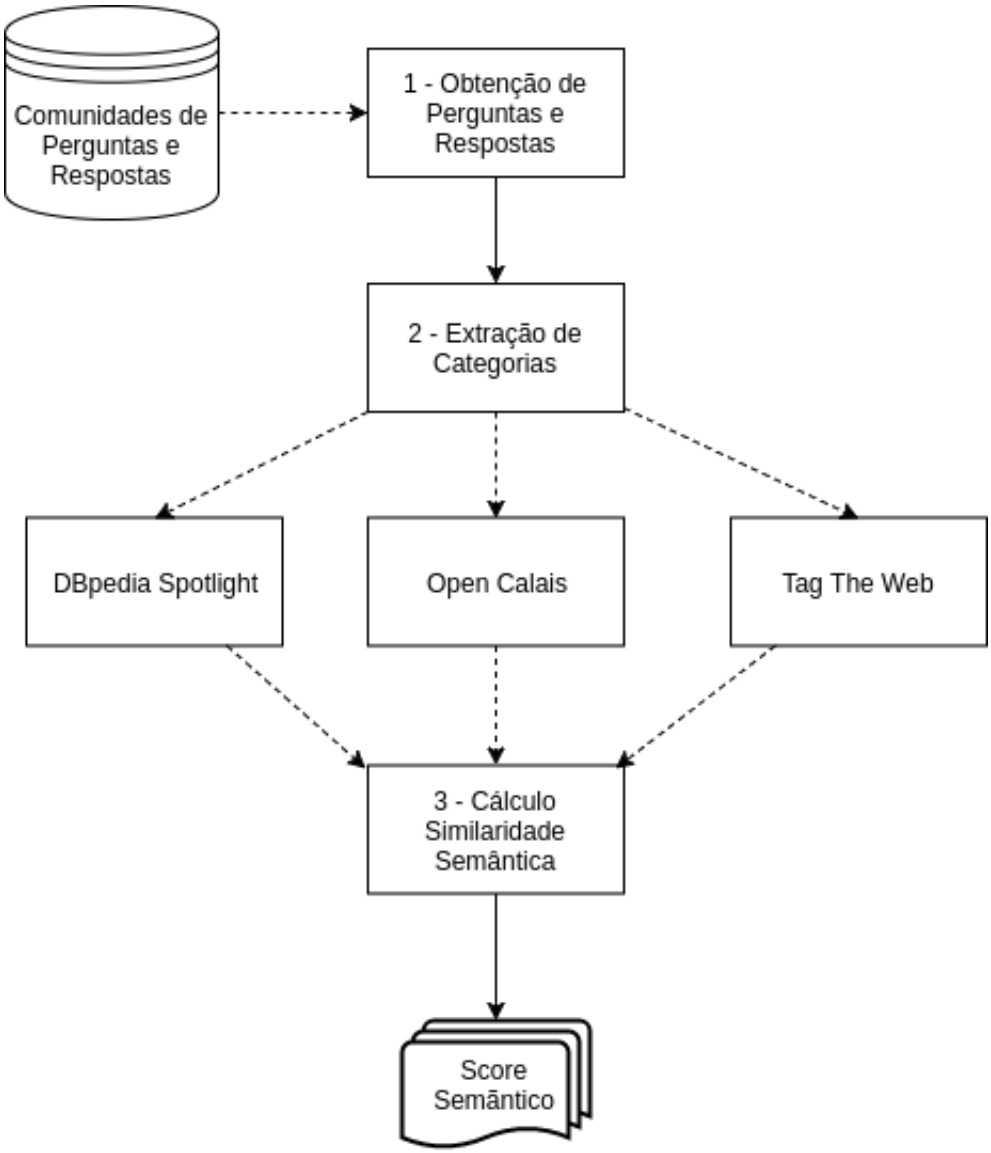

Figura 4.1: Arquitetura conceitual, contendo as etapas do processo da solução proposto

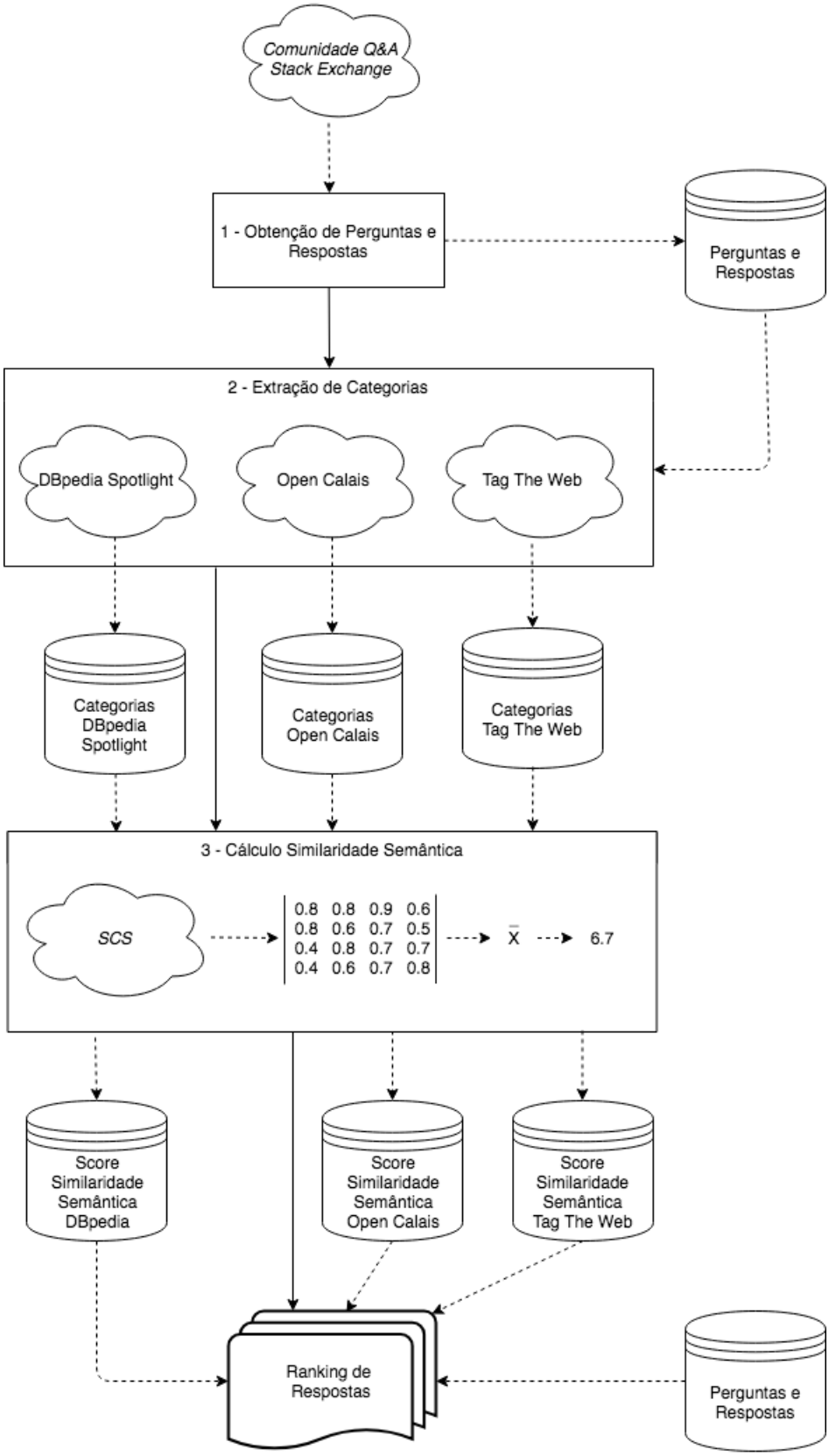

Figura 4.2: Arquitetura implementada, contendo etapas do processo da solução proposto

Na primeira etapa são obtidas as perguntas e respostas de comunidades Q&A, a partir do *Stack Exchange*, porém outras comunidade Q&A, como o *Reddit* e *Quora* por exemplo, ou até mesmo qualquer outro tipo base de dados textual poderia ser utilizado para realizar a análise semântica. Após obtido os dados da comunidade, seja por meio de um *dump*, API, ou qualquer outro meio que possibilite a obtenção dos dados, estes dados são armazenados em uma base de dados relacional, contendo assim as perguntas e respostas (ou qualquer outro texto) para serem analisadas. Na segunda etapa, os textos das perguntas e respostas são submetidos a três ferramentas de extração de categorias, o *DBpedia Spotlight*, *Open Calais* e *Tag The Web*, estas ferramentas disponibilizam uma API, que tem como entrada um texto qualquer e como saída as categorias extraídas a partir do texto. Por fim as categorias extraídas são armazenadas em um banco de dados não relacional separadamente para posterior uso na etapa 3. Nesta etapa ocorre o cálculo de similaridade semântica, em que duas categorias são comparadas obtendo um score semântico. Foi utilizado o *Semantic Connectivity Scores* (*SCS*), que disponibiliza uma API para realizar este cálculo, ele realiza individualmente o cálculo entre duas categorias. Para realizar o cálculo de similaridade entre uma pergunta e uma resposta, todas as categorias extraídas da pergunta e resposta são combinadas e então submetidas para a API realizado o cálculo individualmente para cada par de categoria da pergunta-resposta. Para um melhor entendimento a combinação de todas as categorias da pergunta e resposta pode ser representado como uma matriz, onde as linhas são categorias da pergunta e as colunas são categorias das respostas. Obtém-se todos os scores de cada par de categorias pergunta-resposta e então é realizado uma média aritmética entre todos os scores de cada categorias a fim de obter um valor de score para a pergunta e resposta. Com o score semântico de todas as perguntas e respostas é possível realizar um ranqueamento com base no valor de score, indicando que o maior valor representa uma relação semântica maior entre os dois textos comparados. As próximas seções apresentam um detalhamento do funcionamento de todas as etapas envolvidas do método proposto.

4.2 Obtenção de perguntas respondidas de comunidades Q&A

Nesta etapa, são obtidas perguntas respondidas de forma aleatória de comunidades Q&A. Foi escolhido o Stack Exchange por conter diversas comunidades Q&A de diversos temas em áreas variadas. O Stack Exchange é uma plataforma online de perguntas e respostas, gratuito, construído e gerenciado pelos próprios usuários, conta atualmente com uma rede de 133 comunidades Q&A, no ano de 2015, superou a marca de 5 milhões de usuários registrados e 3,7 milhões de perguntas respondidas, cada comunidade abran-

gendo um tema específico, no qual perguntas, respostas e usuários estão sujeitos a um processo de premiação de reputação.

Na Figura 4.3, é apresentado um exemplo de uma pergunta publicada no Stack Exchange.

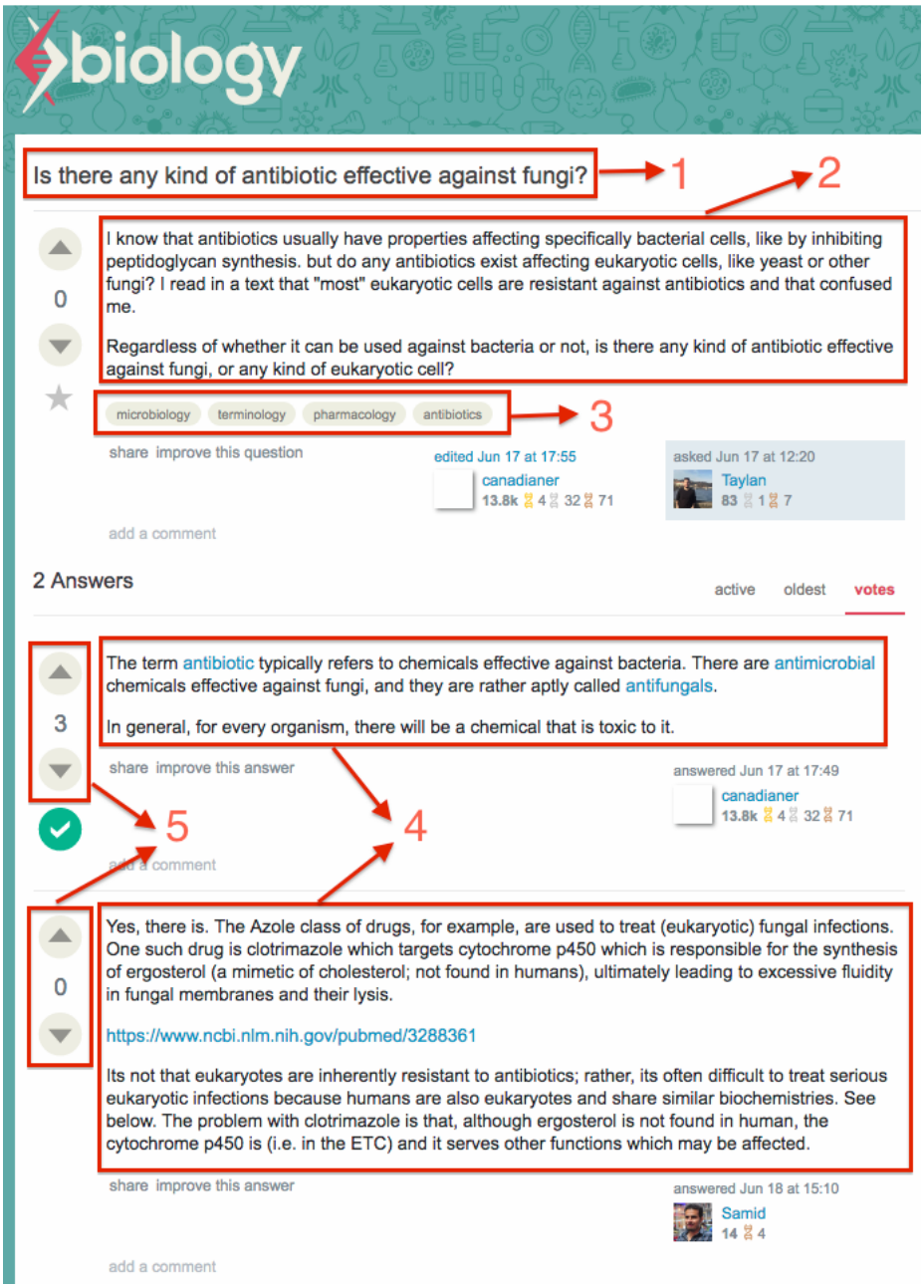

Figura 4.3: Exemplo de questão explorada no Stack Exchange

Para facilitar sua descrição, a Figura 4.3 foi dividida em áreas demarcadas com números, que destacam, respectivamente: (1) o título da pergunta; (2) a descrição da pergunta; (3) as tags atribuídas à pergunta, realizada pelos próprios usuários; (4) as respostas para a pergunta; (5) a pontuação atribuída para as respostas dada pelos usuários, desta maneira é possível identificar a melhor resposta para a pergunta, segundo a votação dos usuários.

O Stack Exchange oferece acesso aos dados de suas comunidades por meio de um *dump*[1] regular de todos os seus dados públicos. O *dump* possui dados de questões, respostas, comentários, informações de usuários (os dados públicos apenas), badges e votação. Os dados são disponibilizados para download por comunidade em formato XML, para cada comunidade são 8 arquivos XML:

- Badges.xml
- Comments.xml
- PostHistory.xml
- PostLinks.xml
- Posts.xml
- Tags.xml
- Users.xml
- Votes.xml

O Stack Exchange também possui uma API[2], no qual é possível realizar demais consultas, como por exemplo lista de tags, número total questões por tags, tags sinônimas e tags relacionadas. A API retorna um arquivo no formato JSON que pode ser interpretado e armazenado em banco de dados.

Neste trabalho foi utilizado o *dump* disponibilizado pelo Stack Exchange em formato XML e por meio de um *script* de importação[3] foi armazenado em uma base relacional. Este *script* de importação cria as seguintes tabelas:

- allposttags
- answers
- badges
- closeasofftopicreasontypes
- closereasontypes
- comments

[1] https://archive.org/details/stackexchange
[2] https://api.stackexchange.com/
[3] https://github.com/Networks-Learning/stackexchange-dump-to-postgres

- flagtypes
- posthistory
- posthistorytypes
- postlinks
- postlinktypes
- posts
- posttags
- posttypes
- questionanswer
- questions
- reviewtaskresulttype
- reviewtasktypes
- tags
- users
- usertagqa
- votes
- votetypes

Foram modeladas as tabelas: "Posts"e "QuestionAnswer"para aplicação consumir os dados, conforme a Figura 4.4, descrita a seguir.

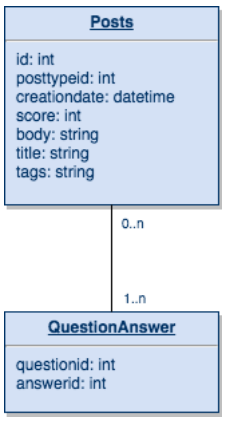

Figura 4.4: Modelo dos dados

A tabela "Posts"contém todas as postagens realizadas na comunidade, o campo "id"é o identificador único de cada postagem na comunidade; "posttypeid"identifica o tipo de postagem, no caso deste trabalho foram utilizados as perguntas e as respostas. As perguntas são identificadas com valor 1 e as respostas com o valor 2); o campo "creationdate"indica a data de postagem; "score"indica a quantidade de pontos obtidos advindo da votação dos usuários da comunidade ao avaliar a postagem, por meio deste campo serão identificadas as melhores respostas para a pergunta; o campo "body"contém a descrição da postagem; "title"contém o título da postagem; e por último o campo "tags"que

contém as categorias relacionadas a postagem, categorizadas pelos próprios usuários. A tabela "QuestionAnswer"contém o relacionamento entre as postagem registrando o par pergunta-resposta, cujo o campo "questionid"é o id da pergunta e o "answerid"é o id da resposta.

4.3 Extração de categorias de perguntas e respostas

Esta etapa é responsável por obter as categorias a partir da análise textual das perguntas e respostas. No Stack Exchange os tópicos das questões abordadas nas comunidades estão divididas em categorias, onde cada *tread* pode conter mais de um assunto, ao qual os membros da comunidade inserem as categorias relacionadas ao assunto, porém é considerado apenas para cada publicação (uma pergunta e as respostas associadas). Neste trabalho são utilizadas três ferramentas para a extração de categorias, o *DBpedia Spotlight*, o *Tag The Web* e o *Open Calais*. Considerar estas três ferramentas permite comparar o uso de entidades da Wikipédia para a categorização (*DBpedia Spotlight* e *Tag The Web*), bem como de outra base (Thomson Reuters). Além disto, vale observar a diferença no mecanismo de categorização provido pelo *DBpedia Spotlight* e pelo *Tag The Web*, sendo este último capaz de utilizar não apenas a identificação da entidade no texto, mas obter parte do grafo correspondente. Essas ferramentas utilizam algoritmos sofisticados com uso de ontologias, Linked Data e uso da Wikipédia que adicionam semântica as categorias, melhorando a qualidade da descrição das perguntas e respostas.

O exemplo abaixo apresenta a extração de categorias de duas perguntas obtidas da comunidade de biologia do Stack Exchange. Os termos sublinhados foram os termos extraídos pela ferramenta *DBpedia Spotlight*.

(i) Human disease and associated phenotype database? Does anyone know of any good databases that contain symptoms for diseases and other ailments in humans? I've tried working with UMLS, but thats been impossible to work with. I've also tried working with the disease ontology website (http://www.disease-ontology.org/), but that doesn't seem to be very...comprehensive...in terms of symptoms for the diseases they have.

(ii) Good source that explains the evolution of single-celledorganisms "from scratch"? Are there any books or sites that detail, step-by-step, the evolution of the first single-celled organisms (bacteria, archaea) from a Miller-Urey-like beginning? That is, assumes only amino acids, then from there to self-replicating proteins, until culminating in the formation of a basic cell?

O *DBpedia Spotlight* é uma ferramenta para anotação automática de menções de recursos da DBpedia em texto, fornecendo uma solução para vincular fontes de informação não estruturada ao Linked Open Data através da DBpedia. O *DBpedia Spotlight* reconhece nomes de conceitos ou entidades, em alguns idiomas, que foram mencionados (por exemplo, "Michael Jordan") e, subsequentemente, compara esses nomes a identificadores exclusivos (por exemplo, dbpedia:Michael_I._Jordan, o professor de inteligência artificial ou dbpedia:Michael_Jordan, o jogador de basquete) . Esta ferramenta disponibiliza uma API[4] no formato API REST, permitindo que textos não estruturados sejam anotados com recursos encontrados no DBpedia. Com uma abordagem de 4 passos realiza a *extração de entidades nomeadas*, incluindo a *detecção de entidade* e *resolução de nomes*, também *reconhecimento de entidades mencionadas* (Named Entity Recognition - NER). A API oferece três funcionalidades: (i) *spot*, que consiste na identificação a partir do texto de entrada as entidades que possam ser mencionadas; (ii) *candidates*, que utiliza o resultado do spot e escolhe o identificador para o candidato mais provável; e por último o (iii) *annotate*, que utiliza o resultado do *spot* e do *candidates*, realizando também a desambiguação e link.

Neste trabalho é utilizado a funcionalidade *annotate* do *DBpedia Spotlight*, o exemplo abaixo apresenta o retorno em formato text/XML. No XML a tag <Resource URI> contém os recursos encontrados baseados na DBpedia. Para cada recurso encontrado são disponibilizados algumas informações, a informação de referência no texto original (surfaceForm=), a posição no texto (offset=), o identificador associado a DBpedia (URI=), o tipo de recurso (types=). Outras propriedades da busca são apresentadas, indicando o quanto a entidade é proeminentemente baseada no número de *inlinks* da Wikipedia (suport=), o quanto a entidade está "na liderança"(percentageOfSecondRank=) baseado no quanto a entidade vencedora ganhou ao usar o contextualScore_2ndRank / contextualScore_1stRank, o que significa que quanto menor essa pontuação, mais a primeira entidade classificada está ”na liderança”, e por último a similaridade entre a entidade anotada e a entidade registrada (similarityScore=) de acordo com o Lucene[5], que é o motor de busca textual utilizado pelo *DBpedia Spotlight*

```
<Annotation text="President␣Obama␣called␣Wednesday␣on
␣␣Congress␣to␣extend␣a␣tax␣break␣for␣students␣included
␣␣in␣last␣year's␣economic␣stimulus␣package,␣arguing
␣␣that␣the␣policy␣provides␣more␣generous␣assistance."
  confidence="0.2" support="20">
  <Resources>
```

[4]https://www.dbpedia-spotlight.org/api

[5]http://succeed-project.eu/wiki/index.php/DBPedia_Spotlight

```
    <Resource
      URI="http://dbpedia.org/resource/Barack_Obama"
      support="5761"
      types="Person,Politician,President"
      surfaceForm="President␣Obama" offset="0"
      similarityScore="0.31504717469215393"
      percentageOfSecondRank="-1.0" />
    <Resource
      URI="http://dbpedia.org/resource/United_States_Congress"
      support="8569" types="Organisation,Legislature"
      surfaceForm="Congress" offset="36"
      similarityScore="0.2348192036151886"
      percentageOfSecondRank="0.8635579006818564" />
    <Resource
      URI="http://dbpedia.org/resource/Tax_break"
      support="32" types="" surfaceForm="tax␣break"
      offset="57" similarityScore="0.35041093826293945"
      percentageOfSecondRank="-1.0" />
    <Resource
      URI="http://dbpedia.org/resource/Student"
      support="1701" types="" surfaceForm="students" offset="71"
      similarityScore="0.32534149289131165"
      percentageOfSecondRank="-1.0" />
    <Resource
      URI="http://dbpedia.org/resource/Policy"
      support="557" types="" surfaceForm="policy" offset="148"
      similarityScore="0.3228176236152649"
      percentageOfSecondRank="-1.0" />
  </Resources>
</Annotation>
```

O *Open Calais*, é serviço Web da Thomson Reuters que anexa tags de metadados inteligentes a conteúdos não estruturados, permitindo a análise de texto. Utiliza o mecanismo de processamento de linguagem natural para analisar e marcar automaticamente seus arquivos de entrada de forma que possa identificar facilmente dados relevantes e obter informações contidas no texto. O *Open Calais* analisa o conteúdo semântico de seus arquivos de entrada usando uma combinação de métodos estatísticos, de aprendizado de

máquina e baseados em padrões personalizados. Para isso, utiliza metadados desenvolvidos por membros da equipe de dados da Thomson Reuters, o grupo *Text Metadata Services* (*TMS*).

O *Open Calais* analisa o texto de entrada e realiza os seguintes processos: (i) *Nomear Entidade e Reconhecer Relacionamento*, que consiste em identificar e marcar menções (strings de texto), possibilitando identificar empresas, pessoas, negócios, localizações geográficas, indústrias, organizações, produtos, eventos, etc., com base em uma lista de tipos de metadados predefinidos; e (ii) *Aboutness Tagging*, nesta etapa atribui tags que descrevem o documento de entrada como um todo. Na etapa de *Aboutness Tagging* as tags são divididas em: tags sociais (*social tagging*), que classificam os documentos com base na folksonomia da Wikipédia; tags de tópicos (*topic tagging*), que identifica os tópicos discutidos no documento, a lista de referência de tópicos é retirada das taxonomias do *Thomson Reuters Coding Schema* (*TRCS*) e do *International Press Telecommunications Council* (*IPTC*); e por último, as tags industriais (*Industry Tagging*), que identificam as indústrias relacionadas ao texto, a lista de setores que podem ser identificados é definida pela taxonomia *Thomson Reuters Business Classification* (*TRBC*). O *Open Calais* disponibiliza uma API[6] com uma interface API REST, possibilitando a análise textual assim atribuindo tags para a classificação dos documentos.

Neste trabalho é utilizado a tag de tópico (*topic tagging*), o exemplo abaixo apresenta a extração das tags sociais, extraídos pelo *Open Calais"* do artigo sobre a Apple desenvolvendo um carro autônomo, em destaque o atributo ”importance”[7], que indica como o tópico nomeado pela tag social é centrado no documento como um todo. O valor do atributo de importância pode ser 1 (muito cêntrico), 2 (um tanto cêntrico) ou 3 (menos cêntrico).

```
<rdf:Description
      rdf:about="http://d.opencalais.com/dochash-1/
␣␣␣␣7586b818-40af-3d55-ac16-␣bf3520cddfa6/SocialTag/1">
  <rdf:type
  rdf:resource="http://s.opencalais.com/1/type/tag/SocialTag"/>
  <c:docId
      rdf:resource="http://d.opencalais.com/dochash-1/
␣␣␣␣7586b818-40af-3d55-ac16-bf3520cddfa6"/>
  <c:socialtag
```

[6]http://www.opencalais.com/opencalais-api/

[7]http://www.opencalais.com/wp-content/uploads/folder/ThomsonReutersOpenCalaisAPIUserGuideR11_9.pdf, pag 7

```
   rdf:resource="http://d.opencalais.com/genericHasher-
␣␣1/1205cb52-d703-34d2-83b2-
␣␣a09d4d47575c"/>
   <c:forenduserdisplay>true</c:forenduserdisplay>
   <c:name>Apple Inc.</c:name>
   <c:importance>1</c:importance>
   <c:originalValue>Apple Inc.</c:originalValue>
   </rdf:Description>
   <rdf:Description
   rdf:about="http://d.opencalais.com/dochash-1/
␣␣7586b818-40af-3d55-ac16-␣bf3520cddfa6/SocialTag/2">
   <rdf:type
   rdf:resource="http://s.opencalais.com/1/type/tag/SocialTag"/>
   <c:docId
      rdf:resource="http://d.opencalais.com/dochash-1/
␣␣7586b818-40af-3d55-ac16-bf3520cddfa6"/>
   <c:socialtag
   rdf:resource="http://d.opencalais.com/
␣␣genericHasher-1/ccc60460-211d-3b02-b13e11fba4449fbd"/>
   <c:forenduserdisplay>true</c:forenduserdisplay>
   <c:name>Autonomous car</c:name>
   <c:importance>1</c:importance>
   <c:originalValue>Autonomous car</c:originalValue>
</rdf:Description>
```

O *Tag The Web*, é uma ferramenta que utiliza a estrutura taxonômica da Wikipédia, baseado na geração de um *fingerprint* por meio da relação semântica entre os nós do grafo de categorias da Wikipédia. A relação semântica entre os nós das categoria da Wikipédia é dada por meio da medição da distância entre os nós e então gerado uma impressão digital (*fingerprint*) com base na influência de cada categoria de topo no documento classificado. A classificação calculada é representada como um vetor multidimensional, facilitando a recuperação e a comparação de documentos. Apresenta uma cadeia de processamento para gerar a categorização genérica com base em três etapas: (i) anotação textual, as quais são identificadas as entidades a partir do texto de entrada, de acordo com os recursos do DBpedia, é utilizado o *DBpedia Spotlight* para realizar a extração e enriquecimento das entidades encontradas dos recursos de web; (ii) extração de categorias, a partir das entidades encontradas são realizadas consultas no DBpedia com a finalidade

de encontrar relacionamentos para tornar uma representação mais genérica das entidades, obtendo assim as categorias relacionadas; (iii) Geração de *Fingerprint*, o objetivo desta etapa é atribuir um conjunto de tópicos principais dentro das categorias da Wikipédia a um determinado recurso da web.

O Tag The Web disponibiliza uma API[8] em formato REST, oferecendo três funcionalidades: (i) *Generate Fingerprint,* que dado qualquer texto ou URL como entrada, retorna a distribuição percentual de tópicos ao longo das categorias de tópicos da Wikipédia e também possibilitando escolher o nível no grafo ao longo das categorias por meio do parâmetro "depth"; (i) *Get Paths*, retorna as subárvores usadas na composição da distribuição, considerando qualquer texto; e (iii) *Generate Paths*, retorna as subárvores usadas na composição da distribuição, dada uma lista de categorias separadas por pipes (|).

Neste trabalho é utilizado a funcionalidade *Get Paths* do *Tag The Web* e por meio do parâmetro "*depth*"com valor igual 4, assim descendo 5 níveis a partir das categorias de tópicos do Wikipédia, o que possibilita enriquecer as categorias com obtenção das categorias relacionadas. Abaixo um exemplo do JSON de retorno, contendo a distribuição percentual ao longo das 19 categorias de topo de tópicos da Wikipédia.


```
{
   "Culture": 0.14035087719298,
   "Religion": 0.087719298245614,
   "Matter": 0.0058479532163743,
   "Life": 0.011695906432749,
   "Law": 0.1812865497076,
   "Industry": 0.017543859649123,
   "Games": 0,
   "Arts": 0.029239766081871,
   "Science_and_technology": 0.017543859649123,
   "Society": 0.093567251461988,
   "Humanities": 0.14035087719298,
   "Health": 0.070175438596491,
   "Reference_works": 0.0058479532163743,
   "Nature": 0.0058479532163743,
   "Geography": 0.011695906432749,
   "History": 0.035087719298246,
   "Philosophy": 0.052631578947368,
   "People": 0.093567251461988,
```


[8] http://tagtheweb.com.br/wiki/getFingerPrint.php

```
  "Mathematics": 0
 }
```

4.3.1 Método de Extração de categorias

Foram submetidos o texto das postagens para obter as categorias a partir da análise textual das perguntas e respostas. Um algoritimo foi desenvolvido utilizando a linguagem de programação Python com a função de enviar várias requisições HTTP3 as APIs das ferramentas, para cada serviço, em seguida, salvar os dados retornados. As categorias foram armazenadas em um banco de dados não relacional MongoDB, utilizando uma coleção (collection) para cada ferramenta de extração de categorias utilizada, no caso três coleções, nomeadas de: ”dbpedia”, ”opencalais” e ”tagtheweb”. Os dados são armazenados em formato JSON, conforme a estrutura abaixo.

```
{
"id" : int,
 "status" : int,
 "terms" : [],
 "extra" : {},
}
```

A estrutura de dados contém os seguintes elementos: o ”id” é o identificador único da pergunta ou respostas; o ”status” indicando o status da execução, o valor 0 foi atribuído quando ocorre algum erro na execução da extração das categorias, indicando que deverá ser realizado novamente, o valor 1 foi atribuído quando houve sucesso na extração das categorias; o elemento ”terms” é uma lista com as entidades extraídas dos texto; e por último ”extra”, são dados extras particulares de cada ferramenta de extração de categorias, que posteriormente foram utilizados para fazer análises dos dados. No caso do DBpedia Spotlight foram armazenados os campos ”offset”, ”similarityScore”, ”percentageOfSecondRank” e ”support”. O Tag The Web foi armazenado todo o dicionário retornado pela ferramenta, que possui distribuição percentual de tópicos ao longo das categorias. E por último no Open Calais foi armazenado o campo ”importance”.

4.4 Cálculo da similaridade semântica

Nesta etapa é realizado o cálculo da similaridade semântica, utilizando o *Semantic Connectivity Scores* (*SCS*), com objetivo de verificar a relação semântica entre as pergun-

tas e respostas. O Semantic Connectivity Scores (SCS) é uma medida baseada em co-ocorrência e Web Semântica para descobrir relacionamentos entre entidades. Apresenta um score de conectividade semântica baseado no índice de Katz [13], para mensurar o parentesco entre entidades por meio da estrutura de grafos de categorias da Wikipédia. O SCS apresenta o score semântico em uma escala entre 0 e 1, que indica a similaridade entre dois termos. O SCS oferece uma API REST[9], tendo como base o DBpedia para realizar o cálculo de score semântico entre dois termos, que são entidades presentes no DBpedia.

Neste trabalho para cada termo extraído do texto da pergunta e da resposta é calculado o score de conectividade semântica, o SCS, sobre as respectivas entidades do par perguntaresposta, tendo como entrada dois termos, um da pergunta e outro da resposta, e como retorno um score semântico SCS. É realizado este cálculo para todos termos da pergunta e resposta, que são combinadas entre todos os termos então submetidas para a API realizado o cálculo individualmente para cada par de termo da pergunta-resposta. Por fim é obtido todos os scores semântico de cada par de termo, que pode ser representado como uma matriz, contendo todos os scores semânticos entre cada termo pergunta-resposta. Ao final para se obter um valor de score entre a pergunta e resposta é realizado uma média aritmética entre todos os scores de cada par de termo, obtendo assim o score semântico entre a pergunta e resposta. Com o score semântico de todas as perguntas e respostas, é possível realizar um ranqueamento com base no valor de score, ao qual o valor mais alto representa que há uma maior relação semântica entre os dois textos comparados. A Tabela 4.1 apresenta a matriz com os valores de score entre os termos da pergunta (Eukaryote, Intron, Macronucleus e Exon) e da resposta (Germline, Ribosomal_RNA e Intron).

	Germline	Ribosomal_RNA	Intron
Eukaryote	0.88	0.5	0.82
Intron	0.56	0.88	0.87
Macronucleus	0.92	0.89	0.92
Exon	0.94	0.91	0.91

Tabela 4.1: Matriz de score semântico entre os termos extraídos do par pergunta-resposta

Após a execução os valores de score semântico foram armazenadas em um banco de dados não relacional MongoDB, o que utiliza uma coleção (collection) para cada ferramenta de extração de categorias utilizada, DBpedia, Open Calais e Tag The Web. Os dados são armazenados em formato JSON, conforme a estrutura a seguir:

```
{
"id" : "10_22",
```

[9]http://semanticweb.inf.puc-rio.br/similarities.json

```
"general_status" : 1,
"metrics" : {
"qut_items" : 12,
"average_similarity" : 0.9265242951682083,
"sum_similarity" : 11.1182915420185
},
"terms" : {
"Eukaryote" : {
"Germline" : {
"status" : 1,
"scs_score" : 0.93333333333333
},
"Ribosomal_RNA" : {
"status" : 1,
"scs_score" : 0.94366197183099
},
"Intron" : {
"status" : 1,
"scs_score" : 0.94117647058824
},
},
"Intron" : {
"Germline" : {
"status" : 1,
"scs_score" : 0.89473684210526
},
"Ribosomal_RNA" : {
"status" : 1,
"scs_score" : 0.93846153846154
},
"Intron" : {
"status" : 1,
"scs_score" : 1 },
},
"Anatomical_terms_of_location" : {
"Germline" : {
"status" : 1,
```

```
  "scs_score" : 0.89473684210526
  },
  "Ribosomal_RNA" : {
  "status" : 1,
  "scs_score" : 0.88235294117647
  },
  "Intron" : {
  "status" : 1,
  "scs_score" : 0.89473684210526
  },
  },
  "Exon" : {
  "Germline" : {
  "status" : 1,
  "scs_score" : 0.91304347826087
  },
  "Ribosomal_RNA" : {
  "status" : 1,
  "scs_score" : 0.93333333333333
  },
  "Intron" : {
  "status" : 1,
  "scs_score" : 0.94871794871795
  },
  },
  },
}
```


A estrutura de dados contém os seguintes elementos: o "*id*" é o identificador único do par pergunta-respostas em uma string única ("*questionid_answerid*"); o "*general_status*" indicando o status da execução de todas as entidades do par pergunta-resposta, o valor 0 foi atribuído quando ocorre algum erro na execução de pelo menos um cálculo de similaridade, indicando que deverá ser realizado novamente, o valor 1 foi atribuído quando houve sucesso em todos os cálculos de similaridade; o elemento "*metrics*" são alguns cálculos realizados sobre o resultado, contendo: "*qut_items*", que contém a quantidade de pares de entidades pergunta-respostas; "*average_similarity*", que engloba a média aritmética entre todos os *scores* de similaridade; "*sum_similarity*", que inclui a soma dos scores de simila-

ridade; o elemento ”*terms*” é uma dicionário contendo os termos da pergunta como chave e como valor outro dicionário abarcando os termos das respostas como chave, dentro dele temos como valor um dicionário contendo: ”*status*”, que indica se aquele execução do cálculo da similaridade obteve sucesso ou não; e o ”*scs_score*” que é o valor de *score* de similaridade.

Após realizar o cálculo de similaridade entre os termos das perguntas e respostas é então obtido o score de similaridade semântica entre a pergunta e resposta. Com este score é possível realizar um ranqueamento com base no valor de score, onde o valor mais alto indica uma maior relação semântica entre os dois textos comparados. Em cenários de recomendação de respostas, para cada pergunta é realizado o cálculo entre as possíveis respostas obtendo-se um ranqueamento das melhores respostas com base no valor de similaridade. As respostas com maior score de similaridade podem ser consideradas as respostas com maior chance de responder a pergunta. O capítulo 5 apresenta de forma detalhada os experimentos que foram realizados, utilizando todas a etapas do método proposto e a avaliação do método com dados de comunidades reais utilizando este ranqueamento com base no score de similaridade, comparado aos pontos atribuídos pelos usuários das comunidades às respostas.

5. Experimentos e Resultados

Este capítulo tem como objetivo apresentar os experimentos realizados e os resultados alcançados. Primeiro serão apresentados algumas métricas e características extraídas das comunidade online de perguntas e respostas. Posteriormente serão analisados os resultados da extração de categorias a partir das perguntas e respostas. Em seguida é apresentado o cenário de caso utilizado para avaliar o método proposto, realizando o cálculo da similaridade semântica. Por último uma discussão dos resultados obtidos e reflexões sobre o método abordado sobre os experimentos realizados.

5.1 Dataset e Características Gerais das Comunidades

Com a finalidade de avaliar a proposta dessa dissertação, primeiramente, é necessário extrair um conjunto de dados de comunidades online reais. Para isso, foram escolhidas três distintas comunidades online de perguntas e resposta do Stack Exchange. As comunidades escolhidas para esse estudo foram as seguintes:

- Biologia[1]: uma comunidade destinada para pesquisadores, acadêmicos e estudantes de biologia;
- História[2]: uma comunidade voltada para historiadores e entusiastas de história.
- Legislação[3]: uma comunidade destinada para profissionais jurídicos, estudantes e outros com experiência ou interesse em lei.

A principal motivação para seleção destas comunidades foi pelo por apresentarem uma grande quantidade de postagens e o conteúdo mais textual, o que facilita a extração de

[1] https://biology.stackexchange.com/
[2] https://history.stackexchange.com/
[3] https://law.stackexchange.com/

categorias, comunidades com elementos contendo muitas fórmulas ou trechos de códigos requerem um pré-processamento, como relatado no trabalho em [53] e [54]. A Tabela 5.1 mostra os dados coletados e algumas características gerais das comunidades como: número de usuários, número de mensagens, número de perguntas, número de respostas, número de comentários, etc.

Tabela 5.1: Características gerais das comunidades

Comunidade	**Número de mensagens**	**Número de perguntas**	**Número de respostas**	**Número de respostas**	**Quantidade média de caracteres / postagens**	**Número de usuários**
Biologia	120.522	19.939	22.933	77.650	929	30.646
História	102.603	8.787	17.495	76.321	1.232	21.033
Legislação	59.180	9.490	11.840	37.850	1.111	13.706

Na Tabela 5.1, pode-se observar que há uma grande troca de mensagens entre os membros das comunidades ao responderem as perguntas, o número de comentários supera o número de perguntas e respostas. As respostas são editadas conforme a necessidade de correção e aprimoramento, isso mostra que os comentários, além das respostas, tem uma importante contribuição no aprimoramento das respostas às questões.

A quantidade média de caracteres escritos nas postagens é bem parecida nas três comunidades. Como texto da pergunta foi considerado o título da pergunta juntamente com sua descrição, enquanto como resposta foi considerado o texto da resposta. Os comentários não foram considerados nem para as perguntas e nem para as respostas. Observamos também que a média de caracteres por postagem ultrapassa a mil caracteres, o que pode indicar que são respostas bem elaboradas e não apenas textos curtos.

A Tabela 5.2 mostra alguns dados com relação às respostas das perguntas na comunidade, como: número de perguntas respondidas, média de respostas por pergunta, número de perguntas com apenas uma respostas, etc.

Tabela 5.2: Características das respostas das comunidades

Comunidade	**Número de perguntas**	**Número de perguntas respondidas**	**Percentual perguntas respondidas**	**Média de respostas por pergunta**	**Número de perguntas com uma resposta**	**Número máximo de respostas por perguntas**
Biologia	19.939	15.718	79%	1,5	10.652	11
História	8.787	7.985	91%	2,2	5.146	34
Legislação	9.490	7.945	84%	1,5	3.661	13

A Tabela 5.2 ilustra que uma comunidade grande, como a de biologia, a taxa de perguntas sem respostas ultrapassa os 20%. Percebe-se que a quantidade de usuários não

está diretamente relacionada ao maior número de respostas e comparando-se as três comunidades, observa-se que apesar da comunidade de biologia apresentar maior número de membros, ela apresenta um percentual maior de perguntas não respondidas. Observa-se também que em geral as comunidades apresentam poucas respostas por perguntas, grande parte delas apenas uma resposta por pergunta.

Os tópicos das questões abordadas nas comunidades estão divididas em categorias, onde cada *tread* pode conter mais de um assunto, ao qual os membros da comunidade inserem as categorias relacionadas ao assunto. Na Tabela 5.3 são apresentados alguns dados com relação às categorias.

Tabela 5.3: Características das categorias das comunidades

Comunidade	**Número de treads**	**Número de categorias**	**Média de categorias por tread**
Biologia	19939	716	2,5
História	8787	747	2,8
Legislação	9490	628	2,5

Na tabela 5.3 se pode perceber que há um grande número de categorias, com destaque para a comunidade de história, que apresenta 747 Categorias. Observa-se que o tamanho da comunidade e o número de *treads* não está relacionado à quantidade de assuntos abordados na comunidade. Os membros da comunidade, em geral, atribuem para cada *tread* mais de duas categorias.

5.2 Extração de categorias

Os textos das perguntas e respostas das três comunidades foram utilizados para realizar a extração de categorias, conforme descrição na etapa 2 do método proposto na Seção 4. O resultado encontrado em cada comunidade é apresentado na Tabela 5.4 para a comunidade de biologia, Tabela 5.5 para a comunidade de história e Tabela 5.6 para a comunidade de legislação.

Tabela 5.4: Sumário extração de categorias de perguntas e respostas da comunidade de Biologia

Ferramenta de extração categorias	**Número de perguntas/respostas**	**Percentual de perguntas/respostas**	**Número de categorias extraídas**	**Média categorias por pergunta/respostas**
DBpedia Spotlight	40296	94%	314392	7,8
Open Calais	35716	83%	275838	7,7
Tag The Web	41102	96%	2023053	49,2

Na comunidade de biologia (Tabela 5.4) foi possível extrair categorias para a grande maioria das perguntas e respostas, com destaque para a ferramenta de extração Tag The Web que identificou na maioria dos os textos, com 96% de cobertura. Com relação a quantidade de categorias obtidas, o DBpedia Spotlight e o Open Calais apresentaram uma média muito próxima, entre 7,7 e 7,8 de categorias por pergunta/resposta. O Tag The Web apresentou um grande número de categorias, com média de 49,2 categorias por pergunta/resposta, explicado pelo fato dele também retornar categorias relacionadas com base no grafo da Wikipédia.

Tabela 5.5: Sumário Extração de categorias de perguntas e respostas da comunidade de História

Ferramenta de extração categorias	**Número de perguntas/respostas**	**Percentual de perguntas/respostas**	**Número de categorias extraídas**	**Média categorias por pergunta/respostas**
DBpedia Spotlight	25.837	98%	265.000	10,3
Open Calais	23.977	91%	13.5849	5,7
Tag The Web	26.009	99%	2.791.036	107,3

Na comunidade de história (Tabela 5.5) também foi possível extrair categorias para a grande maioria das perguntas e respostas, com destaque para a ferramenta de extração *DBpedia Spotlight* e *Tag The Web* que identificou na maioria dos os textos, com quase 100% de cobertura. Com relação a quantidade de categorias obtidas, o *DBpedia Spotlight* apresentou uma média de 10,3 categorias por pergunta/resposta, enquanto que o *Open Calais* apresentou 5,7 categorias por pergunta/resposta. O *Tag The Web* apresentou um grande número de categorias, com média de 107,3 categorias por pergunta/resposta.

Tabela 5.6: Sumário Extração de categorias de perguntas e respostas da comunidade de Legislação

Ferramenta de extração categorias	**Número de perguntas/respostas**	**Percentual de perguntas/respostas**	**Número de categorias extraídas**	**Média categorias por pergunta/respostas**
DBpedia Spotlight	19.562	91%	95.400	4,9
Open Calais	19.512	91%	107.423	5,5
Tag The Web	21.221	99%	1.078.447	50,8

Na comunidade de legislação (Tabela 5.6), assim como nas demais foi possível extrair categorias para a grande maioria das perguntas e respostas, com destaque também para a ferramenta de extração *Tag The Web* que quase identificou na maioria dos os textos, com 99% de cobertura. Com relação a quantidade de categorias obtidas, o *DBpedia Spotlight* e o *Open Calais* apresentaram uma média muito próxima, entre 4,9 e 5,5 de categorias por pergunta/resposta. O *Tag The Web* apresentou um grande número de categorias, com média de 50,8 categorias por pergunta/resposta.

Como podemos observar a extração de categorias nas três comunidades, as três ferramentas conseguiram extrair categorias para a grande maioria das perguntas e respostas,

na maioria dos casos superando 90% das perguntas e respostas, com destaque ao *Tag The Web* que quase identificou na maioria dos os textos, com 99% de cobertura em duas comunidades. Com relação a quantidade de categorias obtidas, o *DBpedia Spotlight* e o *Open Calais* apresentaram resultados semelhantes, com médias de categorias por pergunta e resposta muito próximas, média geral de 7,6 e 6,3 respectivamente. O *Tag The Web* apresentou um grande número de categorias, com média geral de 69,1 categorias por pergunta/resposta, explicado pelo fato dele também retornar categorias relacionadas com base no grafo da Wikipédia.

5.3 Cálculo de similaridade semântica

Para avaliar o método proposto nesta dissertação, usando o cálculo da similaridade semântica, os resultados gerados nesta abordagem foram comparados com o resultado da votação de usuários da comunidade ao avaliar uma resposta, utilizando as perguntas e apenas suas respostas, de modo que as relações semânticas possam ser confrontadas. Este cenário define se é possível estabelecer relações semânticas entre os termos extraídos das perguntas e respostas a partir do cálculo de similaridade semântica de cada par do termo pergunta-resposta.

Na etapa 3 da arquitetura implementada, foi criado um cenário para avaliar o método para recomendar a melhor resposta, para isto foram utilizadas perguntas respondidas contendo mais de uma resposta. Para selecionar a melhor resposta foram utilizados os pontos atribuídos às respostas pelos usuários ao avaliá-las, de modo a confrontar ao *score* semântico obtido pelo cálculo do SCS. Neste cenário foram utilizadas as categorias obtidas pelo *DBpedia Spotlight* e *Open Calais*, e então calculado o *score* semântico entre cada par de termo pergunta-resposta. Foram calculados para as três comunidades, Q&A Biologia, Q&A História e Q&A Legislação e os resultados encontrados em cada comunidade são apresentados na Tabela 5.7 para a comunidade de biologia, na Tabela 5.8 para a comunidade de história e na Tabela 5.9 para a comunidade de legislação.

Tabela 5.7: Sumário cálculo de similaridade semânticas de perguntas e respostas da comunidade de Biologia

Ferramenta de extração categorias	**Número de perguntas**	**Número de acerto**	**Percentual acerto**	**Valor médio do *score SCS***
DBpedia Spotlight	4889	2077	42,4 %	0,77
Open Calais	5066	2559	50,5%	0,81

Na comunidade de biologia (Tabela 5.7) foi observado que o *score* semântico obteve

um valor de 0,77 de média com as categorias extraídas do *DBpedia Spotlight* e 0,81 de média com as categorias extraídas do *Open Calais*, o que indica que os termos extraídos das perguntas e respostas possuem uma relação semântica. Comparando os resultados gerados por nossa abordagem com o resultado advindo da votação dos usuários da comunidade ao avaliar a melhor resposta, o modelo proposto encontrou a melhor resposta com as categorias do *DBpedia Spotlight* para 2.077 perguntas de um total de 4.889 e com as categorias do *Open Calais* para 2.559 perguntas de um total de 5.066. Assim, nesta comunidade encontrou o maior score de similaridade para a melhor resposta avaliada pelos usuários, com uma taxa de sucesso entre 42% e 50% de acerto, com melhor resultado para as categorias extraídas do *Open Calais*.

Tabela 5.8: Sumário cálculo de similaridade semânticas de perguntas e respostas da comunidade de História

Ferramenta de extração categorias	**Número de perguntas**	**Número de acerto**	**Percentual acerto**	**Valor médio do *score SCS***
DBpedia Spotlight	763	271	35,5%	0,81
Open Calais	1792	727	40,1%	0,73

Na comunidade de história (Tabela 5.8) foi observado que o *score* semântico obteve um valor de 0,81 de média com as categorias extraídas do *DBpedia Spotlight* e 0,73 de média com as categorias extraídas do *Open Calais*, o que indica que os termos extraídos das perguntas e respostas possuem uma relação semântica. Comparando os resultados gerados por nossa abordagem com o resultado advindo da votação dos usuários da comunidade ao avaliar a melhor resposta, o modelo proposto encontrou a melhor resposta com as categorias do *DBpedia Spotlight* para 271 perguntas de um total de 763 e com as categorias do *Open Calais* para 727 perguntas de um total de 1792. Assim, nesta comunidade encontrou o maior *score* de similaridade para a melhor resposta avaliada pelos usuários, com uma taxa de sucesso entre 35% e 40% de acerto, com melhor resultado para as categorias extraídas do *Open Calais*.

Tabela 5.9: Sumário cálculo de similaridade semânticas de perguntas e respostas da comunidade de Legislação

Ferramenta de extração categorias	**Número de perguntas**	**Número de acerto**	**Percentual acerto**	**Valor médio do *score SCS***
DBpedia Spotlight	1479	673	45,5%	0,82
Open Calais	1479	672	45,4%	0,70

Na comunidade de legislação (Tabela 5.9) foi observado que o *score* semântico obteve um valor de 0,82 de média com as categorias extraídas do *DBpedia Spotlight* e 0,7 de

média com as categorias extraídas do *Open Calais*, o que indica que os termos extraídos das perguntas e respostas possuem uma relação semântica. Comparando os resultados gerados por nossa abordagem com o resultado advindo da votação dos usuários da comunidade ao avaliar a melhor resposta, o modelo proposto encontrou a melhor resposta com as categorias do *DBpedia Spotlight* para 673 perguntas de um total de 1479 e com as categorias do *Open Calais* para 672 perguntas de um total de 1479. Assim, nesta comunidade encontrou o maior *score* de similaridade para a melhor resposta avaliada pelos usuários, com resultado semelhante com as categorias extraídas pelas duas ferramentas, com uma taxa de sucesso de 45% de acerto.

5.3.1 Discussão dos resultados

Considerando que neste cenário as respostas são relacionadas a pergunta, foi observado que entre as três comunidades o *score* semântico obteve um valor médio entre 0,7 e 0,8 entre as categorias extraídas pelo *Dbpedia Spotlight* e pelo *Open Calais*, o que indica que os termos extraídos das perguntas e respostas possuem uma relação semântica entre eles.

Comparando os resultados gerados por nossa abordagem com o resultado advindo da votação dos usuários da comunidade ao avaliar a melhor resposta, o trabalho aqui apresentado encontrou o maior score de similaridade para a melhor resposta avaliada pelos usuários, com uma taxa de sucesso entre 35,5% e 50,5% de acerto. Deste modo, a abordagem proposta nesta dissertação pode ser utilizada não apenas para identificar a correlação entre perguntas e respostas, mas também consegue encontrar as melhores respostas em muitos casos. Comparando os resultados entre as categorias extraídas, no geral apresentaram resultados semelhantes, com uma pequena margem de vantagem para as categorias extraídas pelo *Open Calais*.

Os resultados obtidos indicam que é possível estabelecer relações semânticas a partir da extração de categorias e com cálculo de similaridades semântica dos termos extraídos, indicando que este método pode ser utilizados como suporte à recomendações de respostas para questões em aberto em comunidades Q&A, sem a intervenção de especialistas.

6. Conclusão

Neste capítulo serão apresentadas as conclusões finais deste trabalho, onde serão descritas as principais contribuições e suas limitações. Além disso, serão levantadas algumas sugestões de trabalhos futuros para a continuação da pesquisa realizada nesta dissertação.

6.1 Comentários Finais e Conclusão

Esta dissertação teve como foco principal realizar um estudo a fim de identificar relações semânticas entre categorias textuais das perguntas e respostas em comunidades de perguntas e respostas. Investigando se a partir da similaridade semântica de categorias obtidas entre perguntas e respostas é possível estabelecer relações semânticas entre eles. Também foram estabelecidos os referenciais teóricos para estudar os conceitos fundamentais que nortearam todo o trabalho conduzido.

Primeiramente, foi apresentado um novo modelo de busca de informação, que surgiu através da Consulta Social, que consiste num esforço de modificar as relações sociais na busca de informação com benefício do "conhecimento da multidão". No Paradigma de Vila através da Consulta Social a forma das pessoas partilharem suas dificuldades em um contexto de cooperação apareceu nos fóruns online e Sites de Perguntas e Respostas. Plataformas online de Q&A atuam como importante papel na educação, principalmente no contexto de cenários de aprendizagem informal. Participantes de comunidade de Q&A colocam suas questões e obtém respostas, com *feedback* e sugestões de outros usuários em um curto espaço de tempo depois de publicar a sua questão a ser respondida. Esse sistema atua como um ambiente dinâmico que promove uma aprendizagem social e colaborativa, o que contribui para o crescente número de usuários para plataformas, no entanto, muitas questões permanecem sem resposta, ocasionando um problema de falta de *feedback* para o usuário, o que pode gerar desmotivação para uso da plataforma, e uma metodologia automática poderia recomendar respostas relacionadas.

Neste contexto, foram apresentados trabalhos da comunidade científica cujo objetivo era reduzir o número crescente de perguntas sem respostas em comunidade Q&A e de sistemas de perguntas e respostas. Foram apresentadas as abordagens, técnicas e metodologias dos trabalhos relacionados em três perspectivas: identificar questões semelhantes, responder perguntas com respostas anteriores, ambas no contexto de comunidades de perguntas e respostas, e por último no contexto de sistemas de perguntas e respostas.

Foi apresentado a arquitetura da solução proposta, que consiste em três etapas: obtenção de perguntas e respostas, extração de categorias e cálculo de similaridade semântica. Detalhando a arquitetura com os modelos de dados e ferramentas utilizadas em cada processo.

Em seguida, iniciou-se o estudo nas três comunidades selecionadas para as análises que este trabalho se propôs a fazer com a finalidade de verificar a hipótese apresentada no Capitulo 1. Inicialmente foi apresentado como os dados dessas três comunidades foram coletados e também foram mostradas algumas características de cada uma das comunidades.

Para investigar a hipótese comparamos os resultados gerados por nossa abordagem com o resultado advindo da votação dos usuários das comunidades ao avaliar a resposta, de modo que as relações semânticas possam ser confrontadas, verificando se as respostas recomendadas para uma questão sem resposta poderiam contribuir para responder a questão.

Observa-se ainda que este trabalho permitiu comparar os resultados de três diferentes abordagens para identificar entidades à partir de texto, a saber: *DBpedia Spotlight, Open Calais* e *Tag The Web*. Embora *DBpedia Spotlight* e *Open Calais* tenham obtido resultados similares, com leve vantagem para o último, o *Tag The Web* trouxe um conjunto de categorias (ou entidades) bem superior. Este resultado pode ser ao mesmo tempo positivo (por permitir um conjunto maior de entidades a serem processados e portanto encontrar as correspondências entre perguntas e respostas) e negativo (devido ao custo de processamento e o excesso de entidades, que também poder gerar distorções nos resultados, o que poderia ser configurado na ferramenta, considerando a priorização dos resultados).

Considerando o escopo deste trabalho, pode-se considerar que a hipótese apresentada foi comprovada. Os resultados obtidos indicam que é possível estabelecer relações semânticas a partir da extração de categorias e com cálculo de similaridades semântica dos termos extraídos, indicando que este método pode ser utilizados como suporte à recomendações de respostas para questões em aberto em comunidades Q&A, sem a intervenção

de especialistas.

Parte dos resultados obtidos nesta dissertação foram apresentados no artigo:

Exploring the Correlation of Semantic Entities Between Questions and Answers in Q&A Communities. Euro American Conference on Telematics and Information Systems (EATIS), 2018.

Este trabalhou se limitou a análise de três comunidades apenas. Contudo, sabe-se que análises em maior número de comunidades são necessárias para que seja possível alcançar resultados mais seguros, tornando assim possível a elaboração de um método que pode ser utilizado como suporte à recomendações de respostas para questões em aberto em qualquer comunidade online (ou talvez algumas com determinadas características).

6.2 Contribuições

A principal contribuição dessa pesquisa foi o estudo a fim de identificar relações semânticas entre categorias textuais das perguntas e respostas em comunidades de perguntas e respostas. Desenvolvendo uma arquitetura para avaliar o uso de categorias extraídas partir da análise textual das perguntas e respostas para representação semântica com uma métrica de similaridade semântica para confrontar as relações semânticas. Através dessa arquitetura foi possível comprovar a hipótese apresentada nesta dissertação. Além disto, as contribuições técnicas da pesquisa são:

- Construção de *scripts* linguagem de programação Python para a extração de dados das comunidades online estudadas neste trabalho.
- Construções de algoritmos linguagem de programação Python para tornar possível a realização das análises apresentadas.
- Um levantamento bibliográfico detalhado visando mostrar características de diversos trabalhos já realizados na comunidade científica sobre o tema e também a comparação entre eles.
- Estudo de algumas características de comunidades online, ressaltando suas semelhanças e diferenças.
- Comparação de três ferramentas para identificar entidades à partir de texto.

6.3 Trabalhos Futuros

Como trabalhos futuros podem ser citados o desenvolvimento de um mecanismo de recomendações de perguntas similares ou respostas compatíveis às que estão sendo postadas, com base no trabalho apresentado nesta dissertação. Também seria possível considerar não apenas a identificação de correlações entre perguntas e respostas, mas a semântica do relacionamento entre as postagens (perguntas que complementam outras, respostas que complementam outras, perguntas/respostas que são mais genéricas ou específicas do que outra e assim por diante).

Finalmente, em [35], os autores descrevem uma análise do comportamento do usuário e a influência de usuários especialistas nas discussões da comunidade. Outro trabalho futuro poderia ser usar o trabalho apresentado nesta dissertação para investigar o comportamento dos usuários e recomendar perguntas para especialistas.

Referências Bibliográficas

[1] ALLEN, R. "Search engine statistics", *Smart Insights*, , 2017.

[2] ANDY, A., SEKINE, S., RWEBANGIRA, M., et al. "Name variation in community question answering systems". In: *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pp. 51–60, 2016.

[3] AROUSSI, S. A., EL HABIB, N., EL BEQQALI, O. "Improving question answering systems by using the explicit semantic analysis method". In: *Intelligent Systems: Theories and Applications (SITA), 2016 11th International Conference on*, pp. 1–6, 2016.

[4] BAEZAYATES, R., RIBEIRONETO, B., OTHERS, *Modern information retrieval*. vol. 463. ACM press New York, 1999.

[5] BANERJEE, A., BASU, S. "A social query model for decentralized search". In: *Proceedings of the 2nd Workshop on Social Network Mining and Analysiss. ACM, New York*, 2008.

[6] BERNERS-LEE, T. Weaving the web: The original design and ultimate destiny of the world wide web.[sl], 2000.

[7] BERNERS-LEE, T., HENDLER, J., LASSILA, O. "The semantic web", *Scientific american* v. 284, n. 5, pp. 34–43, 2001.

[8] BIZER, C., HEATH, T., BERNERS-LEE, T. "Linked data: The story so far". In: , pp. 205–227, IGI Global, 2011.

[9] BOGDANOVA, D., SANTOS, D., C., BARBOSA, L., et al. "Detecting semantically equivalent questions in online user forums". In: *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pp. 123–131, 2015.

[10] CAO, X., CONG, G., CUI, B., et al. "A generalized framework of exploring category information for question retrieval in community question answer archives". In: *Proceedings of the 19th international conference on World wide web*, pp. 201–210, 2010.

[11] CAO, X., CONG, G., CUI, B., et al. "Approaches to exploring category information for question retrieval in community question-answer archives", *ACM Transactions on Information Systems (TOIS)* v. 30, n. 2, p. 7, 2012.

[12] CHUA, A. Y., BANERJEE, S. "Answers or no answers: Studying question answerability in stack overflow", *Journal of Information Science* v. 41, n. 5, pp. 720–731, 2015.

[13] COSTA, R. D. "Por um novo conceito de comunidade: redes sociais, comunidades pessoais, inteligência coletiva", *Interface-Comunicação, Saúde, Educação* v. 9pp. 235–248, 2005.

[14] EFRON, M., WINGET, M. "Questions are content: a taxonomy of questions in a microblogging environment". In: *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem-Volume 47*, p. 27, 2010.

[15] ELLISON, N. B., GRAY, R., VITAK, J., et al. "Calling all facebook friends: Exploring requests for help on facebook.". In: *ICWSM*, 2013.

[16] GABRILOVICH, E., MARKOVITCH, S. "Computing semantic relatedness using wikipedia-based explicit semantic analysis.". In: *IJcAI*, pp. 1606–1611, 2007.

[17] HARPER, F. M., RABAN, D., RAFAELI, S., et al. "Predictors of answer quality in online q&a sites". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 865–874, 2008.

[18] HILTZ, S. R., TUROFF, M., *The network nation: Human communication via computer*. Mit Press, 1993.

[19] HILTZ, S. R., TUROFF, M., JOHNSON, K. "Experiments in group decision making, 3: Disinhibition, deindividuation, and group process in pen name and real name computer conferences", *Decision Support Systems* v. 5, n. 2, pp. 217–232, 1989.

[20] HOROWITZ, D., KAMVAR, S. D. "The anatomy of a large-scale social search engine". In: *Proceedings of the 19th international conference on World wide web*, pp. 431–440, 2010.

[21] HUBERMAN, B. A., ROMERO, D. M., WU, F. "Social networks that matter: Twitter under the microscope", *arXiv preprint arXiv:0812.1045*, , 2008.

[22] KATZ, L. "A new status index derived from sociometric analysis", *Psychometrika* v. 18, n. 1, pp. 39–43, 1953.

[23] KOLLOCK, P. "The economies ol online cooperation", *Communities in cyberspace* v. 220, 1999.

[24] LAKHANI, K. R., VON HIPPEL, E. "How open source software works:"free" user-to-user assistance". In: , pp. 303–339, Springer, 2004.

[25] MANORANJITHAM, G., VEERASELVI, S. "Mobile question and answer system based on social network", *International Journal of Advanced Research in Computer and Communication Engineering* v. 2pp. 3620–3624, 2013.

[26] MEDEIROS, J. F., SIQUEIRA, S. W. M., NUNES, B. P., et al. "Tagtheweb: Using wikipedia categories to automatically classify resources on the web". In: *Extended Semantic Web Conference*, Crete, 2018.

[27] MOLINO, P. "Semantic models for question answering", *Consortium*, p. 28, 2012.

[28] MORRIS, M. R., TEEVAN, J., PANOVICH, K. "What do people ask their social networks, and why?: a survey study of status message q&a behavior". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 1739–1748, 2010.

[29] MUI, Y. Q., WHORISKEY, P. "Facebook passes google as most popular site on the internet, two measures show", *The Washington Post*, , 2010.

[30] MUSSOI, E. M., FLORES, M. L. P., BEHAR, P. A. "Comunidades virtuais: um novo espaço de aprendizagem", *RENOTE: revista novas tecnologias na educação [recurso eletrônico]. Porto Alegre, RS*, , 2007.

[31] NUNES, B. P., KAWASE, R., FETAHU, B., et al. "Interlinking documents based on semantic graphs", *Procedia Computer Science* v. 22pp. 231–240, 2013.

[32] PODGORNY, I., GIELOW, C. Semi-automated prevention and curation of duplicate content in social support systems, 2018.

[33] PRATES, J. C., FRITZEN, E., SIQUEIRA, S. W., et al. "Contextual web searches in facebook using learning materials and discussion messages", *Computers in Human Behavior* v. 29, n. 2, pp. 386–394, 2013.

[34] PREECE, J. “Online communities: designing usability, supporting sociability”, *Industrial Management & Data Systems* v. 100, n. 9, pp. 459–460, 2000.

[35] PROCACI, T. B., SIQUEIRA, S. W., NUNES, B. P., et al. “Modelling experts behaviour in q&a communities to predict worthy discussions”. In: *Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference on*, pp. 291–295, 2017.

[36] RHEINGOLD, H. *Lisboa: Gradiva*, .

[37] RHEINGOLD, H., *The virtual community: Homesteading on the electronic frontier*. MIT press, 2000.

[38] SHADBOLT, N., BERNERS-LEE, T., HALL, W. “The semantic web revisited”, *IEEE intelligent systems* v. 21, n. 3, pp. 96–101, 2006.

[39] SHAH, C., KITZIE, V., CHOI, E. “Modalities, motivations, and materials–investigating traditional and social online q&a services”, *Journal of Information Science* v. 40, n. 5, pp. 669–687, 2014.

[40] SHEN, Y., RONG, W., SUN, Z., et al. “Question/answer matching for cqa system via combining lexical and sequential information.”. In: *AAAI*, pp. 275–281, 2015.

[41] SHTOK, A., DROR, G., MAAREK, Y., et al. “Learning from the past: answering new questions with past answers”. In: *Proceedings of the 21st international conference on World Wide Web*, pp. 759–768, 2012.

[42] SINGH, P., SHADBOLT, N. “Linked data in crowdsourcing purposive social network”. In: *Proceedings of the 22nd International Conference on World Wide Web*, pp. 913–918, 2013.

[43] SINGH, P., SIMPERL, E. “Using semantics to search answers for unanswered questions in q&a forums”. In: *Proceedings of the 25th International Conference Companion on World Wide Web*, pp. 699–706, 2016.

[44] SOUZA, C., MAGALHÃES, J., COSTA, E., et al. “Social query: a query routing system for twitter”. In: *Proc. 8th International Conference on Internet and Web Applications and Services (ICIW)*, pp. 147–153, 2013.

[45] SOUZA, C., REMíGIO, J., ARAGÃO, F., et al. “Investigating how"good"characteristics’ presence are related with questions’ performance: An empirical study on a programming community”. In: *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*, pp. 289–294, 2016.

[46] SOUZA, C. C. D. *Investigando como Características Desejáveis afetam o Desempenho de Perguntas sobre Programação em Sites de Perguntas e Respostas*. Universidade Federal de Campina Grande, UFCG, 2018.

[47] SRBA, I., BIELIKOVA, M. "Why is stack overflow failing? preserving sustainability in community question answering", *IEEE Software* v. 33, n. 4, pp. 80–89, 2016.

[48] SUROWIECKI, J., SILVERMAN, M. P., OTHERS. "The wisdom of crowds", *American Journal of Physics* v. 75, n. 2, pp. 190–192, 2007.

[49] TECHONOLOGY, I. I., ASSOCIATES), T. State of the internet 2000. Tech. rep., US Internet Council, 2000.

[50] MANSILLA, T. I., A., DE LA ROSA I ESTEVA, J. L. "Survey of social search from the perspectives of the village paradigm and online social networks", *Journal of Information Science* v. 39, n. 5, pp. 688–707, 2013.

[51] VASILESCU, B., CAPILUPPI, A., SEREBRENIK, A. "Gender, representation and online participation: A quantitative study of stackoverflow". In: *Social Informatics (SocialInformatics), 2012 International Conference on*, pp. 332–338, 2012.

[52] ZHANG, J., ACKERMAN, M. S., ADAMIC, L. "Expertise networks in online communities: structure and algorithms". In: *Proceedings of the 16th international conference on World Wide Web*, pp. 221–230, 2007.

[53] ZHANG, W. E., SHENG, Q. Z., LAU, J. H., et al. "Detecting duplicate posts in programming qa communities via latent semantics and association rules". In: *Proceedings of the 26th International Conference on World Wide Web*, pp. 1221–1229, 2017.

[54] ZHANG, W. E., SHENG, Q. Z., LAU, J. H., et al. "Duplicate detection in programming question answering communities", *ACM Transactions on Internet Technology (TOIT)* v. 18, n. 3, p. 37, 2018.

[55] ZHOU, G., LIU, Y., LIU, F., et al. "Improving question retrieval in community question answering using world knowledge.". In: *IJCAI*, pp. 2239–2245, 2013.

[56] ZHOU, T. C., LYU, M. R., KING, I. "A classification-based approach to question routing in community question answering". In: *Proceedings of the 21st International Conference on World Wide Web*, pp. 783–790, 2012.

[57] ZOLAKTAF, Z., RIAHI, F., SHAFIEI, M., et al. “Modeling community question-answering archives”. In: *Proceedings of the Workshop on Computational Social Science and the Wisdom of Crowds at NIPS*, 2011.