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

Padrões Comportamentais e Feature Learning para Analisar o Prestígio de Usuários em Comunidades Online de Perguntas e Respostas

Thiago Baesso Procaci

Orientadores
Sean Wolfgand Matsui Siqueira
Bernardo Pereira Nunes

RIO DE JANEIRO, RJ - BRASIL
MAIO de 2019

Padrões Comportamentais e Feature Learning para Analisar o Prestígio de Usuários em Comunidades Online de Perguntas e Respostas

Thiago Baesso Procaci

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

Aprovada por:

Sean Wolfgand Matsui Siqueira - UNIRIO

Bernardo Pereira Nunes - PUC-Rio

Mariano Pimentel - UNIRIO

Adriana Cesário de Faria Alvim - UNIRIO

Jonice de Oliveira Sampaio - UFRJ

Claudia Lage Rebello da Motta - UFRJ

RIO DE JANEIRO, RJ - BRASIL

MAIO de 2019

Catalogação informatizada pelo(a) autor(a)

B963 Baesso Procaci, Thiago
Padrões Comportamentais e Feature Learning para Analisar o Prestígio de Usuários em Comunidades Online de Perguntas e Respostas / Thiago Baesso Procaci. -- Rio de Janeiro, 2019.
176

Orientador: Sean Wolfgand Matsui Siqueira.
Coorientador: Bernardo Pereira Nunes.
Tese (Doutorado) - Universidade Federal do Estado do Rio de Janeiro, Programa de Pós-Graduação em Informática, 2019.

1. Comunidades online. 2. Análise de Redes Sociais. 3. Feature Learning. 4. Machine Learning. 5. Prestígio Social. I. Wolfgand Matsui Siqueira, Sean, orient. II. Pereira Nunes, Bernardo, coorient. III. Título.

Aos meus avós Henrique e Maria dedico este trabalho. Pessoas inesquecíveis que partiram durante meu curso de doutorado. Aprendi com eles: ao contrário do desespero, a serenidade ajuda a lidar com as adversidades.

Agradecimentos

Agradeço ao professor Sean pelo trabalho de orientação. Sua generosidade para ensinar é imensa. Foi muito bom te conhecer e trabalhar todo esse tempo em conjunto.

Agradeço também ao professor Bernardo pelo trabalho de coorientação e ajuda nas escritas dos artigos.

Agradeço ao grupo SaL da UNIRIO, pelas excelentes discussões que me fizeram pensar.

Agradeço aos professores que compõem a banca desta tese pelo trabalho de avaliação.

Agradeço à minha esposa Gabriela pelo constante apoio durante todo o doutorado.

Agradeço aos professores do PPGI da UNIRIO pelas disciplinas lecionadas, fundamentais para a conclusão deste trabalho.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) pelo suporte fornecido através do financiamento (código 001).

Agradeço à FAPERJ pelo apoio nas publicações através do projeto E-26-102.256/2013 - JCNE: Associa: Explorando um Ambiente Semântico e Social de Ensino-Aprendizagem.

Agradeço ao CNPQ por também apoiar nas publicações através do projeto 312039/2015-8 – Bolsa DT: Integrando Práticas Pedagógicas e Métodos e Ferramentas de Análise de Dados Educacionais.

Por fim, agradeço a todos os funcionários do PPGI da UNIRIO, por sempre ajudarem os alunos no que for necessário.

Procaci, Thiago Baesso. **Padrões Comportamentais e Feature Learning para Analisar o Prestígio de Usuários em Comunidades Online de Perguntas e Respostas**. UNIRIO, 2019. 158 páginas. Tese de Doutorado. Departamento de Informática Aplicada, UNIRIO.

RESUMO

As comunidades online de perguntas e respostas desempenham um papel importante no contexto da apredizagem informal. Diversas pessoas utilizam estes ambientes colaborativos visando suprir alguma lacuna de conhecimento. Em várias delas, seus membros postam perguntas e podem receber *feedbacks* confiáveis. Este processo de postagem de perguntas e espera por respostas é denominado *social query*. Perguntas postadas em comunidades podem ser esclarecidas pelos mais diversos usuários, porém, em alguns casos, podem exigir um *feedback* especializado. Assim, surge a necessidade de detecção dos tipos de usuários nestas comunidades, com a finalidade de colocar em contato pessoas com dúvidas com outras capacitadas e disponíveis para ajudar. Há dois problemas envolvidos. O primeiro tem relação com a detecção de tipos de usuários, isto é, por meio de algum critério, devemos saber quais são aqueles que têm um grau de competência adequado para esclarecer determinadas questões. O segundo está relacionado com a recomendação de conexões entre questões e pessoas competentes para resolvê-las. O foco desta tese será o primeiro problema. Em síntese, este trabalho visa contribuir na solução do problema de identificação de tipos de usuários, propondo duas abordagens:

(i) uma baseada em análises multi-perspectiva; (ii) e outra baseada *feature learning*. Através das análises multi-perspectivas, elucidou-se comportamentos e características importantes de membros que se destacam em comunidades online, tornando mais fácil a identificação destes. Já as abordagens de *feature learning* se mostraram excelentes alternativas para a solução do problema, superando abordagens similares do estado da arte.

Palavras-chave: Comunidades online, Análise de Redes Sociais, Feature Learning, Machine Learning.

ABSTRACT

Online Q&A communities have emerged as important contributors to informal education. Members can post questions at any time with realistic expectations of receiving reliable feedback. This process is known as Social Query. This collaborative and informal learning environment allows participants to benefit from collective knowledge. Questions asking for recommendations or contextualized questions are adequately answered by eliciting the knowledge of the crowd, but others require the feedback, or specialist knowledge of a domain-expert. In these cases, someone may prefer to find a person who has related and notable experience (e.g. outstanding user) and Q&A communities have emerged as one of the most important places for people to seek advice or help. Consequently, the detection of the types of users in these communities arises in order to promote pairings and creating contact between people who want to ask questions with others available, who have the skill needed to solve the problem. There are two problems involved. The first is related to the detection of the types of users. It means, considering some criterias, we must know which are those who have an adequate degree of competence to clarify certain questions. The second is related to the recommendation of connections between questions and qualified people to solve it. The focus of this thesis is the first problem. In summary, this thesis aims at creating solutions for the problem of the identification of types of users, proposing two approaches: (i) one based on the multi-perspective analysis; (ii) the other, based on feature learning. Through the multi-perspectives analysis, important behaviors and characteristics of members who excel in online communities have been elucidated, making it easier to identify them.

The feature learning approaches have proved to be excellent alternatives for solving the problem, overcoming similar approaches of the state of the art.

Keywords: Online Communities, Social Network Analysis, Feature Learning, Machine Learning.

Sumário

Lista de Figuras

Lista de Tabelas

1. Apresentação da Pesquisa

Neste Capítulo serão introduzidos os elementos que motivaram este trabalho. Além disso, o Capítulo objetiva apresentar o problema de pesquisa, as questões que serão investigadas, a solução proposta, as contribuições e, por fim, a metodologia adotada.

1.1 Introdução

A Internet vem provocando uma revisão em paradigmas de interação e colaboração. Os computadores conectados em rede, desenvolvidos a partir da metade do século XX, se disseminaram por todo o sistema social e, desde então, vêm provocando profundas transformações na vida contemporânea [40]. Com a criação da Web na década de 1990 e, posteriormente, a ampliação de suas capacidades, se estabeleceu um novo lugar para interações humanas denominado espaço digital. Tal espaço é também conhecido como ciberespaço que, por sua vez, possibilitou ampliar as interações e experiências humanas dando origem a cibercultura, ou seja, a cultura do ciberespaço. [75].

O fácil acesso à tecnologia nos últimos anos como, por exemplo, ao computador pessoal, aos dispositívos móveis e à Internet, promoveu impactos significativos no modo de pensar, de comunicar e até mesmo de viver das pessoas [160]. Muitas atividades que antes eram realizadas sem o apoio

direto de sistemas computacionais, hoje são realizadas com tal suporte e, muitas vezes, substituem completamente as tecnologias predecessoras não digitais [94, 34]. Como consequência, novos hábitos e atividades surgiram. Aparentemente, pessoas escrevem menos cartas, compram menos discos de músicas e, até mesmo, interagem menos pessoalmente. Em vez disto, pessoas passaram a enviar mensagens via aplicativos de celular, escutar rádios online, assistir vídeos na Web, conversarem em sistemas online de bate-papo etc.

O universo online possibilitou a ampliação das relações humanas, ou seja, a criação de novos espaços de socialização e, consequentemente, novas possibilidades de interação. O próprio conceito de redes sociais que tradicionalmente tinha como objeto de estudo as relações entre pessoas no mundo real, atualmente se encontra completamente acoplado ao ciberespaço. Várias questões sociais que enfrentamos na vida real se encontram no ciberespaço como, por exemplo, competições para ser reconhecido como influente ou relevante [117], discursos de ódio [96], brincadeiras de mau gosto [74] ou mesmo, olhando sob uma perspectiva positiva, a possibilidade de se aprender com o outro [44] etc.

As redes sociais, tanto tradicionais quanto online, são meios que possibilitam interações entre pessoas, trocas de experiências e propagação de ideias. Entretanto, a capilaridade e a rapidez com que ideias e informações são difundidas no universo online parece ser bem maior [4]. Ademais, pessoas nas redes sociais online deixam 'rastros', isto é, produzem dados durante suas interações que, por sua vez, oportunizam o estudo de padrões de comportamentos humanos [122]. Por exemplo, voltando às eleições de 2004 nos Estados Unidos, através da análise de mensagens de usuários postadas em blogs, foi possível perceber a formação de bolhas de opiniões liberais e conservadoras, como relatado no trabalho clássico de Lada Adamic e Natalie Glance: *'The*

Political Blogosphere and the 2004 U.S. Election: Divided They Blog' [1]. Este fenômeno, que hoje é bem conhecido, na época era menos perceptível a olho nu e as análises destes 'rastros' sociais o deixou em evidência.

Dentre outras possibilidades de estudos de relações sociais no ciberespaço através de tais 'rastros', há diversas pesquisas que focam na observação de usuários que se destacam em algum círculo social online [72], procurando entender a diferença deles com relação aos demais. Geralmente, tais usuários de destaque constituem uma parcela minoritária, quando comparados a quantidade total de pessoas em uma rede social. A motivação destes estudos residem na importância destes usuários na formação de opiniões [71], difusão de informações e definição de tendências [148], além de, muitas vezes, serem considerados como fontes de informações confiáveis [116] ou como importantes atores para dificultar a disseminação de conteúdos não adequados [50].

Há redes sociais online atuais que se preocupam com distinção entre tipos de usuários. Nos sites de recrutamento, como o LinkedIn[1], conseguir identificar automaticamente pessoas que tenham notável destaque em determinada competência pode significar agilidade nas negociações e contratações [30]. Nas comunidades online voltadas para a aprendizagem informal, tais como o Stack Overflow[2] e o Quora[3], os usuários de destaque exercem um papel fundamental para manter a qualidade das postagens e também o ritmo de crescimento da comunidade [117, 113]. Na área de *Learning Analytics*, em especial os estudos que focam nos *Massive Open Online Courses* (MOOCs) ou nos ambientes de ensino à distância, é primordial entender o que distingue o aluno com bom desempenho com relação aos demais [127, 133]. Desta forma,

[1] `https://www.linkedin.com/`
[2] `https://stackoverflow.com/`
[3] `https://www.quora.com`

entender os comportamentos promissores quanto a aprendizagem, isto é, os dos estudantes com bom desempenho, pode servir para saber previamente, por exemplo, uma provável evasão ou reprovação daqueles que divergem de tais comportamentos. Ou seja, prever se um estudante está com dificuldades, constitui uma oportunidade para o professor intervir. Ainda que os MOOCs e os ambientes de ensino a distância não sejam comumente vistos como uma rede social online habitual, é inegável que se possa estabelecer relações entre pessoas em tais lugares.

Através dos exemplos citados, é possível perceber que a definição do termo 'usuários de destaque' parece ser elástico, pois, pode significar alguém influente, um bom aluno, uma pessoa confiável, um profissional com o perfil que um recrutador busca etc. O que seria um 'usuário de destaque'? Na verdade, a definição do termo depende do contexto analisado. Se queremos entender o processo de aprendizagem em uma rede social online, o termo estará intimamente ligado àqueles que melhor demonstram o conhecimento naquele momento. Se desejarmos combater *Fake News*, o termo provavelmente estará conectado aos que mais desmentem boatos e apresentam argumentos devidamente fundamentados. Em síntese, cabe ao pesquisador interpretar o que seria o termo destaque dados suas questões e objetivos de pesquisa, assim como, seus objetos de estudo que, no contexto exposto, seriam as redes sociais online e seus participantes.

Esta tese tem como objetivo principal ampliar as formas de se encontrar os usuários que se destacam nas comunidade online de perguntas e respostas. Um estudo detalhado sobre características e comportamentos de usuários, sob diversas perspectivas, será apresentado. Serão também discutidas formas automáticas para caracterizar tais usuários, bem como classificá-los. Ademais, neste trabalho, definimos usuários de destaque como aqueles que deram con-

tribuições relevantes para a comunidade sendo isto reconhecido pelos demais (os outros usuários).

Nas Seções subsequentes deste Capítulo serão apresentadas com detalhes as motivações para se estudar as comunidades de perguntas e respostas, bem como, uma melhor definição do problema que será tratado. Ainda neste Capítulo, o escopo, a metodologia e a organização do trabalho também serão mostradas, de forma a dar uma visão ampla da tese para os leitores.

1.2 Motivação

Por qual razão se deve estudar as comunidades online e seus participantes? Na Seção 1.1, foi introduzido o tema da pesquisa, mostrando sua relevância, através de algumas de suas possibilidades de estudo e aplicação. Neste momento, serão observados os fatores que motivaram a realização deste trabalho. Assim, se dividiu os fatores motivacionais nas seguintes categorias:

1. Motivação pessoal, por mostrar as conexões desta pesquisa com a trajetória de seu autor;

2. Redes sociais e comunidades online, pela popularidade e também serem fonte de dados para as análises deste trabalho;

3. Colaboração, por ser um fator importante no processo de aprendizagem e muito presente em comunidades online de perguntas e resposta.

Tais fatores são apresentados em detalhes a seguir e, por sua vez, reforçam a relevância deste trabalho.

1.2.1 Motivação Pessoal

O que me levou a estudar comunidades de perguntas e respostas e seus participantes? Desde que ingressei no Programa de Pós-Graduação em Infor-

mática da UNIRIO, participo do grupo de pesquisa Semantics and Learning, onde tive a oportunidade de trocar experiências com os colegas. Assim, acabei me interessando pelas análises de redes sociais, tema bem discutido naquela época e ainda hoje. Fui apresentado ao tema através do grupo e logo vi sua interseção com a Teoria dos Grafos, assunto que me fascina desde a graduação em Ciência da Computação.

Desde então, aprofundei meus estudos em redes sociais online buscando entender suas particularidades como um ecossistema vivo, em constante transformação. As análises de redes sociais funcionaram como uma lupa que me permitiram visualizar relações que a olho nu eram imperceptíveis. Em paralelo, iniciei também estudos de modelos preditivos de Inteligência Artificial (*Machine Learning*) mesclado com conceitos de análises de redes sociais. Esses estudos deram origem a um conjunto de pesquisas quantitativas, voltadas para comunidades online destinadas à aprendizagem informal. A escolha por se fazer pesquisas em comunidades online para a aprendizagem se deve ao foco do grupo Semantics and Learning que, sempre quando possível, busca utilizar técnicas da Computação aplicadas na Educação (mesmo que sejam voltadas para o contexto informal). Assim, diversas relações ocultas pertinentes à aprendizagem online foram elucidadas através destas pesquisas.

Alguns trabalhos relacionados foram marcantes [165, 74, 26, 60]. A leitura destes me estimulou a aprofundar no tema, de forma que, pude amadurecer ideias para conduzir uma pesquisa científica. Acima de tudo, mergulhei no universo das comunidades online de perguntas e respostas sendo, inclusive, um participante ativo[4] no Stack Overflow. O fato é que a pesquisa me faz pensar em modelos, em abstrações focada no entendimento de um fato enquanto a participação real em um círculo social me aproxima das pessoas, da

[4]`https://stackoverflow.com/users/5854925/thiago-procaci`

realidade simplesmente como ela é (às vezes opaca, sem o completo entendimento). Esta oscilação entre a pesquisa teórica e a imersão na realidade foram fundamentais para a pesquisa chegar no estágio atual deste trabalho e isto, sem dúvidas, foi um fator motivador.

1.2.2 Redes Sociais e Comunidades Online

Esta Seção tem como finalidade mostrar dados que evidenciam o sucesso das redes sociais e das comunidades online, sob a perspectiva da popularidade. Este sucesso é de tamanha relevância que vários grupos de pesquisas utilizam tais ambientes como objeto de estudo. Isto, por sua vez, constitui um fator motivacional importante para este trabalho. É importante salientar que redes sociais online e comunidades online são conceitos distintos, apesar de estarem relacionados. Redes sociais online são espaços digitais onde pessoas simplesmente podem interagir entre si. Já em comunidades online, há também interação entre pessoas, contudo, geralmente, tais pessoas compartilham interesses comuns [48]. Um fato interessante é que algumas redes sociais permitem a criação de comunidades como, por exemplo, os grupos (comunidades) do Facebook (rede social online).

1.2.2.1 Popularidade das Redes Sociais e Comunidades

O Facebook é a maior plataforma de rede social online do mundo. Como mostrado em seus relatórios do segundo trimestre de 2016 [45], o número médio de usuários ativos mensalmente do Facebook foi de 1,71 bilhão. Além disso, este mesmo número tem crescido 15% ano a ano. Similarmente, no Twitter, o número médio de usuário ativos por mês foi de 310 milhões no segundo trimestre de 2016 [150]. Em um contexto diferente, o LinkedIn[5], que é uma plataforma de rede social online para contato profissional, contem

[5] `https://www.linkedin.com/`

mais de 450 milhões de membros [77]. De acordo com o Alexa Internet Global Ranking[6], em setembro de 2016, o LinkedIn foi o décimo quarto site Web mais acessado do mundo.

Contudo, apesar das pessoas utilizarem as plataformas de redes sociais online por diversos motivos (entretenimento por exemplo), existem pessoas que preferem usá-las para aprender ou se manter informado [118, 144, 44]. A Universidade Federal do Estado do Rio de Janeiro (UNIRIO), por exemplo, usa grupos do Facebook para apoiar a aprendizagem. Como é o caso do grupo de pesquisa Semantics and Learning[7] da UNIRIO que usa um grupo do Facebook para compartilhar informações sobre os temas de suas pesquisas. Foi também o caso das disciplinas Cibercultura[8] e Docência em Sistemas de Informação[9] da UNIRIO.

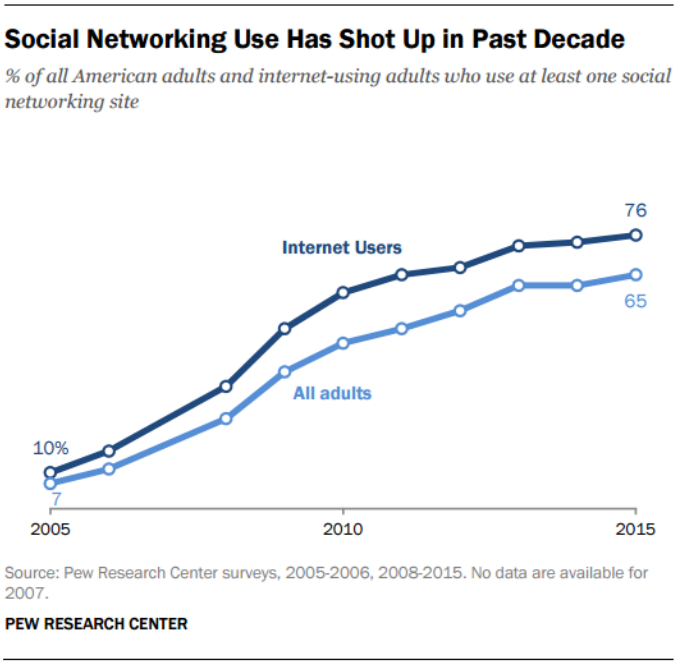
A plataforma Stackexchange Question and Answer (Q&A)[10] consiste em um conjunto de comunidades online de perguntas e respostas onde pessoas podem pedir ou oferecer ajuda. A plataforma é uma grande rede com mais de 160 comunidades sobre diversos domínios de conhecimento. O Stack Overflow[11] é uma das comunidades que pertence ao Stackexchange, onde os assuntos discutidos tem relação com programação de computadores. Essa comunidade tem por volta de 4.7 milhões participantes. A dinâmica básica desta comunidade é bem simples: alguém posta uma pergunta e outra pessoa responde. Existem muitas outras comunidades no Stackexchange, todas com a mesma dinâmica, e cada uma delas especializada em um assunto diferente como biologia, física, matemática, entre outros [119]. Em síntese, a plata-

[6] `http://www.alexa.com/siteinfo/linkedin.com`
[7] `http://bit.ly/2pfPrxr`
[8] `https://www.facebook.com/groups/384634338233143/`
[9] `https://www.facebook.com/groups/950622775034682/`
[10] `http://stackexchange.com/`
[11] `http://stackoverflow.com/`

forma Stackexchange Q&A deu origem a uma grande rede de colaboração onde pessoas podem aprender de maneira informal, através da coautoria de conhecimentos.

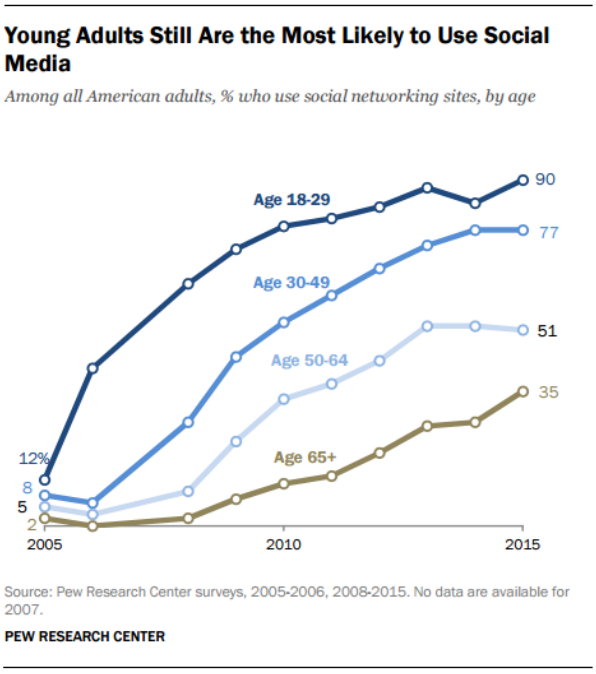
O Pew Research Center é uma instituição que fornece informações sobre questões, atitudes e tendências que estão moldando os Estados Unidos e o mundo. O Pew Research Center afirma que, em 2015, 65% dos adultos estadunidenses usaram pelo menos uma plataforma de rede social online como mostrado na Figura 1.1 [102]. Além disso, a idade é fortemente correlacionada com o uso de mídia social. Quanto mais jovem, maior o uso de mídias sociais como mostrado na Figura 1.2. No entanto, o uso entre os mais velhos tem aumentado significativamente ao longo do tempo.

Figura 1.1: Pew Research Center - Uso de Redes Sociais



Dados os fatos apresentados, é possível notar que as redes sociais e comunidades da Internet estão crescendo e se tornando mais populares ainda. Tais fatos evidenciam que as redes estão muito presente na vida das pessoas.

Figura 1.2: Pew Research Center - Mídias Sociais e Idade



Ademais, as redes sociais podem impactar positivamente ou negativamente a vida das pessoas, como já relatado em [47] e [63]. Foi também mostrados que existem pessoas que usam tais plataformas para aprender. Este impacto e a possibilidade de aprender nesses ambientes são fatores motivacionais significativos para se fazer pesquisas sobre redes sociais e comunidades online, como é o caso deste trabalho.

1.2.2.2 Pesquisa em Redes Sociais e Comunidades

Uma vez que as redes sociais e as comunidades online se tornaram populares, diversos grupos de cientistas desenvolveram pesquisas sobre elas.

A empresa Facebook tem um departamento de pesquisa[12] que objetiva estudar sua própria plataforma de rede social e extrair conhecimentos dela. A missão do Facebook, de acordo com o seu site, é dar às pessoas o poder

[12] `https://research.facebook.com/about/`

para compartilhar e fazer do mundo um lugar mais aberto e conectado. No entanto, para alcançar este objetivo, constante inovação é necessária. Tendo em vista este cenário, o Facebook está sempre desenvolvendo pesquisas visando produzir conhecimentos em diversas áreas[13] como ciência de dados [65], economia [11], aprendizado de máquina [84], interface humano-computador [29] etc. Cada pesquisa e descoberta relevante é uma oportunidade para melhorar sua própria plataforma.

O MIT (Massachusetts Institute of Technology) Media Lab[14] é uma organização na qual um dos seus grupos de pesquisa, cujo nome é Macro Connections, tem como foco estudar redes sociais. De acordo com o seu Web site[15], o grupo foca no desenvolvimento de ferramentas analíticas que podem ajudar e melhorar o entendimento sobre as macro estruturas do mundo, considerando todas as suas complexidades. O grupo Macro Connections já conduziu diversas pesquisas sobre redes sociais, incluíndo temas como meritocracia [25], língua e cultura [128], visualização de dados [142], memória coletiva [70] etc.

De forma similar, a School of Information[16] e o Center for the Study of Complex Systems[17], ambos da Universidade de Michigan, realizaram várias pesquisas sobre análise de redes sociais. Por exemplo, tais instituições pesquisaram sobre os conteúdos produzidos em redes socias relacionados à partidos políticos e candidatos [80], difusão da informação [15] e a dinânica de atividades econômicas [16].

A Microsoft Research[18] é outro grupo que conduz pesquisas em redes sociais. Estudos relacionados a aprendizagem colaborativa [83] e influência [14]

[13] `https://research.facebook.com/publications/`
[14] `https://www.media.mit.edu/`
[15] `https://www.media.mit.edu/research/groups/macro-connections`
[16] `https://www.si.umich.edu/`
[17] `https://lsa.umich.edu/cscs`
[18] `https://www.microsoft.com/en-us/research/`

em redes sociais já foram feitas neste grupo. O Center for Complex Network Research[19] da Universidade de Northeastern tem aplicado conceitos de análises de redes sociais em problemas da área de bioinformática como a criação de modelos para descrever uma doença complexa [58] ou a identificação de genes relacionado a determinadas doenças [153]. Tais estudos demonstram que as técnicas de análise de redes sociais podem ser aplicadas em outros contextos, sem aparente conexão com trabalhos usualmente realizados na área.

Além disso, diversos pesquisadores conduzem pesquisas relacionando redes sociais e educação, como é o caso do grupo de pesquisa Semantics and Learning (SaL)[20] da Universidade Federal do Estado do Rio de Janeiro. No SaL é possível encontrar trabalhos que visam encontrar pessoas dispostas a ajudar o outro a aprender em redes sociais e comunidades online [118, 117, 110, 119]. É também possível encontrar no SaL trabalhos que fazem o enriquecimento semânticos de discussões em redes sociais utilizadas para a educação [54, 107]. Tais trabalhos demonstram que o enriquecimento pode melhor contextualizar discussões e impactar positivamente a aprendizagem. Há ainda trabalhos que buscam encontrar especialistas em tópicos de comunidades online [113, 109] no grupo SaL.

Um outro recurso importante muito presente em comunidades online e redes sociais é o bate-papo. Atento à popularidade e relevância deste recurso, o grupo de pesquisa ComunicaTEC [103] da Unversidade Federal do Estado do Rio de Janeiro vem investigando o uso de sistemas de bate-papo voltado para a Educação. Assim, este vem estudando as potencialidades deste recurso através da caracterização de sessões de bate-papo [125] entendendo, por exemplo, qual é a quantidade ideal de participantes em uma sessão de

[19] `http://www.barabasilab.com/`
[20] `http://bit.ly/2pfPrxr`

forma que todos consigam efetivamente acompanhar as discussões [126].

Concluindo, existem diversos grupos importantes fazendo pesquisas em redes sociais ou em recursos presentes nestas. Além disso, muitos destes grupos estão em plena atividades. Este fato corrobora a importância para se conduzir pesquisas em redes sociais e comunidades online que, por sua vez, é também um fator motivacional para esta pesquisa.

1.2.3 Colaboração

A relação entre colaboração e aprendizagem é também um fator motivador deste trabalho. Esta Seção tem como objetivo mostrar brevemente a importância da colaboração sob a perspectiva de pensamentos pedagógicos, bem como, a interseção de tais pensamentos com as comunidades online.

1.2.3.1 Relação com Aprendizagem

Durante anos, prevaleceu a aprendizagem centrada no professor como polo de transmissão do conhecimento. Sob a pespectiva dos críticos da pedagogia da transmissão [51, 75], ensinar não é transmitir conhecimento, mas criar as possibilidades para sua própria produção ou construção. Parafraseando Paulo Freire, a educação verdadeira não se faz de A para B ou de A sobre B, mas de A com B [51]. A cultura da transmissão perde terreno quando, culturalmente, emerge a valorização das interações e da interatividade [129] que, por sua vez, ficou evidenciado na Seção 1.2.2.1. Estes pensamentos estão em plena sintonia com a ideia de colaboração, interação e diálogo que as comunidades online com seus membros participantes podem promover.

Entende-se atualmente que conhecimento não é um produto fixo e acabado, ele é construído em um contexto de trocas, mediante um tensionamento constante entre as certezas atuais e as dúvidas que recaem sobre essas certezas, conduzindo ao estabelecimento de novas relações ou conhecimento

[46]. Há indicações que se aprende muito através das interações entre as pessoas, por exemplo, resolvendo problemas em conjunto, obtendo explicações sobre problemas já resolvidos, explicando soluções, debatendo sobre vantagens e desvantagens de determinadas escolhas, fazendo ou recebendo críticas, construindo sínteses coletivas, dentre outras atividades realizadas colaborativamente [35].

A aprendizagem colaborativa tem sido defendida por educadores nos diversos níveis escolares, do ensino fundamental à pós-graduação, e também no contexto informal [86]. Esta prática não é uma novidade e a disponibilidade das tecnologias de comunicação e de interação social tem contribuído para adesão de novos interessados [33]. Ademais, os benefícios da colaboração são vários, dentre os quais cita-se: a preparação para a vida em sociedade, o desenvolvimento do pensamento crítico e sofisticado e a competência para resolver problemas de grande porte a partir de contribuições individuais [35].

O advento da Web 2.0[21] e o seu uso, para a realização de atividades colaborativas no trabalho e no lazer, fez despertar nas pessoas o interesse pela incorporação dessas práticas nas atividades de aprendizagem, o que reforça a demanda por práticas colaborativas. A popularização das mídias sociais como blogs, wiki e redes sociais, marcou um novo direcionamento para a geração de tecnologias Web, onde o foco central é a comunicação entre pessoas, a troca de experiências, o compartilhamento e a coautoria de conteúdos [7].

Um estudo envolvendo 30.616 estudantes universitários dos Estados Unidos, constatou que 90,3% dos alunos dedicam parte de seu tempo, diariamente, na utilização de plataformas de redes sociais online [143] que, por sua

[21]Termo para designar uma segunda geração de comunidades e serviços, tendo como conceito a "Web como plataforma", envolvendo wikis, aplicativos baseados em folksonomia, redes sociais e tecnologia da informação.

vez, são importantes ambientes de colaboração [145]. Devido a este caráter colaborativo das redes sociais, que possibilita a junção de pessoas para troca de experiências através de recursos computacionais ricos e cooperativos, elas têm sido vistas como uma tendência para impulsionar transformações nos paradigmas educacionais e na prática da formação à distância ao longo da vida [59].

Desta forma, diante da argumentação exposta que mostra a importância da colaboração para a construção do conhecimento, entende-se como relevante realizar estudos em comunidades online utilizadas para fins de aprendizagem sendo, portanto, um fator motivacional para este trabalho.

1.2.3.2 Social Query

Os ambientes online podem ser bons lugares para se aprender. O processo de postagem de perguntas em um ambiente online e espera por respostas é conhecido como *social query* [144, 90, 18]. *Social query* pode ser vista como uma alternativa aos motores de busca da Web. Alguns trabalhos [68, 91] afirmam que ambientes que permitem a formação de comunidades online com muitos usuários (milhares de usuários no mínimo), como o Twitter e o Facebook, são lugares bons e eficientes para encontrar informações através do uso de *social query*. Isso se deve à presença de muitos usuários que, por sua vez, aumentam as chances de se receber algum tipo de informação ou resposta.

Motores de busca nem sempre são as melhores formas para buscar informações na Web, pois, seus resultados são muitas vezes indesejados ou incompletos, ou seja, não necessariamente refletem o que se busca em um determinado momento [107]. De certa forma, os motores de busca deixam a desejar quando se procura por algo mais contextualizado. Por exemplo, ao se fazer uma busca na Web por uma palavra como “Flamengo”, pode-se

querer obter resultados sobre um time de futebol, um bairro da cidade do Rio de Janeiro, um trecho da música do cantor Djavan ou uma região ao norte da Bélgica [107]. Além disso, alguns problemas são melhores resolvidos por pessoas como, por exemplo, perguntas muito contextualizadas, pedidos de recomendação, pedidos de opiniões, conselhos etc [67]. Assim, uma alternativa aos motores de busca para a resolução de problemas ou dúvidas são, por exemplo, as comunidades online de perguntas e respostas como Stack Overflow, Quora[22] e Yahoo! Answers[23], onde os usuários perguntam e respondem de forma voluntária e colaborativa. Contudo, existem pessoas que preferem postar perguntas para pessoas que pertencem somente ao seu círculo de amizades em vez de postar para pessoas desconhecidas em comunidades de perguntas e respostas [90].

Há resultados confirmando que *social query* é um método viável para se obter respostas em um ambiente online [89]. Tais resultados foram encontrados em um estudo realizado internamente na Microsoft, utilizando suas próprias ferramentas de comunicação. Foi concluído que 93,5% dos usuários tiveram suas perguntas respondidas e, em 90,1% dos casos, os usuários obtiveram respostas em menos de um dia. Estudos similares no Twitter foram feitos [100], porém, com resultados diferentes, pois concluíram que somente 18,7% das perguntas postadas por um usuário do Twitter recebiam respostas. Foi concluído também que o número de respostas recebidas por um usuário tem uma correlação positiva com o seu número de seguidores. Além disso, 67% das perguntas respondidas no Twitter obtinham respostas de modo relativamente rápido (em menos de 30 minutos).

Assim, considerando o apresentado, constata-se que os ambientes online, que permitem a execução de social queries, podem ser lugares eficientes para

[22] `https://www.quora.com/`
[23] `https://answers.yahoo.com/`

a se obter soluções para problemas de forma colaborativa. Este fato é algo que motiva este trabalho, uma vez que, corrobora que é possível aprender em tais lugares.

1.3 Problema

Devido às crescentes demandas por conhecimento dentro das organizações e uma disponibilidade limitada de recursos e competências para suprir tais demandas, muitos profissionais, tanto da indústria quanto da academia, acabam buscando por conhecimento em fontes externas para resolver os seus problemas [157]. Essas fontes externas são muitas vezes os motores de busca da Web, sites ou mesmo comunidades online onde pessoas visam encontrar soluções para seus problemas diários. Em síntese, tais demandas podem impulsionar o uso de *social query*.

Embora existam vantagens no uso de *social query*, ela também tem algumas limitações. Quando uma pergunta é postada em uma comunidade, alguns resultados não esperados podem ser encontrados tais como: receber respostas erradas ou descontextualizadas [144]; continuar recebendo respostas mesmo depois do problema ter sido resolvido [114]; e nunca receber uma resposta, em especial nas comunidades que priorizam a visualização das postagens mais recentes [90].

Em um estudo de *social query* no Twitter [100], ficou demonstrado que a rede social sofre o 'efeito da linha do tempo'. Segundo esse estudo, utilizando o Twitter é possível obter respostas para uma pergunta em um tempo relativamente rápido, porém, a maior parte das perguntas não são respondidas. Uma das explicações da baixa porcentagem de respostas recebidas deve-se ao fato do Twitter priorizar a visualização de postagens mais recentes. Logo, é provável que alguns seguidores (usuários) nem fiquem sabendo da existên-

cia de uma determinada pergunta. Na comunidade de perguntas e respostas Biology Q&A[24], cerca de 22% das perguntas não recebem respostas e na comunidade Chemistry Q&A[25], a parcela de não respondidas é de 24% [114]. Estes números evidenciam algumas das limitações da *social query*.

O fato é que existem perguntas com níveis distintos de dificuldade e competências diferentes para resolvê-las. Em alguns casos, elas podem ser resolvidas pela sabedoria das massas [155], possivelmente por serem mais fáceis, e em outros podem exigir um *feedback* especializado [113]. Neste segundo caso, há mais possibilidades de se receber respostas erradas ou mesmo não receber resposta. Para minimizar isto, alguém pode preferir encontrar uma pessoa com notável experiência no assunto discutido, isto é, um usuário de destaque, objetivando receber uma resposta precisa. Assim, surge a necessidade de detecção dos tipos de usuários nestas comunidades, com a finalidade de promover encontros, ou seja, colocar em contato pessoas com dúvidas com outras capacitadas e disponíveis para ajudar [165].

Como se percebe, há dois problemas envolvidos. **O primeiro tem relação com a detecção de tipos de usuários, isto é, por meio de algum critério, devemos saber quais são aqueles que têm um grau de competência adequado para esclarecer determinadas questões (no inglês, este problema é conhecido como *expertise finding*). O segundo está relacionado com a recomendação de conexões entre questões e pessoas competentes para resolvê-las.** Desta forma, a detecção dos tipos de usuarios é um pré-requisito do problema de recomendação. **Visando limitar o escopo deste trabalho, este irá focar em propostas para a solução do primeiro problema.**

[24] `https://biology.stackexchange.com/`
[25] `https://chemistry.stackexchange.com/`

1.4 Solução

Considerando o primeiro problema apresentado (Seção 1.3), é comum, nos trabalhos anteriores em comunidades online, soluções que foquem na dicotomia entre usuários que buscam e que fornecem ajuda. Para a identificação destes dois tipos de usuários, em diversas oportunidades, são utilizadas técnicas de recuperação de informação [17] ou de análise de redes sociais [165]. Ademais, é comum se referirem aos competentes fornecedores de ajuda como usuários de destaque, excepcionais, especialistas ou notáveis. Desta forma, pesquisas são realizadas para a descoberta de características que diferenciem os usuários que se destacam dos demais. Neste trabalho, o foco também será nesta dicotomia em comunidades online de perguntas e respostas.

A caracterização para a identificação do tipo de usuário é também conhecida como *feature engineering* ou *hand-engineering* [42] que, em síntese, se trata de um processo de observação que objetiva deixar em evidência as particularidades dos usuários. No entanto, em boa parte dos trabalhos relacionados a caracterização dos usuários de destaque consideram uma única perspectiva comportamental associada a eles. Neste trabalho, visamos analisar (através de um estudo empírico) em detalhes os comportamentos dos usuários, considerando diversas perspectivas tais como participação, traços de linguagem, laços sociais, influência e foco, o que já distingue este trabalho dos demais.

Em geral, o processo de *feature engineering* exige conhecimento do domínio analisado. Tendo em vista este cenário, este trabalho também propõe dois métodos de *feature learning*, isto é, soluções que automaticamente entendam as características dos usuários e permitam a detecção de seu tipo, sendo isto a contribuição central deste trabalho.

Concluindo, este trabalho objetiva analisar dois tipos de soluções para

o problema da Seção 1.3. Uma destas é a pesquisa por soluções de *feature learning* para o problema de detecção de tipos de usuários, que não é comum para este tipo de problema. Além disso, nesta pesquisa também será feita uma análise detalhada dos comportamentos dos usuários através de múltiplas perspectivas (*feature engineering*), isto é, uma abordagem distinta da anterior, porém, complementar. Acima de tudo, as análises sob múltiplas perspectivas permite observar os usuários sob vários ângulos.

1.5 Metodologia

Para a verificação das soluções propostas foram realizados experimentos quantitativos, de natureza exploratória e explanatória. Nesta Seção, é brevemente discutido como foi conduzida esta pesquisa.

1.5.1 Dataset

Os dados utilizados para a condução desta pesquisa são oriundos de duas comunidades online de perguntas e respostas. A primeira é uma comunidade online do forum Stackexchange voltada para a aprendizagem de biologia, denominada a Biology Q&A[26]. A segunda também pertence ao Stackexchange, porém, voltada para a aprendizagem de Química, conhecida como Chemistry Q&A[27].

As duas comunidades são destinadas à aprendizagem informal, onde é possível postar uma pergunta sobre algum assunto relacionado à temática de cada uma e receber respostas ou comentários. Os comentários são, em síntese, complementos a uma pergunta ou a uma resposta visando melhor esclarecer algum ponto. No momento da extração dos dados, a comunidade Biology Q&A contava com 22.094 usuários, 15.934 perguntas, 19.009

[26] `http://biology.stackexchange.com/`
[27] `http://chemistry.stackexchange.com/`

respostas e 64.546 comentários. Já a Chemistry Q&A contava com 27.514 usuários, 21.983 perguntas, 25.776 respostas e 79.455 comentários. Os dados desta pesquisa foram obtidos através do Stackexchange dump[28].

1.5.2 Pesquisa Quantitativa e Exploratória

Conforme citado na Seção 1.4, foi conduzido um estudo das características dos usuários, objetivando diferenciar os usuários que se destacam dos demais. O estudo é de natureza exploratória, pois, questões de pesquisas são levantadas. Ademais, não se deseja comprovar hipóteses por meio de análises. Trata-se de um estudo mais abrangente onde, apesar de fazer uso de testes de inferência estatística comumente usadas para validar hipóteses, as questões de pesquisas são amplas. Isto é, por meio de análises estatísticas, são coletadas evidências de forma a permitir elaborar respostas para as questões, com a devida fundamentação. Acima de tudo, busca-se compreender melhor os fatores envolvidos no problema em questão, em sintonia com as sugestões dos pesquisadores Ig Bittencourt e Seiji Isotani [22].

No Capítulo 3, os detalhes sobre este estudo são apresentados, mostrando sua plena descrição e execução. Ademais, os dados usados neste trabalho são os citados na Seção 1.5.1.

1.5.3 Pesquisa Quantitativa e Explanatória

Como também comentado na Seção 1.4, métodos de *feature learning* foram criados como proposta de solução para o problema da pesquisa. Neste caso, trata-se de um estudo mais fechado, isto é, deseja-se verificar se os métodos de *feature learning* propostos são melhores ou piores que outros da literatura. Assim, realizou-se um estudo explanatório, com a elaboração de hipóteses (nula e alternativa), por meio do qual se pretende comprovar que

[28] `https://archive.org/download/stackexchange`

as propostas desta tese superam os demais métodos relacionados.

No Capítulo 4, todas as informações sobre a condução deste estudo são expostas e os dados utilizados também são os descritos na Seção 1.5.1.

1.6 Contribuições

A pesquisa descrita nesta tese já foi avaliada por pares de pesquisa de veículos importantes. O conteúdo das publicações desta pesquisa são as contribuições desta tese [111, 110, 109, 113, 112, 8, 116, 114, 120, 115]. Salienta-se que cinco destas têm QUALIS B1 e uma A1, verificados no período da publicação. É importante ressaltar também que todas as partes da tese estão contidas nos trabalhos citados. Em resumo, as principais contribuições são: (i) a elaboração de uma metodologia para caracterizar usuários de comunidades online, permitindo distinguir os que se destacam com relação aos demais sob diversas perspectivas; (ii) a proposição de dois métodos que captam automaticamente características dos usuários, também permitindo a diferenciação dos que se destacam com relação aos outros. Maiores detalhamentos sobre as contribuições se encontram nos Capítulos 3, 4 e 5.

1.7 Organizaçao do Trabalho

Nesta Seção apresentaremos a organização do trabalho. Além deste Capítulo introdutório, no Capítulo 2 é apresentada a fundamentação desta pesquisa, isto é, conceitos e trabalhos relacionados. No Capítulo 3, é apresentado o estudo do comportamentos dos usuários, de forma a possibilitar a diferenciação entre os tipos de usuários. No Capítulo 4, as abordagens e análises de *feature learning* são expostas e comparadas. Por fim, no Capítulo 5, a conclusão é apresentada.

2. Fundamentação

Este Capítulo é destinado aos fundamentos desta pesquisa. Assim, neste são apresentados os conceitos essenciais como, por exemplo, o de comunidades online, sua relação com compartilhamento de conhecimentos, a noção de prestígio social (que pode ser adquirido pelos membros destes ambientes), dentre outros. Estes conceitos são primordiais, pois, constituem a base para o entendimento do objeto de estudo desta pesquisa. Por fim, alguns trabalhos relacionados que desenvolveram soluções para do problema enunciado na Seção 1.3 são discutidos também neste Capítulo.

2.1 Comunidades Online

As comunidades online são lugares onde indivíduos se reúnem em um espaço na Web, com o objetivo de discutir ideias, socializar, se divertir ou pedir ajuda para outras pessoas [152]. Em geral, pessoas costumam se reunir para formarem grupos das mais diversas naturezas com a finalidade de promover debates sobre assuntos de seus interesses. Assim, possuem um objetivo e utilizam o suporte de alguma tecnologia, além de serem regidas por normas e regras [108].

Apesar do termo 'comunidade' ser amplamente conhecido e utilizado no contexto online, a ideia não é nova. Na década de 1970, por exemplo, já exis-

tia o e-mail, que, embora rudimentar, passavam a noção de agrupamento de pessoas por interesse ou assunto no universo online [82]. Atualmente, muitos indivíduos utilizam o seu tempo livre em comunidades online com o objetivo de realizar alguma atividade de aprimoramento profissional ou pessoal sem necessariamente serem remunerados por isso. Dentre essas atividades, se pode citar, por exemplo, a procura por novas maneiras para projetar ou refinar produtos [56], a busca por ajuda para desenvolver ou depurar um novo software [66] a escrita de textos e espera críticas [130] ou mesmo a exposição de ideias através de artes ou imagens [164].

Além disso, as comunidades online podem ter como público-alvo uma quantidade variada de usuários. Esses usuários podem ser o público geral (por exemplo, os usuários dos grupos de propósito geral do Facebook), profissionais do mercado de trabalho (como os usuários do LinkedIn[1]) ou mesmo um público específico, como profissionais de informática (como os usuários da comunidade Stack Overflow[2]).

Apesar de grande parte dessas comunidades serem destinadas ao compartilhamento de conteúdos, os objetivos de cada uma podem ser bem diferentes. Algumas, por exemplo, têm como objetivo compartilhar fragmentos de código fonte de programas (como o Snipplr[3]), outras são destinadas ao compartilhamento de projetos de software inteiros (como o Github[4] e o Bitbucket[5]), algumas são usadas para compartilhar imagens ou fotografias (Flickr[6]). Outras comunidades têm como objetivo o compartilhamento de conhecimentos através da construção de conteúdos (como a Wikipédia[7]) ou através de per-

[1] `http://linkedin.com/`
[2] `https://stackoverflow.com/`
[3] `https://snipplr.com`
[4] `https://github.com/`
[5] `https://bitbucket.org/`
[6] `https://www.flickr.com/`
[7] `http://www.wikipedia.org/`

guntas e respostas (Yahoo! Answers[8], Quora[9], Stack Overflow).

Dado o apresentado, é notável que comunidades online podem ser utilizadas para vários fins. Em especial, esta pesquisa está interessada nas comunidades de perguntas e respostas utilizadas para o compartilhamento de conhecimentos.

2.2 Compartilhando Conhecimentos

A palavra compartilhar pode ter dois significados[10]: (i) dividir ou repartir algo com alguém; (ii) ou tomar parte em algo, isto é, coparticipar, associar, coautorar. No contexto desta pesquisa, o termo compartilhar se enquadra melhor na segunda definição, pois, as comunidades online podem proporcionar um ambiente rico para interações, colaborações e consequentemente, construção de conhecimentos [141].

Jadin et al. [69] argumentam que as comunidades online se tornaram importantes meios para a troca de experiências e conhecimentos. Isso se deve a grande quantidade de pessoas que participam dessas comunidades que, por sua vez, possuem os mais diversos tipos de conhecimentos e experiências. Desta forma, é possível compartilhar conhecimentos em uma escala muito superior que a forma tradicional.

No contexto das comunidades online, muitas pesquisas são baseadas em técnicas de análises de redes sociais. Estas buscam entender os aspectos de uma comunidade que estão relacionados ao compartilhamento de conhecimentos. Em geral, esses aspectos estão fortemente ligados às interações dos usuários, uma vez que, compartilhar algo necessariamente envolve a participação de mais de uma pessoa. Em um estudo realizado no Yahoo! Answers

[8] `https://answers.yahoo.com/`
[9] `https://www.quora.com/`
[10] `https://www.sinonimos.com.br/compartilhar/`

sobre compartilhamento de conhecimentos [2], um dos aspectos analisados foi o grau de reciprocidade entre usuários, que consiste em uma métrica que objetiva explicitar quantitativamente os usuários que fornecem ajuda (respondem a perguntas de outros) e são ajudados (tem suas perguntas respondidas) nas várias categorias (assuntos) da comunidade. Os resultados desse estudo foram variados. Por exemplo, na categoria "casamento" e "luta greco-romana" o grau de reciprocidade foi alto. Porém, na categoria "programação", o grau de reciprocidade foi baixo. Assim, possivelmente, as características do público de cada categoria determina o grau de reciprocidade. Um outro aspecto interessante analisado no trabalho foi a profundidade do conhecimento dos usuários, analisadas a partir de postagens de usuários aleatoriamente escolhidas para serem avaliadas por profissionais especializados nos assuntos em questão. Foi concluído que, no Yahoo! Answers, os usuários não têm um conhecimento profundo sobre os assuntos comentados.

Em síntese, o estudo do compartilhamento de conhecimento envolve o estudo das interações entre os membros de comunidades. No caso das comunidades online, o estudo quantitativo pode ser rico em detalhes e informações devido à grande quantidade de usuários com múltiplas e diversos tipos interações.

2.3 Comunidades de Perguntas e Respostas

Em comunidades online de perguntas e respostas, em geral, as pessoas entram, fazem alguma pergunta e rapidamente obtêm uma resposta devido ao grande número de usuários [132]. Nessas comunidades, as discussões têm uma estrutura de trilhas (threads): um usuário posta uma pergunta ou tópico e, logo após, outros usuários postam respostas ou comentários relativos à pergunta [3]. Além disso, cada thread pertence a pelo menos uma categoria

da comunidade, que são as palavras-chave que definem o tema da discussão. Ademais, os usuários geralmente recorrem a tais comunidades quando: (i) suas perguntas são muito específicas e requerem uma resposta direta de outros usuários que vivenciaram um problema semelhante no passado; (ii) nenhuma página na Web pôde ajudá-lo; (iii) necessitam se comunicar e conversar com outros usuários [6]. Nestas comunidades, em geral, há esquemas de avaliação ou moderação onde determinado usuário pode ser avaliado por outros usuários baseado em suas perguntas ou respostas postadas.

De forma geral, as comunidades online de perguntas e respostas são definidas como 'comunidades de prática'. Em síntese, comunidades de práticas são compostas por pessoas que possuem interesse em comum e interagem entre si para, através da aprendizagem coletiva, desenvolver competências e evoluir seus repertórios de experiências [161]. Um exemplo de comunidades de práticas são os fóruns de discussões específicos de alguns assuntos. No Brasil, o fórum GUJ[11] é um exemplo de uma comunidade de prática. Nele, profissionais que trabalham com alguma linguagem de programação se reúnem para aprimorarem seus conhecimentos, resolver dúvidas, pedir orientações etc.

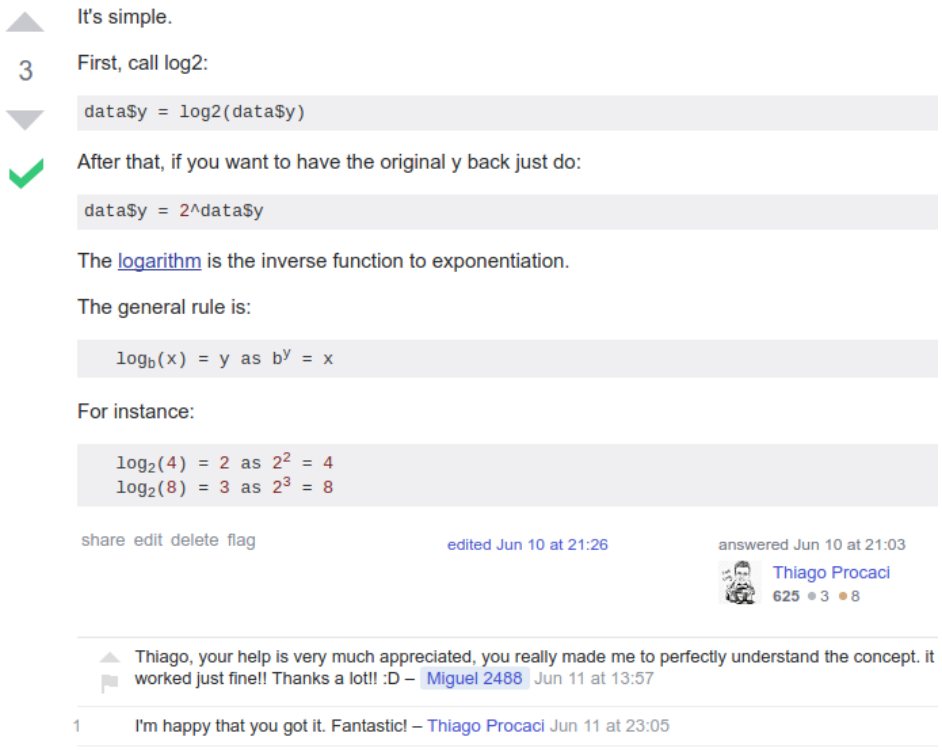
2.4 Prestígio Social: Ocupando Posição de Destaque

Em comunidades online, é comum observar que alguns usuários se destacam mais quando comparados a outros, durante os momentos de interações. Esse destaque ocorre muitas vezes devido às participações que são consideradas importantes, isto é, aquelas que de fato conseguem fazer com que outros usuários tirem alguma lição, conclusão ou aprendizado. Esses usuários de destaque, muitas vezes, são também conhecidos como usuários confiáveis, especialistas, notáveis, excepcionais etc [120, 114]. A Figura 2.1 ilustra

[11] `http://www.guj.com.br/`

uma participação de destaque no Stack Overflow, onde a resposta do usuário agradou a pessoa que postou a dúvida[12]. Este agrado pode ser percebido no comentário da resposta, no qual o *feedback* foi dado. Além disso, o destaque pode ser pontual, isto é, o usuário se destacou em momentos bem específicos ou geral, onde o usuário constantemente faz participações relevantes na comunidade. Usualmente, pesquisas tendem a focar mais no destaque geral, conforme mostrado na Seção 2.6.

Figura 2.1: How to inverse a log2 transformation?



Para saber se alguém sabe alguma coisa, pessoas são submetidas a avaliações. Nas escolas e nas universidades, uma forma comum para avaliar se um aluno deve ser aprovado ou não é através das provas. Há divergências que contestam se prova é a forma mais adequada para avaliar alguém [81], porém,

[12] `https://bit.ly/2MtDmkx`

não será abordada esta discussão neste trabalho. Uma prova consiste em um conjunto de perguntas as quais cada aluno deve responder. O professor avalia estas respostas, com a finalidade de atribuir uma nota para cada aluno e, posteriormente, classificá-lo como aprovado ou reprovado. Além disso, essas notas servem também como parâmetro, tanto para o professor quanto para os próprios alunos, saberem quem precisa melhorar em algum ponto. Uma outra forma para medir o conhecimento de alunos, no meio acadêmico, é através de avaliações colaborativas. Nesse tipo de avaliação, os aprendizes avaliam o próprio trabalho, os trabalhos dos seus colegas, assim como, são avaliados pelo professor e/ou por avaliadores externos [151]. Dividir a responsabilidade de avaliação, entre os diversos papéis, possibilita olhares diferentes para o mesmo trabalho, o que aumenta as possibilidades de identificação de pontos de melhoria e de pontos positivos no trabalho realizado. Em síntese, para avaliar alguém, é preciso que este se expresse de alguma forma, seja de forma escrita, oral, através de desenhos, qualquer coisa que permita que outros (ou ele mesmo) emitam seu julgamento a respeito de tal expressão.

Algumas comunidades online de perguntas e respostas criaram um mecanismo similar ao de avaliações escolares, mas com o objetivo de descobrirem quem são os seus melhores membros. Em geral, nestas comunidades, os usuários podem construir a sua reputação na rede, podendo ser positiva ou negativa. Na Figura 2.1, além do *feedback*, a resposta obteve 3 votos positivos (número 3 entre as setas) e foi escolhida como melhor resposta (através da marcação em verde). Assim, essa reputação é construída com base em avaliações de perguntas ou respostas de um usuário. Em outras palavras, cada usuário é avaliado por outros usuários baseado em suas perguntas ou respostas postadas. Um usuário com alta reputação geralmente é aquele que possui um prestígio social especial na rede, isto é, uma posição de destaque.

Sob a perspectiva do destaque geral, o prestígio de um usuário está relacionado com a sua quantidade de conhecimento exposto em uma comunidade online de perguntas e respostas [165]. Ou seja, à medida que um usuário vai contribuindo com boas participações em uma comunidade, seu prestígio tende a aumentar. Similarmente, nas relações sociais reais, e não virtuais, pessoas que recebem elogios devido as suas capacidades tendem a ter um maior prestígio no meio em que vivem [158].

Em algumas comunidades online, como o Stack Overflow, existem incentivos para um usuário construir uma boa reputação na rede. Em tais comunidades, um usuário com alta reputação possui mais privilégios como, por exemplo, ter a capacidade de moderar tópicos, corrigir respostas de outros usuários ou fornecer comentários esclarecedores. Bosu et al. fizeram um estudo no Stack Overflow e concluíram que a busca por privilégios em uma comunidade acaba sendo um fator motivador e, uma vez os alcançando, o usuário acaba inspirando mais confiança aos outros usuários [27]. Além disso, algumas atividades podem contribuir para um usuário construir a sua reputação mais rapidamente. Dentre elas cita-se: participações substanciais, responder perguntas relacionadas a tópicos pouco explorados, ser um dos primeiros a responder uma pergunta, ser ativo fora do horário de pico da comunidade e prover respostas com detalhamentos. Por outro lado, esta busca por privilégios também pode ser nociva, tornando a comunidade mais competitiva [114] e possivelmente mais hostil[13].

Humanos são seres que vivem em grupo e a busca por um status social diferenciado permeia todos os meios, inclusive o virtual. As razões para alguém querer ser diferente, em especial no contexto das comunidades online, podem ser várias como: (i) a vaidade, a necessidade de ser reconhecido, isto é,

[13] `https://bit.ly/2HuChqk`

a dependência do 'like' alheio para se sentir completo; (ii) a simples vontade de aprender, que naturalmente pode conferir a alguém o destaque merecido; (iii) a generosidade para compartilhar o que se sabe com outros. Nesta linha, e sem entrar nas razões para a busca do destaque, surgem pesquisas que buscam, principalmente, identificar características para determinar como alguém ou um grupo se distingue dos demais.

2.5 Interseções Entre Comunidades Online e Teorias

Esta Seção tem como objetivo traçar paralelos entre teorias de aprendizagem (bem como conceitos correlatos) e comunidades online. As comunidades estudadas neste trabalho são voltadas para aprendizagem. Assim, é relevante discutir suas interseções com as teorias de aprendizagem.

2.5.1 Construtivismo: Dialogando e Aprendendo na Rede

Segundo as palavras proferidas pelo professor Marco Silva da Universidade do Estado do Rio de Janeiro e da Universidade Estácio de Sá, em um encontro na Universidade Federal do Estado do Rio de Janeiro no auditório Paulo Freire em abril de 2016, os ambientes online tornaram possível materializar o 'desejo' de autores clássicos da pedagogia como Paulo Freire, Anísio Teixeira e Vygotsky, que é a possibilidade de interação plena entre todos sem as limitações que uma sala de aula presencial impõe. Além disso, interação online para a aprendizagem tem relação com o estilo diferenciado do artista Hélio Oticica [139], onde ele convida o espectador a interagir com o artista para construir sua obra[14].

Voltando para as comunidades online, o fato é que elas têm grande aceitação, principalmente entre os mais jovens[15], constituindo espaços de colabo-

[14]Arte parangolé: `http://www.saladeaulainterativa.pro.br/texto_0004.htm`
[15]Como mostrado na Seção 1.2.2.1

ração, permitindo a geração de capital social [124] e inteligência coletiva [75]. De acordo com Vygotsky [154] e Bakhtin [12], o processo de aprendizagem é essencialmente social, isto é, acontece por meio da interação de um com outro. Assim, pode-se entender as comunidades online como espaços de diálogo com possibilidade de aprendizagem. Apesar da possibilidade, aprender nestes espaços demandam atitudes responsiva de seus membros [121] e quando não acontecem, verifica-se a existência de problemas como os da *social query*, conforme discutido nas Seções 1.2.3.2 e 1.3.

A ideia de atitude responsiva estão nos textos de Bakhtin [12]. Em síntese, tal ideia diz que a relação com o outro pressupõe ativismo e responsividade da parte do interlocutor [166]. Em outras palavras, o locutor não espera uma compreensão passiva do interlocutor mas uma resposta (seja de concordância, objeção ou discordância). Voltando às características das comunidades online, e evidenciando a interseção com Bakhtin, uma pergunta postada espera por uma resposta, podendo ela ser verbal ou por meio de avaliações subjetivas do tipo 'Gostei' ou 'Não Gostei' [121]. O conceito de resposividade também se conecta com a ideia de comunicação de Marco Silva. Ele pontua a necessidade de se estabelecer a comunicação de todos para todos [140], livre das amarras tradicionais, ou seja, do estabelecimento de um pólo único de transmissão de conhecimento. Assim, o interlocutor é convidado a dar sentido a mensagem através de intervenções. Por outro lado, Marco Silva argumenta que participação do estilo 'Gostei' ou 'Não Gostei', 'Sim' ou 'Não' constituem formas pobres de intervenção em um diálogo.

Siemens e Weller caracterizam comunidades online e redes sociais como ferramentas construtivistas que atendem aos objetivos de novos métodos de participação de ensino, capazes de impactar na autonomia do estudante, contrapondo a utilização dos ambientes online ao modelo de educação tra-

dicional, estruturado na figura central do professor e no fluxo unilateral de conteúdo [138]. Claramente, Siemens e Weller dialogam com a proposta de comunicação de todos para todos de Marco Silva. No construtivismo, o diálogo é fundamental para a aprendizagem, porém, qualquer diálogo[16] ajuda? Siemens e Weller têm uma visão idealizada das comunidades online? Confiança em ambientes de aprendizagem é essencial [115]. Alguém em um ambiente online deve se sentir livre para perguntar, começar debates, criticar e ser criticado. Boas interações, isto é, aquelas respeitosas e focadas no assunto discutido devem ser incentivadas [92]. Ademais, há esforços para inibir vandalismos, postagens fora de contexto, ataques *ad hominem* por meio de ferramentas automáticas em ambientes online [116].

Dado o exposto, entender que comunidades online estabelecem sempre um espaço ideal para o debate, livre de postagens não adequadas[17] é uma visão ingênua [23]. Na perspectiva do autor desta pesquisa, alinhado com o conceito de responsividade de Bakhtin, procurar maneiras para minimizar alguns dos problemas da *social query* (discutidos nas Seções 1.2.3.2, 1.3, 1.4) constituem formas para tornar ambientes onlines mais construtivistas.

2.5.2 Conectivismo

Um dogma central das teorias de aprendizagem pré-digital (como o construtivismo) é que elas estão mais preocupadas com o processo de aprendizagem em si, e menos com o valor do que está sendo aprendido [137]. Ainda de acordo com Siemens [137], isso provavelmente acontece pois foram formuladas dentro de uma lógica de escassez de informação. Em um mundo conectado em rede, com abundância de informação, saber avaliar o valor da informação que se adquire é muito relevante [43]. Além disso, cada vez mais rapidamente

[16]Considerando que diálogo é interação.
[17]https://bit.ly/2HuChqk

informações se tornam obsoletas devido, principalmente, a sua velocidade de produção e de disseminação atualmente. Desta forma, surge a necessidade de avaliar a importância de aprender algo como uma espécie 'meta-habilidade' aplicada anteriormente ao próprio processo de aprendizagem [137]. Numa sociedade em rede, onde o conhecimento é abundante e a quantidade de informação cresce exponencialmente, a capacidade de sintetizar e reconhecer conexões e padrões é uma competência valiosa [137] [135]. Muitas vezes, a interação entre pessoas pode não ser necessária, pois, já existe muita informação disseminada no espaço digital sendo, no entanto, necessário somente encontrá-la.

Questão atual: o que os mais jovens devem aprender considerando que vivem na era digital de rápidas transformações? Segundo Zygmunt Bauman, sociólogo que descreve nossa atualidade sob uma visão fascinante, afirma que os tempos são 'líquidos' pois tudo muda rapidamente, isto é, nada é feito para durar, para ser 'sólido' [19]. Há previsões[18] indicando que 65% das crianças que entraram na escola em 2011 trabalharão em profissões que sequer foram inventadas. Consequentemente, é provável que sejamos a primeira geração da história que não sabe precisamente o que ensinar às crianças na escola ou aos estudantes na faculdade, conforme Yuval Harari explica em sua obra Homo Deus [64]. Este ponto se conecta com a meta-habilidade descrita por Siemens [137] pois há abundancia de informação e mudanças rápidas acontecem.

Tendo em vistas estas colocações, o conectivismo surge como uma teoria de aprendizagem pós-digital, que argumenta que o conhecimento não existe somente na mente do indivíduo mas também no caos de informações que o mundo digital viabilizou. Desta forma, dentro da lógica do conectivismo, mais importante que o ato de aprender é saber navegar na rede de informações

[18] `http://raley.english.ucsb.edu/wp-content2/uploads/234/CDavidson.pdf`

existentes (por exemplo, em comunidades online, redes sociais ou ambiente de aprendizagem online tradicionais) para aprender aquilo que lhe for mais agregador. Ou seja, o aprendiz deve saber se posicionar na rede, através da navegação e também da contribuição para ampliação da rede, de forma que ele consiga fazer as associações que mais lhe agregará conhecimento.

Existem teorias de análises de redes sociais que, mesmo em outro contexto, corroboram a ideia central do conectivismo. Por exemplo, há trabalhos que afirmam que o bom posicionamento de uma pessoa em uma rede pode lhe trazer mais conexões e, consequentemente, possibilidade de conhecimentos [25]. Além disso, existem trabalhos que afirmam que a Internet está provocando mudanças no cérebro humano, uma vez que, o espaço digital serve como uma memória externa a ele [159].

Sem entrar no mérito das controvérsias do conectivismo [32], verifica-se que pessoas bem posicionadas em uma rede estimulam a geração de espaços com boas interações e com discussões mais longas, conforme descrito adiante neste trabalho. Assim, sob o ponto de vista conectivista, tais espaços especiais podem ser detectados (inclusive de forma preditiva como na Seção 3.2.6.3) e serem úteis para filtrar informações. Assim, possivelmente, a detecção de tais lugares pode apoiar na exploração de formas de aquisição da informação e no estabelecemento de conexões entre pessoas-pessoas e pessoas-conteúdos.

2.5.3 Zona de Desenvolvimento Proximal

O conceito de zona de desenvolvimento proximal é bem conhecido na pedagogia. Em síntese, a zona de desenvolvimento proximal de um indivíduo consiste na distância entre o seu conhecimento atual e os conhecimentos potenciais que ele pode vir a ter, se obtiver colaboração de outras pessoas [154]. Paradoxalmente, apesar da definição usual de zona de desenvolvimento proximal ser associado a uma distância, trabalhos de educação argumentam que

elas podem ser construídas através de atividades de caráter dialógico, que tem o potencial de gerar aprendizagem, sendo, portanto também associada a espaços propícios para aprender. Entretanto, apesar disto, este conceito é amplamente aceito nas formas de ensino tradicionais, cabendo, muitas vezes, ao professor identificar o nível de conhecimento de um aluno e propor situações interativas de aprendizagem para que ele possa evoluir a um novo patamar.

Existem evidências que as zonas de desenvolvimento proximal se formam também em ambientes online. Cíntia Rabello [121] apresenta um estudo exploratório qualitativo acerca da utilização de um grupo do Facebook, para o ensino de língua inglesa na educação de nível superior. O objetivo disto, segundo a autora, era expandir as interações realizadas em sala de aula. Os participantes desse estudo foram a professora-pesquisadora e 57 alunos de um curso de graduação em relações internacionais no ano de 2012. Para evidenciar a existência de zonas de desenvolvimento proximal no grupo, a professora solicitou que os alunos comentassem os vídeos de especialistas postados por ela, emitindo sua apreciação crítica. Através desta atividade, foi possível constatar a formação de pequenas zonas de desenvolvimento proximal entre os alunos e os especialistas, no caso os locutores dos vídeos, e entre os próprios alunos que, em colaboração, comentaram os vídeos interagindo uns com os outros. Apesar da subjetividade da questão analisada, uma vez que, a identificação do avanço do nível de conhecimento se deu através de interpretações das expressões dos alunos no grupo, os resultados corroboram aquilo que já é conhecido no ensino tradicional [13, 52, 10]. Em outro trabalho [57], foi apresentado um estudo similar, porém, com estudantes que utilizavam ferramenta de groupware (sistema colaborativo) para interagir. Neste, através da leitura dos logs do sistema e uso de técnicas de análise de redes sociais,

foi ressaltada a importância de fomentar a presença social dos alunos para a construção de zonas de desenvolvimento proximal.

Nota-se que as zonas de desenvolvimento proximal têm profunda relação com a teoria construtivista apresentada na Seção 2.5.1. No entanto, sob a perspectiva conectivista, é possível que já exista uma zona de desenvolvimento proximal em uma comunidade online e esta pode ser identificada através dos registros deixados pelos aprendizes. Neste cenário, possivelmente, um novo participante simplesmente pode identificar que ela se inicia em seu nível atual de conhecimento e termina em um patamar superior e suficiente, de acordo com seu objetivo educacional. Neste caso, cabe ao participante ler e refletir sobre as informações disponibilizadas para aprender.

Concluindo, entender o sucesso de alguns usuários de comunidades online pode ajudar na construção de zonas de desenvolvimento proximal. Adiante, será debatido como tais usuários podem exercer influência em uma rede de pessoas e criarem lugares online promissores para aprender.

2.6 Trabalhos Relacionados

Diversos trabalhos relacionados buscam identificar pessoas que se destacam em um determinado ambiente online. Em geral, nestes trabalhos, quando no contexto das comunidades online de perguntas e respostas, o destaque usualmente significa competência em determinados tópicos. Como é o caso do trabalho de Streeter e Lochbaum [147] e Krulwich et al. [73], que focaram em técnicas de recuperação de informação e processamento de linguagem natural para encontrar os usuários de destaque em assuntos. Neste cenário, os textos produzidos no ambiente online são representados através de vetores de termos (palavras ou tokens), com suas respectivas frequências. A partir disto, foi possível detectar as possíveis competências que uma pessoa

tem, baseado em seus discursos. No entanto, as técnicas de recuperação de informação e processamento de linguagem natural são limitadas para captar o nível de competência de cada usuário. Ou seja, é dificil determinar se uma pessoa, por exemplo, posta uma boa resposta somente analisando e processando seus textos postados em fóruns ou comunidades [165, 78].

Balog et al. propuseram um modelo com duas etapas para identificar os usuários de destaque em uma comunidade [17]. Este estudo foi baseado em consultas executadas no ambiente (por exemplo, alguém buscando por uma questão de um tópico específico) e também em uma coleção de textos associados aos usuários que têm potencial para se destacarem. Esta abordagem foi construída com base em técnicas de recuperação de informação e métodos probabilísticos, objetivando estabelecer a ligação entre as pessoas que buscam ajuda na comunidade com a lista de pessoas capazes de auxiliá-las. Apesar desta proposta ter sido empiricamente validada, a abordagem não deixa claro quais são os parâmetros que permitem dizer se determinado usuário realmente se destaca em algum assunto, isto é, se ele é capaz de resolver questões específicas de tópicos. Em outras palavras, como limitação do trabalho, os autores fizeram suposições simplificadas, assumindo que as coleções de textos relacionadas aos usuários representam a área de especialidade de cada.

Outra abordagem similar foi proposta por Liu et al. [79], através da elaboração de um framework que automaticamente gerava o perfil especializado do usuário que ressaltava as competências em que os usuários se destacavam. Estes perfis foram construídos com base nas semelhanças entre os tópicos de interesse da comunidade e a descrição individual fornecida por cada usuário em seu perfil pessoal, isto é, era uma conjugação entre interesses de todos e o perfil pessoal. Uma limitação evidente deste trabalho era a suposição que

a descrição fornecida no perfil pessoal refletia as competências do usuário. Esta limitação é parecida com a apresentada por Balog et al. [17] sendo, possivelmente, uma evidência de que não é facilmente contornada.

Campbell et al. [31] e Dom et al. [41] utilizaram o algoritmo de ranqueamento HITS em grafos para encontrar os usuários de destaque que faziam parte de uma lista de e-mail. Os resultados desses estudos foram animadores, uma vez que a abordagem baseada em grafos se mostrou eficiente. Contudo, esses estudos tinham uma fraqueza que residia no tamanho da rede analisada. A rede era relativamente pequena e os resultados poderiam não ser generalizáveis para outros contextos.

Zhang et al. propuseram a construção de um algoritmo baseado em grafos para o mesmo fim, porém, aplicado em um fórum de discussão online tradicional [165]. Apesar da abordagem ter se mostrado interessante, os autores do trabalho concluíram, através de simulações, que comunidades com diferentes características devem ser analisadas separadamente, pois, as características podem influenciar nos resultados obtidos, sendo necessárias adaptações nas medidas ou nas técnicas utilizadas. Isto é, a proposta parece não ter capacidade de generalização, assim como a de Campbell et al. [31] e Dom et al. [41], porém, demonstrada explicitamente pelas simulações.

Wang et al. [156] propuseram uma forma de identificar os usuários de destaque, construindo um modelo híbrido, combinando técnicas de recuperação de informações com algoritmos de ranqueamento em grafos. A limitação principal desta abordagem foi o uso de um método simples para encontrar e indexar todas as palavras-chave dos textos encontrados no perfil de cada usuário, assumindo que isto descreve sua competência. De acordo com os próprios autores da pesquisas, existem outras fontes de evidência mais relevantes para julgar se uma pessoa se destaca em algum tópico, como verificar

se a quantidade de postagens sobre determinado tema é relevante. Apesar de apresentar uma solução distinta de Liu et al. [79] e Balog et al. [17], as limitações são parecidas.

Souza et al. [144] propuseram uma ferramenta para encontrar os usuários de destaque que faziam parte lista de seguidores de um usuário do Twitter. A ideia desse trabalho era encontrar o usuário seguidor com o perfil mais adequado para responder a uma pergunta no Twitter. Os resultados dessa pesquisa foram interessantes, pois, no contexto analisado a proposta foi bem-sucedida. Contudo, a avaliação deste algoritmo foi feita com poucos usuários e uma avaliação mais robusta, comparando com outras abordagens, deveria ter sido realizada constituindo, portanto, uma limitação da pesquisa.

Um sistema para identificar pessoas que se destacam em determinadas competências foi proposto em 2012 por Li et al. [76]. Para construir o 'perfil das competências' de cada pessoa, informações da profissão do usuário, da confiabilidade percebida pelos outros, intimidade social e popularidade eram consideradas. Este estudo, além disso, modela o compartilhamento de conhecimento através de grafos, utilizando cadeia de Markov, para melhorar o processo de recomendação. Entretanto, como limitação, a pesquisa não compara a abordagem proposta com outras baseadas em grafo, aplicadas para o mesmo problema como em Zhang et al. [165] ou Campbell et al. [31].

Utilizando técnicas de análises de redes sociais, Odiete et al. [95] investigaram o relacionamento entre pessoas que se destacavam em diferentes linguagens de programação em uma comunidade online. Estas análises fizeram parte de um sistema de recomendação. Assim, os autores criaram um grafo, descrevendo os relacionamentos entre tópicos e usuários. Desta forma, as métricas extraídas do grafo poderiam ser indicadores de especialidade. Em outras palavras, eles exploraram este grafo, onde cada nó representava uma

linguagem de programação e seu tamanho, o número de pessoas que se destacavam nela. Ademais, a espessura da aresta simbolizava quantos usuários se destacavam em duas linguagens. Desta maneira, também conseguiram inferir qual era a linguagem de programação mais relevante para os membros da comunidade. Apesar de interessante, uma limitação deste estudo foi se fundamentar unicamente em observações (visualizações) no grafo para responder às questões de pesquisa. Claramente, assim como Souza et al. [144], ainda há lacunas a serem preenchidas nos experimentos realizados, o que pode comprometer as conclusões.

Fu et al. [55] endereçaram o problema para detectar usuários que se destacam, porém, durante suas primeiras participações na comunidade. Os autores propuseram um modelo de classificação (*Machine Learning*), que identificava tais usuários baseadas nas primeiras atividades. Como limitação desta abordagem pode-se citar: (i) o próprio aspecto temporal da solução, que elimina a possibilidade de detectar usuários que tiveram sua notoridade tardia; (ii) outras dimensões para detectar os usuários poderiam ser consideradas, além da interação.

Yang et al. [162] propuseram o algoritmo NEWHITS para identificar usuários de destaque. A proposta se trata de uma evolução do algoritmo tradicional HITS, que é usualmente utilizado neste mesmo problema [31, 41]. Embora o algoritmo NEWHITS tenha obtido melhores resultados que o HITS tradicional, é importante lembrar que o NEWHITS é um algoritmo de ranqueamento. Logo, há outros algoritmos que poderiam ser testados e comparados com o NEWHITS, como o proposto por Zhang et al.[165]. Pal et al. [98] propuseram uma abordagem similar para encontrar usuários que se destacam em determinados assuntos no Instagram. Apesar do foco ser diferente dos demais, isto é, o contexto não é encontrar pessoas para responder perguntas

ou com determinadas competências, as técnicas utilizadas são parecidas com demais trabalhos. Pal et al. [98] inferiram os interesses dos usuários a partir de suas biografias publicamente disponíveis, que eles mesmos reportaram.

Yeniterzi and Callan [163], diferentemente de Fu et al. [55], propuseram um modelo temporal para melhor estimar a competência dos usuários. Eles se basearam em abordagens clássicas para identificar usuários que se destacam, considerando métricas como o número de resposta ou o z-score, combinado com a informação temporal para melhor direcionar questões à usuários com capacidade para respondê-las. A motivação do uso da temporalidade reside na argumentação que as pessoas simplesmente mudam com o tempo. Por exemplo, alguém que se destacou no passado pode não se destacar hoje. Mukherjee et al. [93] utilizaram a combinação entre experiência do usuário, interesse em tópicos específicos, estilo de escrita e a avaliação de seu comportamento por outros usuários para captar sua evolução temporal. Tudo isto para identificar o nível de maturidade do usuário e, por fim, propor recomendações de forma similar à Yeniterzi and Callan [163]. Srba et al. [146] elaboraram um método que considera o uso de informações de serviços externos à comunidade como (blogs, microblogs ou outras redes sociais) com o objetivo de melhor identificar a competência do usuário. Esta abordagem minimiza as limitações do trabalho de Liu et al. [79].

2.7 Diferencial da Pesquisa

Com base nos trabalhos relacionados, se percebe que há diferentes soluções para o mesmo problema que, de forma geral, é a busca por usuários de destaque. Além disso, o contexto de aplicação pode variar e isto, por sua vez, varia também a noção de destaque. Entretanto, as técnicas utilizadas nas soluções são parecidas e algumas pesquisas têm inclusive limitações simi-

lares. Algumas têm relação com a quantidade de fatores considerados para a definição do 'destaque', isto é, poucos fatores são considerados podendo levar a uma visão distorcida do destaque de um usuário. No estudo empírico do Capítulo 3 são apresentadas algumas das várias perspectivas que podem estar associadas aos que se destacam, demonstrando um diferencial desta pesquisa.

Um ponto comum nas pesquisas mostradas é a necessidade de análises de características dos usuários ou da comunidade que, como definido no Capítulo 1, é também denominada de *feature engineering*. Claramente há espaço para sofisticação na etapa de descoberta de características. Assim, este trabalho propõe dois mecanismos automáticos para a descoberta de tais características, que são usualmente conhecidos como *feature learning*. Como exposto, os trabalhos relacionados giram em torno de soluções com foco em recuperação de informação, processamento de linguagem natural, algoritmos em grafos, porém, nenhum aborda a possibilidade do uso de *feature learning*. Neste sentido, este trabalho visa contribuir incrementando o 'cardápio de abordagens' possíveis para este tipo de problema. Há na literatura outros métodos de *feature learning*, porém, não aplicados no contexto deste trabalho. Assim, uma contribuição importante deste trabalho são os resultados das comparações entres as abordagens propostsa com as outras de *feature learning* existentes, aplicadas no problema endereçado.

2.8 Comentários Finais

A ideia deste Capítulo foi apresentar os conceitos fundamentais que permeiam a temática desta tese. Em resumo, foi apresentado o conceito de comunidades online, seus objetivos, públicos etc. Ademais, mostrou-se a relação de comunidades online com compartilhamento de conhecimentos e,

posteriomente, foi apresentada a descrição de comunidades online de perguntas e respostas. A noção de 'destaque' em comunidades foi também discutida e, brevemente, comentou-se sobre as razões pelas quais as pessoas alcançam ou buscam se destacar. Na sequência, discutiu-se os pontos de interseção do objeto de estudo deste trabalho com os teorias de aprendizagem.

Diversos trabalhos relacionados que abordam o problema da tese foram apresentados, discutindo suas propostas, limitações e pontos de similaridades entre eles. Por fim, o diferencial desta pesquisa foi ressaltado, evidenciando que esta se justifica e que há contribuições importantes.

3. Estudo Comportamental dos Usuários

Este Capítulo tem como finalidade discutir as características dos usuários que se sobressaem em comunidades online. Em síntese, busca-se identificar 'o que há por trás do destaque'. Para isto, um estudo sobre os comportamentos dos usuários é realizado, mostrando as diferenças daqueles que se destacam com relação aos demais.

3.1 Descrição do Estudo

Neste instante, se inicia a apresentação das etapas do estudo empírico conduzido neste Capítulo. Como mencionado, este faz parte do escopo desta pesquisa, conforme comentado na Seção 1.4. Ademais, se trata de um estudo exploratório, com a finalidade de se evidenciar os fatores que compõem o destaque de usuários.

O objetivo deste estudo é **analisar** os usuários **no contexto** das comunidades online, **com o propósito de** caracterizá-los **em relação** as diferenças que levaram uns obterem destaque e outros não, **sob o ponto de vista** das perspectivas comportamentais associadas aos usuários.

Neste cenário, se busca responder a algumas questões de pesquisas, com base em resultados obtidos por experimentos quantitativos.

3.1.1 Questões de Pesquisa

As questões de pesquisa que delineiam as análises do Capítulo podem ser enunciadas da seguinte forma:

- Q_1: Quais são as diferenças entre os usuários que se destacam com relação aos ordinários?
- Q_2: Qual é o conjunto de características mais preditivas quando se classifica um usuário em 'de destaque' ou ordinário?
- Q_3: O quão generalizáveis são os modelos que classificam os usuários em 'de destaque' ou ordinário?
- Q_4: A predição da melhor resposta de uma pergunta postada na comunidade está mais relacionada com as características das postagens da trilha ou com as características dos usuários participantes?

Detalhando as questões, a Q_1 objetiva investigar as características gerais que permitem dizer que alguém se destaca ou não em uma comunidade. Neste estudo, adotou-se a denominação 'ordinários' para aqueles que não se destacam. Sabe-se que na língua portuguesa o termo ordinário também pode significar falta de decência ou pessoas de caráter duvidoso[1]. **Entretanto, neste trabalho, o uso do termo ordinário se refere somente aos usuários comuns, isto é, aqueles que não possuem destaque.**

Assim, esta investigação visa salientar as diferenças entre os usuários que se destacam em relação aos demais, considerando as seguintes perspectivas: (i) a da participação; (ii) a dos traços de linguagem; (iii) a dos laços sociais; (iv) a da influência exercida; (v) e a do foco em assuntos. É importante mencionar que, para cada perspectiva, um conjunto distinto de dados são

[1] `https://www.dicio.com.br/ordinario/`

considerados e analisados. Frisa-se que as escolha de tais perspectivas não foi aleatória. Diversos trabalhos relacionados abordam tais perspectivas, porém, sem apresentá-las em conjunto. Por exemplo, alguns abordam os laços sociais [165, 95, 119], outros em laços sociais e foco [1, 117], outros em influência [97, 116], alguns em traços de linguagem [147, 113] etc. A única perspectiva que parece estar 'onipresente' é a da participação (mesmo que indiretamente), possivelmente por ser a mais elementar. Desta forma, o autor desta pesquisa buscou 'unificar' algumas das perspectivas encontradas mais recorrentes na literatura e consolidar as análises.

A questão Q_2 está relacionada com a criação de modelos de *Machine Learning* que recebem como entrada diferentes conjuntos de características dos usuários, isto é, os dados das distintas perspectivas, objetivando classificar os usuários. Acima de tudo, tais modelos poderão comprovar se o estudo empírico conseguiu 'entender o que há por trás do destaque', caso as classificações sejam bem-sucedidas.

A questão Q_3 tem como finalidade verificar se os modelos de classificação dos usuários criados em uma comunidade podem ser utilizados com sucesso em outra comunidade. Assim, espera-se ter uma ideia sobre a capacidade de generalização destes.

Por fim, a questão Q_4 tem como objetivo elaborar um modelo de predição para a próxima melhor resposta de uma pergunta, bem como, acentuar a melhor maneira para realizá-la. Para isto, uma análise das características dos textos das postagens e das características dos usuários associados à trilha são considerados.

3.1.2 Dados dos Experimentos

Os dados utilizados nos experimentos deste Capítulo são os mesmos daqueles comentados na Seção 1.5.1. Reiterando, as duas comunidades online

Tabela 3.1: Dataset

Comunidade	Usuários	Perguntas	Respostas	Comentários	Revisões
Biology Q&A (BQA)	22.094	15.934	19.009	64.546	62.610
Chemistry Q&A (CQA)	27.514	21.983	25.776	79.455	87.690

analisadas são de perguntas e respostas do Stackexchange. Estas são: a Biology Q&A (BQA) e a Chemistry Q&A (CQA). Conceitualmente estas comunidades podem ser descritas, conforme apresentado na Seção 2.3. Um detalhe ainda não comentado sobre os dados, é que é possível revisar uma postagem (perguntas ou respostas, por exemplo), sugerindo melhorias no texto, geralmente para melhor entendimento.

A visualização consolidada dos dados utilizados nos experimentos estão na Tabela 3.1, onde podem ser observados o número de usuários, perguntas, respostas, comentários e revisões de cada comunidade.

3.1.3 Definição dos Grupos

Para examinar as características de cada tipo de usuário, foi elaborada a seguinte definição: 15% dos usuários com maior reputação (top 15%), isto é, aqueles com maiores pontuações oriundas das avaliações de outros membros da comunidade, foram considerados usuários de destaque. Para obter estes usuários, eles foram colocados em ordem descendente com base em suas respectivas pontuações e, os que estivessem no top 15% desta lista, foram considerados de destaque. É importante ressaltar que os mecanismos de avaliação de comunidades online foram discutidos em detalhes na Seção 2.4, em especial, na explicação que envolve a Figura 2.1.

Os demais, não incluídos no top 15%, foram considerados ordinários, ou seja, comuns. Como comentado na Seção 1.4 e na 2.6, o foco nesta dicotomia (destaque vs ordinários) é comum em trabalhos relacionados e tem sido

utilizada em trabalhos anteriores [120, 97, 114, 165]. Baseado em tais trabalhos, foi verificado que a definição do grupo dos usuários de destaque, boa parte das vezes, oscila entre os 10% e 20% daqueles com melhores avaliações. Em outras palavras, os trabalhos consideram os usuários de destaque como alguém entre o top 10% e o top 20% com melhores pontuações. Desta forma, se entendeu factível considerar o top 15% das comunidades, como aqueles que obtiveram destaque, isto é, o meio termo.

Em algumas análises, esta pesquisa buscou identificar diferenças dentro do grupo dos que se destacam, considerando outros três subgrupos (top 5%, top 5-10% e top 10-15%). Além disso, em análises onde desejava-se mais detalhamentos, estes três subgrupos foram também comparados com os ordinários. Em síntese, esta Seção teve como finalidade apresentar os grupos de usuários envolvidos nos experimentos deste Capítulo.

3.1.4 Ameaças à Validade

Neste momento, serão elecandas algumas ameaças à validade deste estudo. Além disso, são discutidas as formas para minimizá-las.

3.1.4.1 Validade Interna

A validade interna diz respeito à capacidade de um novo estudo, utilizando os mesmos dados, replicar o comportamento do estudo atual [38]. Para garantir isto, foram utilizados dados públicos do Stackexchange, como relatado na Seção 1.5, e os experimentos desta pesquisa estão todos disponíveis no GitHub[2] para reprodução.

3.1.4.2 Validade Externa

A seleção dos sujeitos, em especial, aqueles que compõem o grupo dos que se destacam, pode não representar a população que de fato é de destaque em

[2] `https://github.com/thiagoprocaci/diff-ourstanding-ordinary`

Biologia ou Química que, por sua vez, são as temáticas das comunidades. Seriam eles somente os 'menos piores'? Em outras palavras, será que as postagens, em geral, são de tamanha baixa qualidade que qualquer um poderia se destacar? Estas questões constituem uma ameaça à validade externa da pesquisa. Apesar de existir esta possibilidade, os mecanismos de controle e moderação das comunidades do Stackexchange já foram reconhecidos como eficientes para manter a qualidades das postagens [104, 49], o que demonstra a viabilidade de estudos neste sentido.

Outra questão sobre a validade externa, relacionada ao argumento anterior, é a possibilidade dos resultados deste estudo serem replicados utilizando outros dados [38], isto é, de outras comunidades. Para minimizar isto, neste trabalho se optou por analisar dados de duas comunidades distintas.

3.1.4.3 Validade de Construção

Há uma ameaça a validade de construção deste estudo, relacionada à definição do que é destaque. Em trabalhos anteriores os usuários de destaque geralmente encontram-se entre os 10% e 20% daqueles com melhores avaliações. Entretanto, esta definição pode ser controversa. Por que não considerar os usuários top 11% como os de destaque? Por que não o top 17%? O fato é que não há uma definição final sobre tal questão, ou seja, é sempre dependente do julgamento do pesquisador que conduz tais estudos. Como consequência, a má definição do destaque pode levar à problemas quanto à condução do estudo. Assim, objetivando minimizar esta ameaça, esta pesquisa optou por escolher o meio termo entre a osciliação padrão averiguada (10% a 20%), definindo os usuários de destaque como aqueles que estão no top 15%. Reiterando, esta decisão foi devidamente embasada em pesquisas anteriores que sugerem definições nesta direção não sendo, portanto, arbitrária.

3.1.4.4 Validade de Conclusão

A validade de conclusão verifica a relação entre as variáveis utilizadas e os resultados obtidos, determinando a capacidade do estudo em gerar uma conclusão [38]. Objetivamente o problema é: será que minhas análises permitem chegar a determinada conclusão? Chegar a conclusões incompletas é um risco que se corre em pesquisas científicas. Por exemplo, a limitação de uma investigação a determinados dados ou perspectivas pode levar o pesquisador a concluir algo controverso. Por outro lado, é impossível considerar 'tudo', isto é, se considerar todos os dados ou perspectivas possíveis, corre-se o risco de se não chegar a lugar algum. Neste sentido, se optou por algo intermediário: nem pela limitação forte de perspectivas associadas aos comportamentos dos usuários, isto é, considerar somente uma perspectiva, e nem pela ampliação irrestrita destas.

Assim, foram consideradas algumas perspectivas dos usuários, como citado na Seção 3.1.1, sendo estas as bases das análises dos comportamentos dos usuários. Além disso, para comparar as diferenças entre os tipos de usuários, foram aplicados testes de inferência estatística, calculando o nível de significância e, por fim, verificando a superioridade de uma distribuição com relação à outra, através de métricas de tamanho de efeito. Há experimentos onde correlações foram usadas e, da mesma forma, seu nível de significância foi calculado para que a interpretação do resultado pudesse ser válida. Tais métodos são amplamente aceitos na comunidade científica [9, 123, 39].

3.1.5 Mecanismos de Análise

O estudo proposto se classifica como uma série de experimentos, onde as variáveis são representadas na escala razão, e estas são apresentadas ao longo das análises. Há diversos mecanismos possíveis de análise para estudos

desta natureza. No estudo deste Capítulo, estatística descritiva foi usada nas análises mais elementares. Ademais, também foi utilizado a correlação de Spearman para verificar associações entre variáveis e o teste de Wilcoxon-Mann-Whitney para verificar se existem diferenças nas distribuições de dados analisados. Em caso de diferenças, o tamanho de efeito é calculado através do método Vargha and Delaney's A12. A opção por tais métodos foi devido as distribuições encontradas serem não gaussianas [9, 123].

Por fim, depois das análises dos comportamentos dos usuários por meio dos métodos descritos acima, modelos de classificação de Aprendizagem de Máquina foram propostos e, para averiguar a qualidade das classificações, foi usada a métrica 'area under the receiver operating characteristics curve' (AUC), que é comumente usada para este fim [106].

3.2 Execução do Estudo

Nesta Seção, as análises relacionadas aos comportamentos dos usuários são descritas. Estas foram divididas em subseções, onde são abordadas as perspectivas comportamentais dos usuários e, por fim, o modelo de classificação proposto.

3.2.1 Perspectiva da Participação

Neste momento, a atenção se volta para a caracterização e entendimento das participações dos usuários. Reiterando, as participações dos usuários consideradas são as postagens, isto é, as variáveis deste estudo são as perguntas, as respostas, os comentários e as revisões.

Além disso, em algumas análises, as participações foram agrupadas por mês. Em outras palavras, desde a criação de cada comunidade, em cada mês, foram consolidadas as participações dos usuários. A comunidade BQA

foi dividida em 72 meses e a CQA em 63.

3.2.1.1 Usuários de Destaque São Criados Mais Cedo

Esta Seção tem como finalidade verificar quando cada tipo de usuário acessou pela primeira vez a comunidade. Neste momento, é utilizada somente estatística descritiva para descrever as análises em questão.

Na comunidade BQA, 13,14% dos usuários top 5 criaram sua conta no primeiro mês de atividade da comunidade, enquanto 0,64% dos ordinários desta mesma comunidade entraram na comunidade neste mesmo período. Na comunidade CQA, 8,19% dos usuários top 5 fizeram o primeiro acesso no primeiro mês, em contraste, com os 0,57% dos ordinários que também se juntaram à comunidade no mesmo período. Em síntese, no primeiro mês, foram criados: (i) 5,38% dos usuários top 5-10 da comunidade BQA; (ii) 3,04% dos usuários top 5-10 da comunidade CQA; (iii) 4,74% dos usuários top 10-15 da comunidade BQA; (iv) 2,25% dos usuários top 10-15 da CQA. Nos meses restantes, em ambas comunidades, foram encontradas porcentagens de criação mais baixas para os usuários de destaque (por volta de 1%) e, com relação aos ordinários, estes tiveram uma porcentagem de criação parecida com a do primeiro mês. Os resultados estão consolidados na Tabela 3.2. Desta forma, pode-se concluir que mais usuários de destaque têm suas contas criadas nas comunidades mais cedo.

3.2.1.2 Primeira Atividade Depois do Primeiro Acesso

Um usuário de destaque tende a esperar menos tempo para escrever e postar sua primeira participação, depois de se juntar à comunidade. Assim, com base nos dados das comunidades, calculou-se o tempo gasto por cada usuário para fazer sua primeira participação.

Primeiramente, se comparou o tempo dos usuários top 5 com o dos ordinários na comunidade BQA. Foi verificado que existe diferença entre os tempos

Tabela 3.2: Porcentagem de Criação de Usuários no Primeiro Mês

Comunidade	Grupo	Porcentagem
BQA	top 5	13,14%
BQA	top 5-10	5,38%
BQA	top 10-15	4,74%
BQA	ordinários	0,64%
CQA	top 5	8,19%
CQA	top 5-10	3,04%
CQA	top 10-15	2,25%
CQA	ordinários	0,57%

Tabela 3.3: Comparação Primeira Atividade BQA

Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
top 10-15	ordinários	<0,01	0,6	0,4
top 5	ordinários	<0,01	0,61	0,39
top 5-10	ordinários	<0,01	0,64	0,36
top 5	top 10-15	>0,05	inconclusivo	inconclusivo
top 5-10	top 10-15	>0,05	inconclusivo	inconclusivo
top 5-10	top 5	>0,05	inconclusivo	inconclusivo

analisados. Com base no teste de inferência descrito na Seção 3.1.5, com o nível de significância menor que 0,01. Posteriormente, o tamanho desta diferença (comumente conhecida como tamanho de efeito) foi calculado com o método também descrito na Seção 3.1.5. Na comunidade BQA, em 61% dos casos (tamanho de efeito), os usuários top 5 fizeram sua primeira postagem em menos tempo que os ordinários. Os ordinários fizeram sua participação mais rapidamente em 39% dos casos. (0,61 top 5 vs 0,39 ord.[3], p<0,01).

Resultados similares foram observados nos subgrupos dos usuários de destaque, quando comparados com os ordinários da comunidade BQA (0.64 top 5-10 vs 0.36 ord., p<0.01) (0.6 top 10-15 vs 0.4 ord., p<0.01). Na comu-

[3]ord. significa ordinário

Tabela 3.4: Comparação Primeira Atividade CQA

Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
top 10-15	ordinários	<0,01	0,65	0,35
top 5	ordinários	<0,01	0,65	0,35
top 5-10	ordinários	<0,01	0,65	0,35
top 5	top 10-15	>0,05	inconclusivo	inconclusivo
top 5-10	top 10-15	>0,05	inconclusivo	inconclusivo
top 5-10	top 5	>0,05	inconclusivo	inconclusivo

nidade CQA, resultados parecidos foram encontrandos. Nesta, em síntese, em 65% dos casos, os usuários de destaque, isto é, o top 15, criam sua primeira postagem em menos tempo que os ordinários (0.65 dest.[4] vs 0.35 ord., p<0.01). Quando comparadas as diferenças entre o tempo de criação da primeira postagem entre os subgrupos dos usuários de destaque (top 5, top 5-10, top 10-15), nenhuma diferença significativa pode ser observada em ambas comunidades. Os resultados estão nas Tabelas 3.3 e 3.4.

3.2.1.3 Nível de Participação

Os usuários de destaque são mais participativos quando comparados aos ordinários. Na comunidade CQA, os usuários de destaque fazem mais perguntas (0,81 dest. vs 0,19 ord., p<0,01), fornecem mais respostas (0,77 dest. vs 0,23 ord., p<0,01) e escrevem mais comentários (0,87 dest. vs 0,13 ord., p<0,01) e também revisam mais as postagens de outros (0,82 dest. vs 0,18 ord., p<0,01). Proporções muitos parecidas foram encontradas na comunidade BQA. Alguém pode esperar que os ordinários tendem a perguntar mais porque supostamente sabem menos. Porém, os resultados mostram o contrário: quanto mais se destacam, mais perguntam. Os resultados destas análises estão nas Tabelas 3.5 e 3.6.

[4]dest. significa usuários de destaque

Tabela 3.5: Nível de Participação BQA

Tipo Part.	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Perguntas	destaque	ordinários	<0,01	0,81	0,19
Respostas	destaque	ordinários	<0,01	0,81	0,19
Comentários	destaque	ordinários	<0,01	0,88	0,12
Revisões	destaque	ordinários	<0,01	0,85	0,15

Tabela 3.6: Nível de Participação CQA

Tipo Part.	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Perguntas	destaque	ordinários	<0,01	0,81	0,19
Respostas	destaque	ordinários	<0,01	0,77	0,23
Comentários	destaque	ordinários	<0,01	0,87	0,13
Revisões	destaque	ordinários	<0,01	0,82	0,18

3.2.1.4 Diferenças nos Subgrupos dos Usuários de Destaque

Analisando tais subgrupos, foram encontradas diferenças no nível de participação. A regra intuitiva é: quanto mais destaque um usuário tem, maior é seu nível de participação. Foi observado que esta hipótese está correta em ambas comunidades. Por exemplo, na comunidade BQA, os usuários top 5 quando comparados com os usuários top 5-10 fazem mais perguntas (0,56 top 5 vs 0,44 top 5-10, p<0,01), fornecem mais respostas (0,85 top 5 vs 0,15 top 5-10, p<0,01), postam mais comentários (0,57 top 5 vs 0,43 top 5-10, p<0,01) e revisam mais postagens de outros (0,62 top 5 vs 0,38 top 5-10, p<0,01). Similarmente, os usuários top 5-10 tendem a participar mais que os usuários top 10-15. As Tabelas 3.7 e 3.8 descrevem estes resultados.

3.2.1.5 Alcançando o Destaque

Conforme as análises anteriores, um nível mais alto de participação parece ser importante para ser alguém de prestígio nas comunidades. Entretanto, o quanto alguém deve participar para estar no grupo dos que se destacam? Considerando esta questão, foi calculada a participação média mensal dos

Tabela 3.7: Diferenças Entre os Que Se Destacam - BQA

Tipo Part.	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Perguntas	top 5	top 10-15	<0,01	0,61	0,39
Perguntas	top 5-10	top 10-15	<0,01	0,57	0,43
Perguntas	top 5-10	top 5	<0,01	0,44	0,56
Respostas	top 5	top 10-15	<0,01	0,89	0,11
Respostas	top 5-10	top 10-15	<0,01	0,6	0,4
Respostas	top 5-10	top 5	<0,01	0,15	0,85
Comentários	top 5	top 10-15	<0,01	0,66	0,34
Comentários	top 5-10	top 10-15	<0,01	0,58	0,42
Comentários	top 5-10	top 5	<0,01	0,43	0,57
Revisões	top 5	top 10-15	<0,01	0,74	0,26
Revisões	top 5-10	top 10-15	<0,01	0,62	0,38
Revisões	top 5-10	top 5	<0,01	0,38	0,62

usuários. Assim, na comunidade BQA, os usuários top 5 fazem em média 0,84 perguntas, fornecem 2,41 respostas, escrevem 7,4 comentários e revisam 5,72 postagens. Ainda na comunidade BQA, os usuários top 5-10 fazem em média 0,80 perguntas ao mês, 0,43 respostas, 1,73 comentários e 0,78 revisões. Os usuários top 10-15 da BQA postam em média ao mês 0,84 perguntas, 0,39 respostas, 1,36 comentários and 0,74 revisões. Por fim, os ordinários da BQA, postam 0,62 perguntas, 0,23 respostas, 0,81 comentários e 0,61 revisões. Resultados parecidos foram observados para a comunidade CQA: os usuários top 5 fazem em média 0,93 perguntas, 3,01 respostas, 7,86 comentários e 7,07 revisões; os usuários top 5-10 fazem 0,94 perguntas, 0,38 respostas, 1,64 comentários e 0,84 revisões; os usuários top 10-15 postam 0,88 perguntas, 0,28 respostas, 1,21 comentários e 0,71 revisões; os ordinários, por fim, fazem em média 0,79 perguntas, 0,20 respostas, 0,72 comentários e 0,63 revisões. Os resultados são apresentados na Tabela 3.9 e 3.10.

De acordo com as análises, se pode concluir que, às vezes, o esforço realizado pelos usuários top 5-10 e top 10-15 é similar ao esforço dos ordinários, embora seja maior. Este esforço extra parece fazer a diferença ao longo do

Tabela 3.8: Diferenças Entre os Que Se Destacam - CQA

Tipo Part.	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Perguntas	top 5	top 10-15	<0,01	0,65	0,35
Perguntas	top 5-10	top 10-15	<0,01	0,61	0,39
Perguntas	top 5-10	top 5	<0,01	0,41	0,59
Respostas	top 5	top 10-15	<0,01	0,9	0,1
Respostas	top 5-10	top 10-15	<0,01	0,61	0,39
Respostas	top 5-10	top 5	<0,01	0,15	0,85
Comentários	top 5	top 10-15	<0,01	0,71	0,29
Comentários	top 5-10	top 10-15	<0,01	0,6	0,4
Comentários	top 5-10	top 5	<0,01	0,4	0,6
Revisões	top 5	top 10-15	<0,01	0,8	0,2
Revisões	top 5-10	top 10-15	<0,01	0,61	0,39
Revisões	top 5-10	top 5	<0,01	0,31	0,69

tempo, colocando uns em posições de destaque na comunidade e outros não.

3.2.1.6 Postagens com Pontuação Zero

Uma grande parte das postagens dos usuários de destaque recebem pontuação zero (neutra). Isto pode significar duas coisas: (i) ninguém as avaliou; (ii) ou o número de avaliações positivas é igual ao número de negativas. Na comunidade CQA, 49,5% das postagens dos usuários top 5, 55,4% dos usuários top 5-10 e 52,2% dos usuários top 10-15 receberam pontuação igual a zero. Uma conclusão muito parecida foi constatada na comunidade BQA, conforme na Tabela 3.11.

As comunidades do Stack Exchange tendem a valorizar mais as avaliações positivas que as negativas, para construir pontuação do perfil do usuário. Caso alguém faça uma postagem que receba uma avaliação positiva e uma negativa, embora a pontuação da postagem seja zero, a pontuação do usuário será positiva, porém menor, caso tivesse saldo positivo de avaliações da postagem. Concluindo, dado a alta porcentagem de postagens com pontuação zero, aproximadamente metade delas não contribui para o melhor entendimento de um assunto. Entretanto, apesar disso, por volta de 1% (variando

Tabela 3.9: Média Participação Mensal BQA

Tipo Part.	Grupo	Média Part. Mensal
Perguntas	top 5	0,84
Respostas	top 5	2,41
Comentários	top 5	7,4
Revisões	top 5	5,72
Perguntas	top 5-10	0,80
Respostas	top 5-10	0,43
Comentários	top 5-10	1,73
Revisões	top 5-10	0,78
Perguntas	top 10-15	0,84
Respostas	top 10-15	0,39
Comentários	top 10-15	1,36
Revisões	top 10-15	0,74
Perguntas	ordinários	0,62
Respostas	ordinários	0,23
Comentários	ordinários	0,81
Revisões	ordinários	0,61

para menos e para mais) das postagens dos usuários de destaque recebem mais avaliações negativas que positivas em ambas comunidades.

3.2.1.7 Tratamentos dos Ordinários

Os usuários ordinários têm menos avaliações positivas quando comparados com os usuários de destaque. Entretanto, será que as postagens dos usuários ordinários recebem mais avaliações negativas que positivas? Na comunidade BQA, somente 1% das postagens dos ordinários recebem mais avaliações negativas que positivas e 49% recebem pontuação zero. Na comunidade CQA, 5% das postagens dos ordinários recebem mais avaliações negativas que positivas e 54% recebem pontuação zero. Estes resultados sugerem que os ordinários não estão realizando participações indesejadas nas comunidades em questão, considerando que menor parte das participações são negativamente avaliadas.

Tabela 3.10: Média Participação Mensal CQA

Tipo Part.	Grupo	Média Part. Mensal
Perguntas	top 5	0,93
Respostas	top 5	3,01
Comentários	top 5	7,86
Revisões	top 5	7,07
Perguntas	top 5-10	0,94
Respostas	top 5-10	0,38
Comentários	top 5-10	1,64
Revisões	top 5-10	0,84
Perguntas	top 10-15	0,88
Respostas	top 10-15	0,28
Comentários	top 10-15	1,21
Revisões	top 10-15	0,71
Perguntas	ordinários	0,79
Respostas	ordinários	0,20
Comentários	ordinários	0,72
Revisões	ordinários	0,63

3.2.1.8 Exposição do Perfil

Analisando o perfil de cada usuário, foi encontrado que os usuários de destaque proveem mais informações sobre si. Na comunidade BQA, 68% dos usuários top 5, 54% dos usuários top 5-10, 53% dos usuários 10-15 fornecem uma pequena descrição no campo *'about me'* enquanto, 31% dos ordinários concedem esta mesma informação. Ainda na comunidade BQA, 59% dos usuários de destaque em contraste com 32% dos ordinários fornecem informações sobre sua localização. Mais de 30% dos usuários de destaque indicam seu website enquanto 17% dos ordinários colocam esta informação. Por fim, também mais 30% dos usuários de destaque informam sua idade e somente 22% dos ordinários colocam este dado.

Resultados parecidos foram encontrados na comunidade CQA, mostrando que os usuários de destaque possivelmente usam a comunidade para se pro-

Tabela 3.11: Pontuação Postagens dos Usuários de Destaque

Comunidade	Grupo	% Aval. Negativa	% Aval. Neutra	% Aval. Positiva
BQA	top 5	0,24%	50,24%	49,50%
BQA	top 5-10	0,98%	52,24%	46,76%
BQA	top 10-15	0,65%	47,57%	51,76%
BQA	ordinários	1%	49%	50%
CQA	top 5	0,43%	49,59%	49,97%
CQA	top 5-10	1,64%	55,48%	42,86%
CQA	top 10-15	1,50%	52,28%	46,20%
CQA	ordinários	5%	54%	41%

moverem. Os detalhes destas constatações estão nas Tabelas 3.12 e 3.13.

3.2.1.9 Melhor Resposta

Cada resposta de uma questão potencialmente pode ser escolhida como melhor resposta. Em geral, a melhor resposta é aquela que consegue prover informações suficientes para solucionar o questionamento realizado. Como esperado, os usuários de destaque têm mais frequentemente melhores respostas escolhidas, quando comparados com os ordinários tanto na comunidade BQA (0,74 dest. vs 0,26 ord., p<0,01) quanto na CQA (0,7 dest. vs 0,3 ord., p<0,01), conforme na Tabela 3.14.

Além disso, um usuário de destaque tende a continuar postando respostas, mesmo quando uma boa candidata a melhor resposta já está postada. Na comunidade BQA, por volta de 70% das respostas postadas depois de uma forte candidata a melhor, são escritas por usuários de destaque (0,7 dest. vs 0,3 ord., p<0,01). A mesma conclusão é observada na comunidade CQA (0,73 dest. vs 0,27 ord., p<0,01). Isto possivelmente indica que os usuários de destaque querem competir para a seleção da melhor resposta, conforme descrito na Tabela 3.15.

Tabela 3.12: Exposição do Perfil - BQA

Grupo	Informação	Percentual
top 5	'about me'	68,53%
top 5	localização	59,05%
top 5	website	34,05%
top 5	idade	39,87%
top 5-10	'about me'	54,95%
top 5-10	localização	55,60%
top 5-10	website	33,18%
top 5-10	idade	36,63%
top 10-15	'about me'	53,87%
top 10-15	localização	50,64%
top 10-15	website	31,89%
top 10-15	idade	32,54%
ordinários	'about me'	31,34%
ordinários	localização	32,68%
ordinários	website	17,89%
ordinários	idade	22,82%

3.2.2 Perspectiva dos Traços de Linguagem

Neste momento, o foco deste trabalho se volta para a caracterização dos traços de linguagem dos usuários, objetivando averiguar diferenças em seus discursos.

3.2.2.1 Diferenças na Escrita dos Usuários

Os traços de linguagem têm sido amplamente utilizados para identificar brincadeiras[5] em debates [37], vandalismos em websites [105] e revisões falsas em sites de e-commerce [92]. Tudo isto baseado na observação da linguagem expressada por pessoas, isto é, nas palavras usadas tais como pronomes empregados, análise de sentimentos das palavras etc.

Neste trabalho, foram feitas observações similares com os usuários das

[5]Também conhecidas informalmente como 'trollagens'.

Tabela 3.13: Exposição do Perfil - CQA

Grupo	Informação	Percentual
top 5	'about me'	55,94%
top 5	localização	49,83%
top 5	website	22,34%
top 5	idade	36,81%
top 5-10	'about me'	53,13%
top 5-10	localização	49,75%
top 5-10	website	27,12%
top 5-10	idade	39,00%
top 10-15	'about me'	50,16%
top 10-15	localização	45,49%
top 10-15	website	24,11%
top 10-15	idade	31,18%
ordinários	'about me'	21,97%
ordinários	localização	22,63%
ordinários	website	10,83%
ordinários	idade	16,44%

Tabela 3.14: Melhor Resposta

Comunidade	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
BQA	destaque	ordinários	<0,01	0,74	0,26
CQA	destaque	ordinários	<0,01	0.7	0,3

comunidades. Primeiramente, foi medida a legibilidade das postagem dos usuários através da métrica *automated readability index* (ARI) [131]. O ARI produz uma representação aproximada no nível de ensino necessário para compreender um texto. Um alto índice ARI significa uma maior competência necessária para compreender um determinado texto (por ser mais sofisticado). Assim, baseado no ARI, na comunidade BQA os usuários de destaque escrevem textos mais sofisticados que os ordinários (0,57 dest. vs 0,43 ord., p<0,01) e na comunidade CQA também (0,59 dest. vs 0,41 ord., p<0,01).

Na comunidade BQA, não se encontrou diferenças no número de palavras

Tabela 3.15: Postagem de Respostas Depois da Melhor

Comunidade	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
BQA	destaque	ordinários	<0,01	0.7	0,3
CQA	destaque	ordinários	<0,01	0.73	0,27

Tabela 3.16: Principais Diferenças nos Traços de Linguagem - BQA

Traços de Ling.	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
ARI	destaque	ordinários	<0,01	0,57	0,43
Nº Palavras	destaque	ordinários	>0,05	inconclusivo	inconclusivo
Nº Palavras Complexas	destaque	ordinários	>0,05	inconclusivo	inconclusivo
Nº Sentenças	destaque	ordinários	<0,01	0,46	0,54
Nº Caracteres	destaque	ordinários	<0,01	0,49	0,51

e no número de palavras complexas[6] entre os usuários de destaque e os ordinários. Por outro lado, na comunidade CQA, se percebeu que os usuários de destaque utilizam mais constantemente palavras complexas (0,51 dest. vs 0,49 ord., p<0,01) e os ordinários escrevem mais palavras em geral[7] (0,48 dest. vs 0,52 ord., p<0,01). Além disso, os ordinários na comunidade BQA escrevem mais sentenças que os usuários de destaque (0,46 dest. vs 0,54 ord., p<0,01) e o mesmo ocorre na comunidade CQA (0,45 dest. vs 0,55 ord., p<0,01). Os ordinários escrevem mais caracteres na comunidade BQA (0,49 dest. vs 0,51 ord., p <0,01) enquanto na comunidade CQA nenhuma diferença entre os usuários de destaque e ordinários foi constatada. Assim, considerando o número de sentenças, os usuários ordinários parecem ser mais prolixos. Os resultados são apresentados nas Tabelas 3.16 e 3.17.

3.2.2.2 Uso de Pronomes

Observando o comportamento dos usuários de destaque, constatou-se que eles tendem a escrever postagens que são mais centradas em si, isto é, utilizando com maior frequência o pronome 'eu' (BQA: 0,57 dest. vs 0,43 ord.,

[6]Foram consideradas complexas aquelas com mais de 3 sílabas.

[7]Todos os tipos de palavras, incluindo as complexas.

Tabela 3.17: Principais Diferenças nos Traços de Linguagem - CQA

Traços de Ling.	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
ARI	destaque	ordinários	<0,01	0,59	0,41
Nº Palavras	destaque	ordinários	<0,01	0,48	0,52
Nº Palavras Complexas	destaque	ordinários	<0,01	0,51	0,49
Nº Sentenças	destaque	ordinários	<0,01	0,45	0,55
Nº Caracteres	destaque	ordinários	>0,05	inconclusivo	inconclusivo

Tabela 3.18: Pronomes - BQA

Pronome	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Eu	destaque	ordinários	<0,01	0,57	0,43
Você	destaque	ordinários	<0,01	0,65	0,35
Ele/Ela	destaque	ordinários	<0,01	0,68	0,32
Nós	destaque	ordinários	<0,01	0,64	0,36
Eles/Elas	destaque	ordinários	<0,01	0,68	0,32

p<0,01) (CQA: 0,56 dest. vs 0,44 ord., p<0,01). Além disso, os usuários de destaque também utilizam mais frequentemente o pronome 'você' e outros pronomes na terceira pessoa como 'ele/ela' indicando possivelmente que eles tendem a abordar pessoas mais diretamente. Mais detalhes podem ser encontrados na Tabela 3.18 e 3.19.

3.2.3 Perspectiva dos Laços Sociais

Nesta Seção, serão caracterizados os laços sociais dos usuários das comunidades.

3.2.3.1 Estrutura da Rede

Para iniciar o estudo sobre os laços sociais, foram examinadas as interações entre usuários nas discussões das comunidades. Assim, primeiramente, criou-se uma rede de respostas, representada através de um grafo direcionado. Desta forma, se o usuário A posta uma pergunta e, o usuário B responde, então o grafo terá um nó A representando o usuário A e um nó B representando o usuário B. Além disso, esse grafo terá uma aresta que sairá do nó A

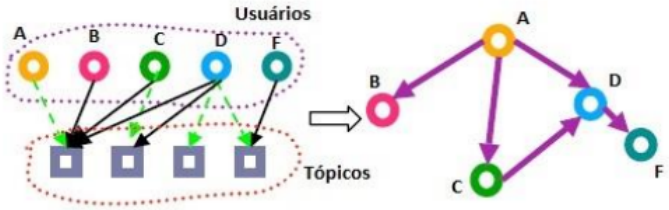
Tabela 3.19: Pronomes - CQA

Pronome	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Eu	destaque	ordinários	<0,01	0,56	0,44
Você	destaque	ordinários	<0,01	0,66	0,34
Ele/Ela	destaque	ordinários	<0,01	0,68	0,32
Nós	destaque	ordinários	<0,01	0,65	0,35
Eles/Elas	destaque	ordinários	<0,01	0,7	0,3

em direção ao B, simbolizando que B respondeu o A.

Esta representação é mostrada na Figura 3.1. As setas em verde (tracejadas) significam que um usuário postou uma pergunta (tópico) e as em preto (linha contínua) significam que um usuário respondeu à pergunta. Do lado direito da Figura é mostrado o grafo correspondente a esse esquema de perguntas e respostas.

Figura 3.1: Grafo



Como nas comunidades analisadas é possível também comentar uma pergunta ou uma resposta, caso um usuário X comente uma pergunta do usuário Y, então uma aresta sairá do usuário Y e chegará no usuário X. Da mesma forma, caso um usuário Z comente uma resposta do usuário K, então uma aresta sairá do usuário K e chegará ao usuário Z.

Uma vez criado os dois grafos (um para cada comunidade), métricas foram extraídas e atribuídas aos nós. Estas foram: (i) o grau de um nó, representando o número de interações distintas de um usuário; (ii) o grau de entrada, representado o número de pessoas distintas que um usuário ajudou;

(iii) o grau de saída, representando o número de pessoas distintas que ajudaram um determinado usuário; (iv) *betweenness*, métrica que mede se um usuário atua como 'ponte', estabelecendo conexões com grupos diversos; (v) *closeness*, métrica que quantifica o quão perto um usuário está dos demais (quanto mais central é o nó, menor é a distância do seu total para todos os outros nós); (vi) *page rank*, métrica que mede a importância de um nó (usuário) em uma rede; (vii) o coeficiente de clusterização, que mede se o nó tende a se agrupar mais com seus vizinhos.

3.2.3.2 Comparando Métricas

Calculadas as métricas dos grafos, comparações foram realizadas nos mesmos moldes da Seções anteriores.

Os resultados sugerem que os usuários de destaque são mais centrais, isto é, mais perto de todos na comunidade BQA (0,77 dest. vs 0,23 ord., p<0,01) e na CQA (0,75 dest. vs 0,25 ord., p<0,01), conforme medido pela métrica de centralidade *closeness*. Além disso, os usuários de destaque são mais importantes, isto é, possuem maior valor da métrica *page rank* (BQA: 0,83 dest. vs 0,17 ord., p<0,01) (CQA: 0,82 dest. vs 0,18 ord., p<0,01). Os usuários de destaque mais frequentemente atuam como 'pontes' nas comunidades, conectando distintos grupos de pessoas como medido pela métrica *betweenness*, que é maior para os usuários de destaque (BQA e CQA: 0,78 dest. vs 0,22 ord., p<0,01). Em geral também, os usuários de destaque interagem com mais distintas pessoas, conforme as centralidades de grau sugerem.

Por fim, foi observados em alguns casos que os ordinários tendem a ser mais conectados com seus vizinhos, como medido pelo coefiente de clusterização (BQA: top 5 0,45 vs 0,55 ord., p<0,01) (CQA: top 5 0,46 vs 0,54 ord., p<0,01). Os resultados de tais análises estão nas Tabelas 3.20 e 3.21.

Tabela 3.20: Métricas Grafo - BQA

Métrica	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Betweenness	destaque	ordinários	<0,01	0,78	0,22
Closeness	destaque	ordinários	<0,01	0,77	0,23
Grau	destaque	ordinários	<0,01	0,55	0,45
Grau Entrada	destaque	ordinários	<0,01	0,82	0,18
Grau Saída	destaque	ordinários	<0,01	0,86	0,14
Page Rank	destaque	ordinários	<0,01	0,83	0,17
Coef. Clusterização	destaque	ordinários	>0,05	inconclusivo	inconclusivo
Coef. Clusterização	top 5	ordinários	<0,01	0,45	0,55
Coef. Clusterização	top 5 10	ordinários	<0,01	0,51	0,49
Coef. Clusterização	top 10 15	ordinários	<0,01	0,56	0,44

Tabela 3.21: Métricas Grafo - CQA

Métrica	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
Betweenness	destaque	ordinários	<0,01	0,78	0,22
Closeness	destaque	ordinários	<0,01	0,75	0,25
Grau	destaque	ordinários	<0,01	0,58	0,42
Grau Entrada	destaque	ordinários	<0,01	0,81	0,19
Grau Saída	destaque	ordinários	<0,01	0,82	0,18
Page Rank	destaque	ordinários	<0,01	0,82	0,18
Coef. Clusterização	destaque	ordinários	>0,05	inconclusivo	inconclusivo
Coef. Clusterização	top 5	ordinários	<0,01	0,46	0,54
Coef. Clusterização	top 5 10	ordinários	>0,05	inconclusivo	inconclusivo
Coef. Clusterização	top 10 15	ordinários	>0,05	inconclusivo	inconclusivo

3.2.3.3 Quem Pede e Quem Fornece Ajuda

Na comunidade BQA, 27% dos usuários somente pedem ajuda, isto é, somente postam perguntas. Ainda nesta comunidade, 14% somente proveem ajuda (repondendo ou comentando). Similarmente, na comunidade CQA, 35% dos usuários somente pedem ajuda e 12% só postam respostas ou comentários. Este resultado possivelmente indica que as comunidades são lugares realmente colaborativos, uma vez que, a maioria das pessoas que participaram pelo menos uma vez tendem a pedir e também fornecer ajuda. Ademais, considerando a representação do grafo da comunidade, os usuários de destaque tendem a ter maior grau de entrada e de saída que os ordinários, indicando que eles interagem com uma maior quantidade de pessoas distintas

Tabela 3.22: Distribuição de Grau

Grau	Mínimo	1ª Quartil	Mediana	Média	3º Quartil	Máximo
Entrada BQA	0	0	1	4.9	2	1534
Saída BQA	0	1	3	4.9	5	441
Entrada CQA	0	0	1	4.7	2	1341
Saída CQA	0	1	3	4.7	5	332

em ambas comunidades.

3.2.3.4 Padrões de Apoio e Ajuda

Uma visão detalhada das interações em uma rede pode ser pela distribuição de graus, tanto o de entrada quanto o de saída, como apresentado na Tabela 3.22. A distribuição de grau foi parecida para ambas comunidades. Em vez de demonstrar padrões igualitários de apoio e ajuda, um pequeno número de usuários que são extremamente ativos proveem ajudas para muitas pessoas (poucos usuários com alto grau de entrada). No entanto, a maioria dos usuários fornece ajuda a um número limitado de pessoas (muitos usuários com baixo grau entrada - 75% com grau de entrada menor que 2 na comunidade BQA, por exemplo). Há também aqueles que recebem ajuda de poucas pessoas (muitos usuários com baixo grau de saída - 75% com grau de saída menor que 5 na comunidade BQA) e um número pequeno de usuários que recebem ajuda de muitos (aqueles poucos com alto grau de saída).

3.2.3.5 Reciprocidade

Foi contabilizado quantos padrões distintos de reciprocidade para cada usuário da comunidade. Ou seja, padrões do tipo: em um instante o usuário A ajuda o usuário B e depois o usuário B ajuda o usuário A. Foi observado, conforme na Tabela 3.23, que os usuários de destaque tendem a apresentar este padrão com mais frequência na comunidade CQA (0,55 dest. vs 0,45 ord., $p<0{,}01$) enquanto nenhuma diferença foi notada na comunidade BQA.

Tabela 3.23: Reciprocidade - CQA

Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
destaque	ordinários	<0,01	0,55	0,45
top 10 15	ordinários	<0,01	0,59	0,41
top 5	ordinários	>0,05	inconclusivo	inconclusivo
top 5 10	ordinários	<0,01	0,56	0,44
top 5	top 10 15	0,0217	0,44	0,56
top 5 10	top 10 15	>0,05	inconclusivo	inconclusivo
top 5 10	top 5	>0,05	inconclusivo	inconclusivo

Tabela 3.24: Subcomunidades

Comunidade	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
BQA	Subcom. tem usuário de destaque	Subcom. só ordinários	<0,01	0,91	0,09
CQA	Subcom. tem usuário de destaque	Subcom. só ordinários	<0,01	0,93	0,07

3.2.4 Perspectiva da Influência

Neste momento, será analisada a influência dos usuários nas comunidades BQA e CQA.

3.2.4.1 Atraindo o Público

Foi utilizada a representação em grafos da comunidade, conforme explicada na Seção 3.2.3, para detectar as subcomunidades [24] com pessoas que com mais frequência interagem entre si. Na comunidade CQA foram detectadas 58 subcomunidades e 42 na comunidade BQA. Como resultado, sempre quando há usuários de destaque em alguma das subcomunidades, seu número de membros tende a ser maior (BQA: 0,91 tem usuário dest. vs 0,09 não tem usuário dest., p<0,01) (CQA: 0,93 tem usuário dest. vs 0,07 não tem usuário dest., p<0,01). Entretanto, a maioria das subcomunidades têm somente ordinários. Os resultados desta análise estão na Tabela 3.24.

Tabela 3.25: Citações

Comunidade	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
BQA	destaque	ordinários	<0,01	0,74	0,26
CQA	destaque	ordinários	<0,01	0,75	0,25

3.2.4.2 Citações

Nas comunidades examinadas, é possível citar um usuário utilizando o padrão '@username' em uma postagem. Os usuários de destaque são mais citados que os ordinários e isto consiste em uma importante noção de popularidade [36]. Na comunidade CQA, a maior parte das citações se refere aos usuários de destaque (0,75 dest. vs 0,25 ord., p<0,01). Este resultado se replica para a comunidade BQA (0,74 dest. vs 0,26 ord., p<0,01). Os resultados podem ser visualizados na Tabela 3.25.

3.2.4.3 Tamanho das Discussões

A análise do tamanho da discussão, isto é, o total de postagens relacionadas a uma pergunta foi usada para investigar a influência dos usuários de destaque nelas. Em ambas comunidades, por volta de 95% dos casos, sempre que pelo menos um usuário de destaque participa de uma discussão, seu tamanho tende a ser maior. Nas comunidades, o tamanho médio de uma discussão é 6 quando há pelo menos um usuário de destaque, e 2 quando há somente ordinários.

3.2.5 Perspectiva do Foco

Toda discussão das comunidades tem *tags* que descrevem sua categoria. Para capturar como as participações dos usuários são focadas nas categorias das comunidades, foi utilizada a métrica entropia [2]. Quanto mais agrupadas forem as participações de um usuário em determinada categoria, menor é sua entropia e maior seu foco.

Além disso, se deseja que a entropia capture a organização das categorias, visto que, uma discussão pode pertencer a diversas categorias (pode ter várias *tags*) onde uma complementa a outra. Assim, um usuário que participa em várias categorias, porém, cada uma destas categorias é complementada por uma única outra categoria, então, este usuário terá entropia menor que aqueles que participam no mesmo número de categorias, porém, complementares. Foi verificado que algumas categorias são amplamente utilizadas e outras não. Desta forma, as categorias foram agrupadas em quatro níveis. O primeiro nível é aquele que contém as categorias mais o populares, o segundo nível com as categorias não tão populares quanto a do primeiro, porém mais populares que aquelas do terceiro nível. Por fim, o quarto nível com as menos populares.

A definição da entropia de nível é dada na fórmula (a). Imagine um usuário que postou 10 respostas na comunidade BQA. Considere que 3 respostas foram relacionadas à categoria botânica e 7 relacionadas à categoria genética, sendo ambas do primeiro nível. Assim o 'P' da fórmula (a) para este usuário na categoria botânica é 0,3, pois, 3 em 10 respostas foram desta categoria. Da mesma maneira, o 'P' para a categoria genética é 0,7, pois, 7 em 10 respostas foram para esta categoria. Assim, conforme a fórmula (c), a entropia do primeiro nível deste usuário pode ser calculada. Por fim, após o cálculo da entropia de cada nível, é calculada a entropia total do usuário como apresentado na fórmula (b).

$$(a)\ E_L = -\sum_i P_{L,i} * \ln(P_{L,i}) \qquad (b)\ E_T = \sum_L E_L$$

$$(c)\ E_1 = -((0.3 * \ln(0.3)) + (0.7 * \ln(0.7))) = 0.61$$

No geral, os usuários de destaque são menos focados. Estes possuem

Tabela 3.26: Foco

Comunidade	Grupo 1	Grupo 2	p	Tam. Efeito Grp. 1	Tam. Efeito Grp. 2
BQA	destaque	ordinários	<0,01	0,88	0,12
CQA	destaque	ordinários	<0,01	0,84	0,16

maior entropia que os ordinários (BQA 0,88 dest. vs 0,12 ord., p<0.01) (CQA 0,84 dest. vs 0,16 ord., p<0,01), conforme na Tabela 3.26. Isto pode indicar que foco não é um fator essencial para estar no topo das comunidades analisadas. Além disso, alguém pode esperar que usuários que são focados em um número limitados de tópicos tendem a ter mais frequentemente respostas escolhidas como melhores. Pode-se esperar que um usuário que gosta de botânica e responda a perguntas desta categoria tenha maior proporção de melhores respostas, pois todas as suas respostas estão focadas em sua especialidade. Os usuários não fornecem respostas melhores (pelo menos de acordo com a melhor contagem de melhores respostas) quando se especializam, pois encontramos uma correlação moderada entre a entropia do usuário e o número de melhores respostas quando consideramos os usuários de destaque (correlação $\approx$ 0,6, p<0,01).

3.2.6 Modelo de Classificação

Nas Seções anteriores, cada tipo de usuário foi descrito, por meio de análises e comparações. Tomando tais análises como base, seria interessante ter ferramentas automáticas que identificassem os tipos de usuários ou postagens com potencial para serem escolhidas como melhores respostas. Assim, essas ferramentas poderiam, de alguma forma, apoiar os usuários durante sua trajetória de aprendizagem em comunidades online, com possivelmente uma menor dependência das avaliações manuais hoje existentes. Neste sentido, foram consideradas duas propostas de classificações: (i) Podemos distinguir

um usuário de destaque dos ordinários? (ii) Podemos identificar a melhor respostas de uma questão?

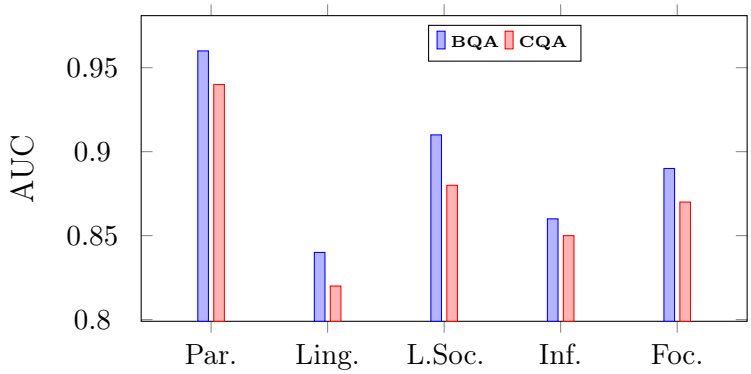
Para isto, foi usado o classificador *Stochastic Gradient Boosting* [53]. Em síntese, este classificador combina diversas árvores de decisão (denomindas classificadores fracos) em um único (classificador forte) de maneira interativa. Os classificadores fracos geram classificações 'imperfeitas' e estes, por sua vez, são utilizados na construção do classificador forte, com um modelo com classificações mais precisas. Além disso, foram selecionados 60% dos dados disponíveis para constituir o conjunto de treinamento e o restante o conjunto de teste. Também foi aplicada a validação cruzada *k-fold* (k = 5) para evitar o *over-fitting*, que é uma técnica de uso habitual em classificações desta natureza [20].

Foram escolhidas também algumas características dos usuários relacionadas a cada uma das perspectivas discutidas: (i) na pespectiva da participação, foram selecionados o número de perguntas, respostas, comentários e revisões de cada usuário; (ii) na perspectiva dos traços de linguagens características foram selecionados o ARI, número de caracteres, número palavras comuns e complexas, número de sentenças, e o número de vezes que pronomes são utilizados; (iii) na dos laços sociais, características como o grau de entrada, saída, *page rank*, *betweenness*, *closeness* e coeficiente de clusterização foram escolhidas; (iv) na da influência, a participação média em discussões e citações; e (v) na do foco, a entropia e número de participações em diferentes categorias. Estas caracterísicas foram as entradas do modelo de classificação.

3.2.6.1 Encontrando Usuários de Destaque

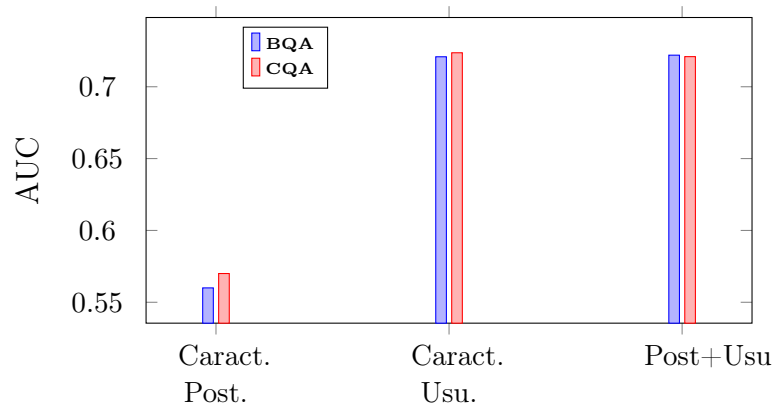
A Figura 3.2(a) mostra os resultados obtidos individualmente por cada conjunto de características. Foi observado que a participação tende a ser mais preditiva, para detectar os tipos de usuários. As comunidades analisa-

Figura 3.2: Classificação - AUC



(a) Classificação usuário

Par. = participação, Ling. = traços de linguagem,
L.Soc. = laços sociais, Inf. = influência, Foc. = foco



(b) Identificação da melhor resposta

Caract. Post. = características da postagem,
Caract. Usu. = características do usuário,
Post+Usu = características do usuário e da postagem

das utilizam mecanismos de gamificação que podem estimular a participação. Porém, somente uma alta participação não é suficiente para estar entre os melhores da comunidade. Outras características obtiveram boas classificações, isto é, alto AUC (>0,8). Por exemplo, se pode classificar alguém como de destaque considerando características relacionadas ao foco em tópicos ou seus traços de linguagem.

3.2.6.2 Generalização dos Classificadores

Nos estudo das perspectivas, pode-se perceber que as duas comunidades apresentam características parecidas. Por exemplo, a diferença entre cada tipo de usuário se replica em ambas comunidades na maior parte dos casos. Embora sejam comunidades distintas, as dinâmicas delas são similares. Assim, criamos uma classificação entre domínios (*cross-domain classification model*), para identificar os tipos de usuários. Para criar este modelo, foi selecionado 60% dos dados de uma comunidade para ser o conjunto de treinamento. Depois do treino, os dados da outra comunidade foram utilizados como conjunto de testes. Os resultados encontrados, que estão na Tabela 3.27, são comparáveis ao da Figura 3.2 (a), sugerindo não somente o uso de distintas características, mas também uma possível generalização deste modelo de classificação para comunidades como as analisadas.

3.2.6.3 Identificando a Melhor Resposta

A Figura 3.2(b) apresenta os resultados da predição da melhor resposta. Para esta tarefa, consideramos dois tipos de entradas para o classificador: (i) as características das postagens, isto é, os traços de linguagem de uma postagem especificamente; (ii) as características dos usuários que escreveram as postagens. Foi verificado que somente usando as características das postagens, o classificador tende a resultar em classificações mais pobres. Por outro lado, quando utilizadas as características dos usuário sozinhas ou combinadas com as características das postagens, as predições tendem a ser melhores (ou seja, AUC>0,7).

Tabela 3.27: Validação Entre Comunidades do Classificador (AUC)

		Treinado em + Conjunto de Características									
		Participation		Ling. Traits		Social Ties		Influence		Focus	
		BQA	CQA	BQA	CQA	BQA	CQA	BQA	CQA	BQA	CQA
Teste em	BQA	0,96	0,9	0,83	0,82	0,90	0,91	0,86	0,83	0,89	0,88
	CQA	0,9	0,94	0,81	0,82	0,88	0,88	0,80	0,85	0,85	0,87

3.3 Discussão

Depois deste estudo, é possível discutir os resultados encontrados. Para isto, as questões de pesquisas enunciadas na Seção 3.1.1 são respondidas. Na sequência, são discutidas as principais contribuições e as limitações do estudo em questão.

3.3.1 Resposta às Questões de Pesquisas

***Q_1*: Quais são as diferenças entre os usuários que se destacam com relação aos ordinários?**

Alguns dos resultados encontrados são comumente citados em trabalhos relacionados. Por exemplo, participação é essencial para a alcançar o destaque [165]. A participação é importante para o processo de ensino e aprendizagem [127, 5, 99], seja formal ou informal, como é o caso das comunidades deste estudo. O usuário participativo, que se destaca, é parecido com o bom aluno também participante nos ambientes educacionais tradicionais. Em geral, se aprende e ensina expressando, seja por meio de uma postagem ou qualquer outra forma, para que possa ser lido, entendido, debatido e internalizado.

Ainda sobre a participação, foi constatado que os usuários de destaque tendem a ser mais participativos ao longo do tempo. Entretanto, como visto, há casos que o esforço realizado pelos usuários ordinários, quando observado mês a mês, são parecidos com alguns usuários de destaque. Este esforço ex-

tra, quando observado na linha do tempo, faz toda diferença. Alguém pode argumentar que mais usuários de destaque entraram antes nas comunidades e, por isso, tiveram mais tempo para participar. De fato, alguns tiveram mais tempo sim, porém, se percebe que uma parcela pequena deles entraram logo no início da comunidade e os demais tiveram suas contas criadas ao longo dos meses. Não se pode concluir que todos os usuários de destaque tiveram mais tempo para estar no topo. Pelos estudos, estes parecem ser mais engajados quando as participações são analisadas mensalmente e também quando se mede o tempo entre o primeiro acesso e a primeira participação. Os usuários de destaque esperam menos para iniciar suas participações. Possivelmente, tais resultados estão alinhados com a teoria contrutivista [154], que argumenta que o conhecimento vem principalmente das interações.

Uma outra questão interessante é que uma parcela significativa das participações recebe avaliações neutras, significando que não basta somente participar. Ou seja, a participação, em princípio, deve ser no mínimo útil para alguém avaliá-la positivamente. Neste sentido, participar por participar não garante que alguém se destaque. Por outro lado, quanto mais alguém participa, mais chances tem de ser avaliado, seja positivamente ou negativamente. Supostamente, os usuários ordinários podem saber menos ou podem ter menos compromisso com a comunidade. Neste sentido, hipoteticamente, se pode imaginar que há um grande número de avaliações negativas para tais usuários. Porém, isto não ocorre. Em boa parte dos casos, os ordinários não recebem avaliações negativas, embora as avaliações neutras sejam boa parte delas.

Além disso, um número maior de usuários de destaque parece utilizar a comunidade para se promover, expondo mais informações sobre si em seu perfil. Um outro dado interessante é que os usuários de destaque são mais

competitivos: sempre que há uma resposta com forte possibilidade de ser escolhida como melhor, eles tendem a continuar postando mais frequentemente, possivelmente com a expectativa de ter sua resposta escolhida como melhor. Apesar da competição ser algo condenado por alguns pesquisadores [62], esta parecer existir em ambientes de aprendizagem, seja formal ou informal.

Com relação aos traços de linguagem, as análises revelaram diferenças entre os usuários de destaque e os comuns. Nestas análises, o tamanho de efeito foi menor quando comparadas, por exemplo, às análises das participações, embora ainda significativa, como verificados pelos testes de inferência aplicados. Em geral, os usuários de destaque possuem um discurso mais sofisticado e tendem a ser mais centrados em si. Por outro lado, os ordinários escrevem mais sentenças, sendo possivelmente mais prolixos.

Sobre os laços sociais, foi verificado que os usuários de destaque tendem a interagir com mais pessoas distintas. Em geral, são mais centrais, isto é, estão mais perto de todos, além de conectar grupos diferentes existentes dentro da própria comunidade. De maneira geral, estes resultados estão em consonância com a teoria conectivista, que argumenta que saber navegar na rede de informações é tão importante quanto 'absorvê-las' [137, 135]. Tal teoria afirma que hoje as informações são abundantes e as teorias de aprendizagem anteriores não consideram este fato, por serem do mundo pré-digital. Assim, as teorias pré-digitais estão mais interessadas no processo de aprendizagem em si do que na avaliação do que aprender [137]. Em síntese, segundo o conectivismo, dominar muitos conceitos profundamente é difícil e uma meta-habilidade importante é saber e entender como sufar em uma rede de informações ou conhecimentos, para adquirí-los conforme a necessidade. Sem objetivo de discutir as vantagens e desvatagens de cada teoria, de alguma forma, os usuários de destaque ficam mais bem posicionados nessa rede

de interações das comunidades. Isto possivelmente os colocam em alguma vantagem sob a perspectiva da aprendizagem conectivista.

Como esperado, seria fácil presumir que os usuários de destaque exercem alguma influência na comunidade. Entretanto, neste trabalho, isto foi quantificado. Foi verificado, através de um algoritmo que detecta subcomunidades, que os usuários de destaque funcionam como 'polos gravitacionais', atraindo público para os lugares onde eles estão posicionados. Nas subcomunidades onde eles estão presentes, o número de membros destas tende a ser maior. Além disso, eles são mais citados que os ordinários e, sempre que estão presente em uma discussão, o tamanho desta tende a ser maior. Tais fatos demonstram a capacidade que estes têm de mobilizar as comunidades envolvidas, mesmo que involuntariamente.

Com relação ao foco, foi verificado que os usuários de destaque participam em uma maior variedade de tópicos. A partir disto, conclui-se que eles têm menos foco, dentro da temática da comunidade. Ademais, os usuários de destaque não têm maior número de melhores respostas quando eles se especializam. Alguém poderia imaginar o contrário, visto que, uma pessoa que estuda determinado assunto com mais foco, tenderia a ter melhores respostas associadas a este assunto. No entanto, as evidências apontam para o lado oposto.

***Q_2*: Qual é o conjunto de características mais preditivas quando se classifica um usuário em 'de destaque' ou ordinário?**

Nas análises realizadas, foram testados diversos conjuntos de características para identificar automaticamente o tipo de usuário. Encontrou-se que a participação é a característica mais preditiva. Este resultado está em sintonia com outros trabalhos [165, 119], que afirmam que características simples tendem a ser mais preditivas. Embora isto seja verdade, há outros conjuntos

de características que podem ser úteis para a detecção do tipo de usuário, isto é, com uma boa assertividade. Foi verificado que as características associadas às outras perspectivas (traços de linguagem, laços sociais, influência e foco) resultam também em boas classificações.

Em especial, estas classificações comprovam que as análises das diferenças estão corretas. Pois, se através das características estudadas é possível chegar a boas classificações, é sinal que o objetivo foi atingido: entendemos o que há por trás do destaque e fecha-se o ciclo.

***Q_3*: O quão generalizáveis são os modelos que classificam os usuários em 'de destaque' ou ordinário?**

Foi averiguado que a construção de um modelo de classificação em uma comunidade pode levar a boas classificações em outra comunidade. Neste ponto reside uma importante contribuição desta pesquisa, sugerindo que a atuação dos usuários de destaque são parecidas. Isto também é corroborado por pesquisas que apresentam como características mais preditivas as mais simples, como comentado na questão Q_2.

De forma alguma se deseja concluir que o classificador proposto seja universal, isto é, plenamente possível utilizá-los nas mais diversas comunidades. No entanto, já há uma indicação neste sentido, tendo em vista que as análises de domínio cruzado mostraram boa preditividade.

***Q_4*: A predição da melhor resposta de uma pergunta postada na comunidade está mais relacionada com as características das postagens da trilha ou com as características dos usuários participantes?**

Encontramos que somente considerar as características das postagens, isto é, os traços de linguagem somente delas, não levam a boas predições da melhor resposta. Em vez disto, é essencial considerar as características dos usuários envolvidos em uma discussão para se obter bons resultados. Além

disso, a combinação entre características das postagens com as características dos usuários também resulta em boas predições, significando que as características das postagens pouco influenciam na classificação nestas condições.

Esse resultado corrobora a importância do estudo dos padrões comportamentais dos membros das comunidades online. Estes, possivelmente, podem ser usados para encontrar, organizar e fornecer métodos para dar suporte aos usuários antes do início do aprendizado nas comunidades. Espera-se que os resultados, no futuro, ajudem a conectar os usuários a comportamentos valiosos (comportamento dos de destaque) ou locais (discussões que estão quase chegando a um consenso devido a presença de possíveis melhores respostas).

3.3.2 Contribuições

As descobertas deste estudo contribuem para entender como os usuários comuns e de destaque atuam na prática em comunidades de perguntas e respostas. Mais especificamente, como principal contribuição, pode-se citar o desenvolvimento de uma metodologia robusta para caracterizar e identificar os tipos de usuários, através do exame detalhado de suas atividades, sob diversas perspectivas. Assim, primeiramente, estudos foram conduzidos de forma a analisar os comportamentos dos usuários. Tais análises revelaram as diferenças entre os usuários de destaque e os ordinários. Acima de tudo, ficam demonstradas as dinâmicas de funcionamento das comunidades, através de evidências fortes e devidamente validadas.

Como contribuição adicional, se pode citar os modelos de classificação propostos para identificar os tipos de usuários e as melhores respostas. Verificou-se que as características relacionadas à perspectiva da participação tende a ser mais preditiva para a classificação de tipos de usuários, corroborando trabalhos anteriores. Porém, como diferencial, demonstrou-se que características de outras perspectivas também levam à boas classificações dos tipos de

usuários. Além disso, o modelo de classificação de tipos de usuários possuem capacidade de generalização, considerando o escopo analisado, conforme discutido na Seção 3.3. Com relação à classificação das melhores respostas, ficou demonstrado que somente considerar os traços de linguagem de uma postagem isoladamente não levam à boas predições. Assim, mostrou-se que deve-se considerar as características dos usuários para a realização de tal tarefa.

Espera-se que este estudo também possa ser utilizado em pesquisas de *Learning Analytics*, em especial, no contexto informal. Embora não seja o foco principal deste trabalho, pesquisadores em *Learning Analytics* argumentam que há carência de métodos que apoiem e entendam o processo de aprendizagem em ambientes informais [136]. Além disso, muitas das pesquisas existentes neste sentido focam em poucas perspectivas associadas aos usuários (no caso, estudantes) [134, 28]. Acredita-se que este estudo possa ser um bom ponto de partida para análises em ambientes de aprendizagem.

3.3.3 Limitações

Como todo trabalho científico, este também tem suas limitações. A primeira é relacionada aos dados utlizados. Existem várias outras comunidades online que poderiam ser analisadas neste trabalho. Assim, através de mais análises, poderíamos ter mais certeza sobre os resultados obtidos e indicar mais precisamente o que é comum ou diferente em comunidades.

A segunda limitação é com relação as análises das perspectivas, que ainda podem ser mais aprofundadas. Este trabalho se limitou em ressaltar as diferenças entre os tipos de usuários. Por exemplo, quando analisamos a perspectiva da participação, podem parecer questões satélites como: (i) Será que os usuários estão mais interessandos na pontuação ou em compartilhar conhecimento? (ii) Querem somente se promover ou são realmente altruístas?

Para responder à tais questões, seria necessário extrapolar o escopo proposto para este estudo. Ainda citando casos de extensão desta segunda limitação, considerando a perspectiva do foco, o ideal teria sido organizar toda a hierarquia de categorias da comunidade previamente. Assim, saberíamos com mais precisão, por exemplo, se uma pessoa fala de uma 'organela celular', ela também fala de 'célula'. Entretanto, apesar de assumirmos algumas simplificações para o cálculo do foco, os resultados corroboram as conclusões de um estudo feito no Yahoo! Answers [2] o que nos deixa em uma posição mais confortável.

Uma terceita limitação tem relação ao classificador utilizado (*Stochastic Gradient Boosting*). Há diversos outros classificadores que podem ser usados para o mesmo fim. Optou-se nesta pesquisa pelo uso de um único, para que todas as classificações realizadas, isto é, os resultados, não ficassem dependentes do tipo de classificador. Estas limitações devem ser tratadas em trabalhos futuros.

3.4 Comentários Finais

Este Capítulo teve como objetivo apresentar um estudo das características dos usuários das comunidades. Em síntese, como colocado no Capítulo 1, tal estudo também é conhecido como processo de *feature engineering*.

No início, se apresentou a descrição geral do estudo, com as questões de pesquisas, dados utilizados, ameaças de validade (bem como formas de minimizá-las) e os mecanismos de análises. Muitas comparações foram realizadas, mostrando explicitamente o que usuários fazem para alcançar o destaque. Por fim, modelos de classificações foram propostos e as questões de pesquisas foram amplamente discutidas.

Acima de tudo, o Capítulo buscou trazer uma definição mais concreta

sobre destaque. A noção de destaque pode ser vaga, isto é, sua definição pode ser muito subjetiva. No entanto, se buscou entender 'o que está por trás do destaque', elencando comportamentos específicos daqueles que têm notoriedade em uma comunidade online.

Por fim, mais detalhes técnicos da implementação dos experimentos estão nos Apêndices A e B.

4. Aprendizagem das Características dos Usuários

Este Capítulo objetiva mostrar as propostas de *feature learning* deste trabalho. Tais propostas tem como finalidade complementar estudos correlatos, como o realizado no Capítulo 3, porém, apresentando uma outra proposta de solução para o mesmo problema.

Em síntese, *feature learning* é uma forma automática para aprender características, inclusive já existente na literatura. No entanto, o uso de propostas de *feature learning* ao problema endereçado nesta tese são escassas. Assim, o Capítulo visa contribuir neste ponto, mostrando como pode ser útil o uso destas técnicas no problema enunciado na Seção 1.3. Além disso, comparações entre abordagens de *feature learning* são realizadas, a fim de demonstrar quais são as mais promissoras no contexto do trabalho.

4.1 Definições Iniciais

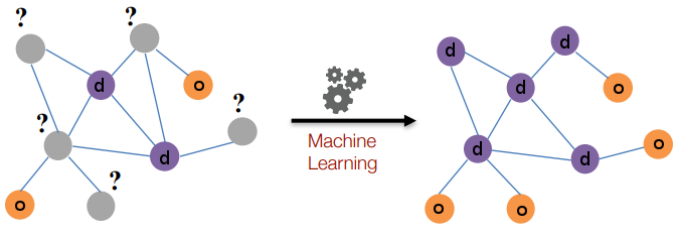
É comum que as interações entre pessoas em comunidades online sejam representadas através de um grafo [114, 165]. Em geral, como na Seção 3.2.3, os usuários são os nós do grafo e as interações são representadas pelas arestas. Diversas tarefas de classificação em grafos envolvem descobrir a categoria de

um nó ou aresta. Por exemplo, na Figura 4.1, é ilustrada uma típica tarefa de classificação de nós. Do lado esquerdo, há alguns nós com o indicador '?', mostrando que estes não pertencem a alguma categoria. No lado direito, após o uso de algum modelo de classificação (*Machine Learning*), tem-se a categoria desejada atribuída ao nó. Conectando a Figura 4.1 com o contexto deste trabalho, os nós com 'd', por exemplo, poderiam ser os usuários de destaque e os nós com 'o' os ordinários (comuns).

Como demonstrado no Capítulo 3, para se concluir que alguém é de destaque ou não, há a necessidade de se analisar os dados disponíveis dos usuários (*feature engineering*) para depois propor um modelo de classificação. No caso do estudo deste Capítulo, se optou por uma solução diferenciada, que consiste em propostas de métodos que automaticamente captem as características dos usuários (*feature learning*), sem a necessidade de uma análise manual. A ideia geral da proposta deste Capítulo está na Figura 4.2, onde é apresentado o que se pretende realizar neste trabalho.

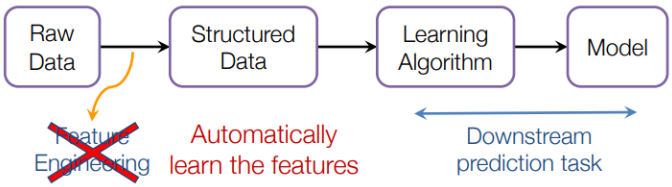
Detalhando a Figura 4.2, nas abordagens de *feature engineering*, deve-se analisar as características de cada usuário (baseado nos dados disponíveis da comunidade online), tais como nível de participação ou traços de linguagem e, depois, estruturá-las para estabelecer as distinções entre cada tipo de usuário através, por exemplo, do uso de testes de inferência estatística. Uma vez sabendo as características que permitam diferenciar os usuários, estas serão as entradas de algoritmos classificação (*Machine Learning* ou *Learning Algorithm* como na Figura 4.2). Estes algoritmos podem ser as Redes Neurais Artificiais, Árvores de Decisão entre outros, e as suas entradas geralmente são números reais ou inteiros (isto é, as características dos usuários representadas por números). No final da execução destes algoritmos, gera-se um modelo que permite realizar classificações futuras. Reiterando, *feature*

Figura 4.1: Classificação de Nós



Fonte: How Powerful are Graph Neural Networks[1]

Figura 4.2: Feature Engineering x Feature Learning



Fonte: Graph Representation Learning[2]

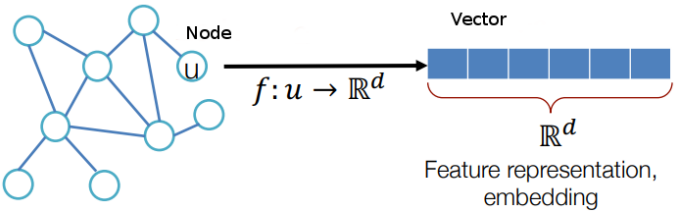
engineering foi o processo realizado no Capítulo 3.

A proposta de *feature learning* vem para contrapor a de *feature engineering*. A ideia por trás da proposta reside em aproveitar a representação em grafos de comunidades (como descrito na Figura 4.1 e na Seção 3.2.3), estimular 'voltas nos caminhos do grafo' e, a partir disto, gerar um vetor de números que serão as entradas para o algoritmo de classificação. A Figura 4.3 ilustra em alto nível a proposta de *feature learning*.

Dada esta ideia geral, neste Capítulo serão propostos dois métodos de *feature learning*: (i) o *Simple Walk*; (ii) e, o *Go Ahead When Necessary*. Mais adiante será apresentando com mais detalhes como estes métodos de *feature learning* funcionam, de maneira a demonstrar suas diferenças e como podem auxiliar na detecção de tipos de usuários.

[1] `https://stanford.io/2PgB72l`
[2] `https://bit.ly/2V401rm`

Figura 4.3: Representação da Feature



Fonte: Graph Representation Learning[3]

4.1.1 Origem do Feature Learning em Grafos

Recentes avanços no processamento de linguagem natural, possibilitaram a existência deste estudo. Contextualizando, o *Skip-gram model* [87] se trata de uma famosa abordagem que converte palavras de um documento em um vetor de números. A ideia geral deste modelo é: (i) ler todas as palavras de um documento; (ii) depois, representar cada palavra como um vetor de números; (iii) por fim, predizer quais são as palavras mais prováveis de serem encontradas ao lado de outra.

Colocando em outra maneira, o *Skip-gram model* pode aprender a representar as palavras como um vetor de números, através do método *Stochastic Gradient Descent* [88]. Este método é semelhante a uma rede neural artificial *feedforward* com uma única camada intermediária [61]. Para fins didáticos, a explicação do método será em função da rede neural. Exemplificando, tendo palavras como entradas, uma rede neural é treinada, porém, o objetivo principal é somente obter os pesos da camada intermediária. Esse conjunto de pesos é o vetor de números que representam a palavra. Depois do treino, a rede deve ser capaz de dizer a probabilidade de uma palavra ser 'vizinha' (próxima) de outra. Por exemplo, considere um documento que fale sobre os Estados Unidos. Assim, para a palavra 'Obama', a probabilidade tende a

[3] `https://bit.ly/2V401rm`

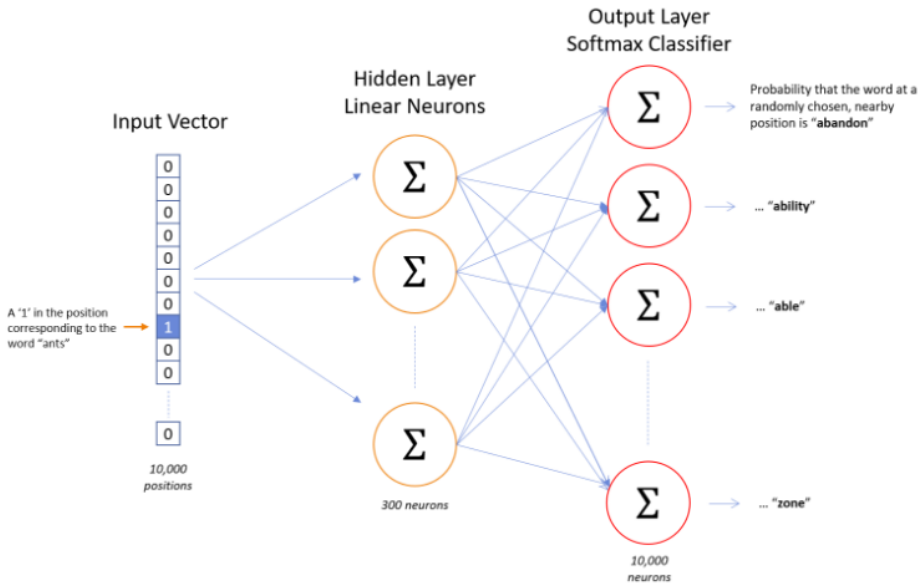
ser maior para as palavras 'Barack' ou 'presidente' do que para outras não relacionadas. Para isto, como camada de saída, esta rede neural usa uma função denominada softmax[4], que retorna probabilidade de uma determinada palavra (entrada da rede) ser vizinha de todas as outras do documento analisado.

Obviamente, isto tudo não ocorre magicamente. Entrando em detalhes mais técnicos, primeiro, não se pode alimentar uma rede neural com palavras textuais. Consequentemente, há a necessidade de ter uma representação inicial para as palavras de entrada. Imagine que um texto tenha 10.000 palavras distintas e, uma delas, é a palavra 'formigas' ('ants' em inglês). Cada entrada da rede neural será um vetor de 10.000 posições, ou seja, o número de posições corresponde à quantidade de palavras do texto inicial. Assim, por exemplo, para a palavra 'formigas' haverá uma posição do vetor preenchida com '1' e as demais com '0' (zero). Esta mesma ideia se aplica para todas as palavras do texto, alternando onde é preenchido com '1', de forma que cada vetor represente unicamente cada palavra. A camada de saída desta rede conterá 10.000 componentes, isto é, um componente para cada palavra do texto, representando a probabilidade da palavra de entrada (no caso, 'formigas') ser encontrada ao lado das outras, conforme a Figura 4.4. Esta camada de saída, como já comentado, é a função softmax.

Um vez entendido como é a entrada e a saída da rede, uma pergunta que surge é: como funciona a camada intermediária? Considere que gostaríamos de representar a palavra 'formigas' como um vetor de número de 300 posições (não confundir com a representação inicial do vetor de 10.000 posições). Isto é, queremos que a palavra 'formigas' seja representada por 300 números. Desta maneira, tem-se 300 componentes (também conhecidos como neurô-

[4] `https://en.wikipedia.org/wiki/Softmax_function`

Figura 4.4: Skip-gram Model



Fonte: Internet[5]

nios) na camada intermediária, conforme na Figura 4.4. Em linhas gerais, a camada intermediária será representada por uma matriz de 10.000 linhas (uma linha para cada palavra) e 300 colunas (representando os neurônios). Esta é comumente conhecida como matriz de pesos. Em síntese, cada linha desta matriz será preenchida com os valores numéricos que representará cada palavra, que é o objetivo principal do método. Uma questão que pode parecer é: qual a relação do vetor de entrada, aquele que é quase todo preenchido com zero, com a matriz 10.000 x 300 ? A relação é simples, pois, quando se multiplica-se o vetor 1 x 10.000 de entrada pela matriz 10.000 x 300, como resultado, se obtém a representação da palavra (pesos aprendidos). Na Figura 4.5, é apresentado exemplo simplificado desta relação, onde a palavra de entrada, inicialmente representada por um vetor 1 x 5, é multiplicada por uma matriz 5 x 3, oriunda da camada intermediária. Em resumo, a matriz

[5] `https://bit.ly/2ZkFLBa`

Figura 4.5: Multiplicação Matrizes

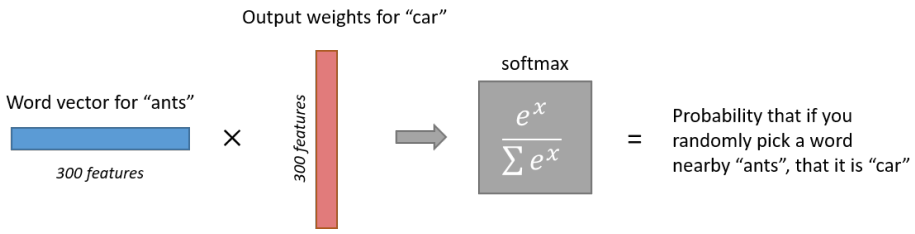
$$[0 \quad 0 \quad 0 \quad 1 \quad 0] \times \begin{bmatrix} 17 & 24 & 1 \\ 23 & 5 & 7 \\ 4 & 6 & 13 \\ 10 & 12 & 19 \\ 11 & 18 & 25 \end{bmatrix} = [10 \quad 12 \quad 19]$$

serve para conectar a representação inicial da palavra, com a representação aprendida pela rede, isto é, os pesos.

Até então, tem-se o fluxo: entram-se com as palavras na rede, calculam-se os pesos e os colocam na linha correspondente da matriz. Uma vez calculado os pesos, estes são as entradas para a camada de saída da rede neural. Considerando o exemplo inicial, um vetor 1 x 300 (linha da matriz) corresponderá a representação aprendida da palavras 'formigas'. Este vetor 1 x 300 será a entrada para cada componente da camada de saída. Especificamente, cada componente da camada de saída tem acesso a representação dos pesos aprendidos das outras palavras. Assim, os pesos vindos da camada intermediária são multiplicados pelos pesos obtidos das outras palavras. Depois disto, o valor desta multiplicação é aplicado à função softmax, que retorna um valor entre 0 e 1, representando a probabilidade de uma palavra ser encontrada ao lado de outra. A Figura 4.6, mostra como é calculada a probabilidade da palavra 'formigas' ser encontrada ao lado da palavra 'carro' (em inglês, 'car').

Mas como a rede sabe que uma palavra vai ser encontrada ao lado de outra? A resposta está no conjunto de treinamento. Em um texto, é comum que certas palavras apareçam juntas. Assim, a medida que as palavras de um texto vão sendo processadas pelo modelo apresentado, as amostras de pareamentos mais frequentes (por exemplo, 'Obama' e 'presidente' ou 'formigas' e 'doce') tenderão a ter probabilidade maior na camada de saída.

Figura 4.6: Softmax



Fonte: Internet[6]

Explicando melhor, o *Skip-gram model* trabalha com o conceito de tamanho de janela do texto que, em resumo, são fragmentos do texto. Por exemplo, considere a frase: ‘O presidente Obama é do partido democrata’. Neste caso, um tamanho de janela possível é três, onde fragmentos como ‘O presidente Obama’, ‘presidente Obama é’ ou ‘é do partido’ são possibilidades de janelas do referido tamanho. Assim, a rede neural tentará maximizar a probabilidade de palavras comumente encontradas em uma mesma janela ao longo das iterações. Concluindo, o modelo é capaz de captar quais palavras aparecem subsequentemente a outra, isto é, à medida que estas são processadas, comumente uma entra no modelo depois ou antes da outra. As explicações dadas sobre o *Skip-gram model* foram adaptadas de diversas fontes [85, 88, 87], onde mais detalhes podem ser obtidos.

Neste trabalho, foi adaptado o *Skip-gram model*, onde as palavras passaram a ser os nós e as palavras vizinhas, os nós vizinhos. Desta forma, para cada nó do grafo da comunidade, estimulamos voltas na vizinhança e armazenamos os caminhos percorridos. Por exemplo, considere que o nó 1 tenha o nó 2 e 3 como vizinhos. Assim, estimulamos voltas de 1 para 2 e de 1 para 3. Desta maneira, os ‘textos’, isto é, o conjunto de palavras, de entrada para o Skip-gram seriam do tipo ‘1 2 1 3’, que são justamente as voltas realizadas no grafo. Fazendo esse processo para todos os nós e utilizando a adaptação do

[6]https://bit.ly/2GqZmGT

Skip-gram model se consegue representar cada nó como um vetor de números.

4.1.2 Formalização

Neste momento é apresentada a formalização do *Skip-gram model*, adaptado para o contexto do trabalho. Ou seja, a ideia é apresentar esta formalização em função dos nós de um grafo, em vez de palavras.

Seja o grafo $G = (V,\ E)$ a representação das comunidades estudadas. Em síntese, trata-se de um grafo direcionado onde V é o conjunto de nós e E o de arestas. Considere a função $f : V \rightarrow \mathbb{R}^d$ como a função que faz o mapeamento de cada nó para um vetor de números reais. Nesta função, a variável d representa o número de dimensões que representará cada nó. Por exemplo, se $d = 5$, isto significa que os nós serão representados por 5 números reais. Dando seguimento, a função f é uma matriz de tamanho $|V|$ x d. Esta matriz é parecida com aquela apresentada na Seção 4.1.1, onde o número de linhas corresponde ao número de palavras e o número de colunas ao número de componentes da camada intermediária. Porém, neste caso, o número de linhas corresponde à quantidade de nós e as colunas às dimensões definidas pela variável d.

Para cada nó $u \in V$, define-se $Ns(u) \subset V$ como a vizinhança do nó u, obtida por algum algoritmo que estimule voltas em grafo, partindo de u. Assim, se busca otimizar a função objetivo descrita na fórmula 1, que maximize a probabilidade de ser observada $Ns(u)$ para o nó u, isto é, a vizinhaça de u, condicionada pela sua representação dada pela função f. De forma coloquial, em analogia com a Seção 4.1.1, o que se deseja é encontrar uma representação para o nó u (vetor de números com d posições) que maximize a probabilidade dele ser encontrado perto de seus vizinhos (similar ao exemplo anterior com as palavras 'Obama' e 'presidente').

$$(1) \quad \max_{f} = \sum_{u \in V} \log Pr(Ns(u)|f(u))$$

Detalhando, primeiramente, para tornar o problema tratável, assume-se que a probabilidade de se observar um vizinho de u é independente da observação de qualquer vizinho de u. Ou seja, a existência de um vizinho não influencia na existência de outro vizinho sendo, portanto, considerados eventos condicionalmente independentes. Assim, parte da fórmula 1 pode ser reescrita conforme a fórmula 2.

$$(2) \quad Pr(Ns(u)|f(u)) = \prod_{n_i \in Ns(u)} Pr(n_i|f(u))$$

Continuando, parte da fórmula 2 ($Pr(n_i|f(u))$) define a probabilidade de um determinado nó n_i ocorrer (ser vizinho), dada a representação de outro nó em função de $f(u)$. Em referência à Seção 4.1.1, parte desta fórmula equivale a função softmax e, baseado no *Skip-gram model*, tal parte pode ser reescrita conforme a fórmula 3.

$$(3) \quad Pr(n_i|f(u)) = \frac{\exp(f(n_i).f(u))}{\sum_{v \in V} \exp(f(v).f(u))}$$

Em resumo, busca-se uma nova representação dos nós, através de um vetor de números, que capte a relação de cada um com seus vizinhos. Ademais, para encontrar a função f aproximada, é comum utilizar algoritmos como o *stochastic gradient* [61, 101].

4.1.3 Simple Walk

Até então, foi apresentada a parte mais conceitual que envolve o processo de *feature learning*. Neste instante, a primeira proposta de *feature learning* deste trabalho é descrita. Como visto na Seção 4.1.2, o processo de *feature learning* requer que exemplos de voltas no grafo sejam gerados para alimentar tal processo. No caso de *feature learning* voltado para o processamento de linguagem natural, isto é, textos, estes exemplos são fragmentos do textos, contendo palavras adjacentes. No contexto da representação em grafos, há a necessidade de algoritmos que gerem tais exemplos. Tendo em vista este cenário, há diversas formas para se caminhar em um grafo e construir estes exemplos.

O algoritmo 1, *Simple Walk*, mostra uma destas possibilidades. Nele, voltas simples são estimuladas em um grafo, de forma a originar tais exemplos. Estas voltas simples se tratam de visitas aos nós mais próximos, que são vizinhos diretos de outro nó. Ou seja, o algoritmo 1 pode ser visto como uma busca em largura que vai até o primeiro nível de profundidade. Explicando melhor, este algoritmo recebe um grafo e um nó de origem como parâmetros. Assim, a partir desta origem, visita-se seus vizinhos diretos (ou seja, não considera os vizinhos dos vizinhos) e armazena as voltas percorridas, isto é, os nós visitados, em uma lista.

O fato é que o algoritmo 1 é somente parte do processo de *feature learning*. Conforme observado na Seção 4.1.2, este processo se trata de um problema de otimização. O algoritmo 2 mostra como ocorrem todas as etapas do processo de *feature learning*. Este algoritmo recebe como parâmetro o grafo, bem como, o número de dimensões que deverão representar cada nó. Assim, para cada nó do grafo, invoca-se o algoritmo *Simple Walk*, visando gerar exemplos de vizinhanças para todos eles. Uma vez gerado isto, chama-se o algoritmo

Algorithm 1 Simple Walk

```
1: function simpleWalk(Graph g, Node startNode)
2:     List walk ← [ ]
3:     List neighbors ← g.neighbors(startNode)
4:     int i ← 0
5:     while i < length(neighbors) do
6:         Node neighbor ← neighbors[i]
7:         int startNodeId ← startNode.id
8:         int neighborId ← neighbor.id
9:         append startNodeId to walk
10:        append neighborId to walk
11:        i ← i + 1
12:    return walk
```

Algorithm 2 Feature Learning + Simple Walk

```
1: function featureLearning(Graph g, int dimensions)
2:     List walkList ← [ ]
3:     List nodes ← g.nodes()
4:     int i ← 0
5:     while i < length(nodes) do
6:         Node node ← nodes[i]
7:         List walk ← simpleWalk(g, node)
8:         append walk to walkList
9:         i ← i + 1
10:    f ← StochasticGradientDescent(walkList, dimensions)
11:    return f
```

Stochastic Gradient Descent[7], que é o responsável por resolver o problema de otimização enunciado na Seção 4.1.2. Por fim, o retorno do algoritmo 2, é a função que transforma um nó em um vetor de números.

4.1.4 Go Ahead When Necessary

A segunda proposta de *feature learning* deste trabalho consiste em uma extensão do método mostrado na Seção 4.1.3. Esta extensão é apresentada nos algoritmos 3 e 4.

Para estimular voltas no grafo e gerar os exemplos, neste caso, foi elaborado o algoritmo *Go Ahead When Necessary*, que objetiva realizar voltas mais profundas no grafo em alguns cenários. Detalhando, como parâmetros este algoritmo recebe o grafo, um nó de origem e um terceiro argumento que representa o nível máximo de profundidade para se percorrer (maxLevel), a partir da origem. Além disso, este funciona da mesma forma que o *Simple Walk* (ver algoritmo 4.1.3), caso o grau do nó de origem seja maior ou igual a média de grau de todos os nós do grafo. Por outro lado, caso o grau do nó de origem seja menor que a média, uma busca em largura é realizada até o nível máximo de profundidade definido. Esta busca em largura mais profunda, por sua vez, irá gerar mais exemplos de voltas que contenham os nós com poucos vizinhos diretos.

Como explicado na Seção 4.1.3, o algoritmo que estimula voltas e gera exemplos é somente parte do processo. Desta forma, foi elaborado o algoritmo 4 que invoca o algoritmo 3 para gerar exemplos para todos os nós e, depois disto, resolver o problema de otimização da Seção 4.1.2, através do *Stochastic Gradient Descent*.

[7]O nome semanticamente correto seria gradient ascent por se tratar de um problema de maximização. Entretanto, o nome do algoritmo é gradient descent, bastando uma configuração para tratar problemas de maximização em vez de minimização.

Algorithm 3 Go Ahead When Necessary

```
1: function GOAHEAD(Graph g, Node startNode, int maxLevel)
2:     List walk ← [ ]
3:     if startNode.degree() < g.avgDegree() then
4:         walk ← BreadthFirstSearch(g, startNode, maxLevel)
5:     else
6:         walk ← simpleWalk(g, startNode)
7:     return walk
```

Algorithm 4 Feature Learning + Go Ahead When Necessary

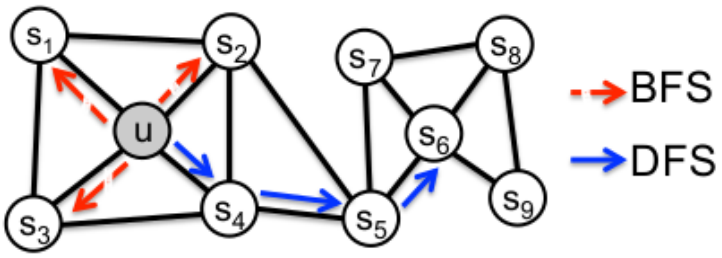
```
1: function FEATURELEARNING(Graph g, int dimensions, int maxLevel)
2:     List walkList ← [ ]
3:     List nodes ← g.nodes()
4:     int i ← 0
5:     while i < length(nodes) do
6:         Node node ← node[i]
7:         List walk ← goAhead(g, node, maxLevel)
8:         append walk to walkList
9:         i ← i + 1
10:    f ← StochasticGradientDescent(walkList, dimensions)
11:    return f
```

4.1.5 Outras Abordagens

Existem na literatura outras abordagens de *feature learning* para nós de grafos. Estas abordagens, assim como as apresentadas neste Capítulo, buscam resolver o problema de otimização mostrado na Seção 4.1.2. Em geral, o que muda, é a maneira com que se estimula voltas no grafo para gerar os exemplos.

Uma destas abordagens é a *node2vec* que transforma nós de um grafo em um vetor de números [61]. Em síntese, o *node2vec* permite enviesar as voltas para que estas sejam em largura (BFS - *breadth-first search*) ou em profundidade (DFS - *depth-first search*). Ou seja, dependendo da configuração inicial do *node2vec*, as voltas podem ser mais parecidas com uma busca em largura ou em profundidade, como mostrado na Figura 4.7. Assim, para cada nó visitado, toma-se uma decisão: devo visitar os nós irmãos (aqueles com a mesma origem do nó atual visitado) ou visitar os nós filhos? É importante ressaltar que, apesar de um viés ser estabelecido, nada impede que o oposto a este ocorra. Por exemplo, se decido que as voltas devem parecer mais com uma busca em profundidade, isto significa que há mais chances disto acontecer e não a completa ausência de voltas parecidas com uma busca em largura.

Figura 4.7: Voltas do node2vec



Uma outra abordagem anterior, porém, similar ao *node2vec*, é a DeepWalk [101]. Nesta, ao contrário do *node2vec*, não é possível enviesar as voltas.

Assim, a probabilidade das voltas da busca ser mais parecida com uma busca em largura ou em profundidade é a mesma. O DeepWalk pode ser visto como um caso especial do *node2vec* onde se atribui o mesmo peso, isto é, a mesma probabilidade de se realizar voltas como busca em largura ou em profundidade.

Uma outra abordagem interessante é o LINE [149], que pode ter duas configurações: (i) o *LINE First Order*; (ii) e o *LINE Second Order*. O *LINE First Order* é bem parecido com o *Simple Walk*, proposto na Seção 4.1.3. Assim *LINE First Order* também visita os vizinhos mais próximos (com pronfundidade nível 1), porém, desconsiderando o sentido das arestas. Somente para reforçar, os autores deixam claro em seu artigo que o *LINE First Order* é somente aplicado para grafos não direcionados, conforme o trecho: *'Note that the first-order proximity is only applicable for undirected graphs, not for directed graphs'* [149].

O *LINE Second Order* é similar ao *First Order* considerando, no entanto, a realização de voltas como busca em largura até o segundo nível de profundidade. Ademais, o *LINE Second Order* tem a capacidade de considerar o sentido das arestas, diferentemente do *First Order*.

4.2 Descrição do Estudo

Nesta Seção se inicia a apresentação das etapas dos experimentos relacionados aos métodos de *feature learning* propostos nas Seções 4.1.3 e 4.1.4. Este estudo faz parte do escopo principal deste trabalho, conforme mencionado na Seção 1.4. Este se trata de um estudo explanatório, com objetivo de comprovar ou refutar uma hipótese.

A ideia dos experimentos deste Capítulo é verificar se os métodos propostos, quando aplicados ao problema enunciado na Seção 1.3, isto é, a detecção

dos tipos de usuários, se mostram como boas alternativas. Isto é, primeiro, as características são geradas através das abordagens propostas e, depois disto, estas serão as entradas de modelos de classificação, nos mesmos moldes daqueles da Seção 3.2.6.1. Ou seja, assim como no Capítulo 3, este também objetiva classificar os usuários em ordinários ou de destaque.

Além disso, as abordagens existentes na literatura, apresentadas na Seção 4.1.5, também serão aplicadas ao problema de pesquisa, com a finalidade de compará-las entre si e com as propostas deste trabalho. Por fim, os resultados são discutidos, buscando ressaltar as vantagens e desvantagens de cada método.

Assim, o objetivo deste estudo é **analisar** os métodos de *feature learning* **no contexto** das comunidades online, **com o propósito de** compará-los **em relação** às características automaticamente captadas dos usuários, **sob o ponto de vista** das classificações dos tipos de usuários.

Neste cenário, se busca a comprovação ou refutação das hipóteses, com base nos resultados obtidos por meio dos experimentos.

4.2.1 Hipótese

Para solucionar o problema enunciado na Seção 1.3, bem como testar as propostas de *feature learning*, foi elaborada a seguinte hipótese que o estudo deste Capítulo tentará comprovar. Assim, busca-se verificar se os dois métodos propostos são melhores que outros relacionados, aplicados ao problema endereçado. Desta forma, a hipótese nula H_0 e alternativa H_1 são:

- H_0: Não existem diferenças significativas entre os métodos de *feature learning* propostos em relação aos demais.
- H_1: Há diferenças significativas entre os métodos de *feature learning* propostos em relação aos demais, em favor das abordagens propostas.

Desta forma, o que se deseja nesta pesquisa é refutar a hipótese nula, isto é, dizer que existem diferenças entre as abordagens, e ainda, afirmar que as propostas deste trabalho são melhores que as demais. A expressão 'em relação aos demais' se refere aos métodos discutidos na Seção 4.1.5, isto é, as outras propostas existentes na literatura.

4.2.2 Dados do Experimento

Os dados utilizados no estudo deste Capítulo são os mesmos da Seção 3.1.2. Reiterando, são dados das comunidades Biology Q&A (BQA) e Chemistry Q&A (CQA). No entanto, não se utilizará os dados brutos, isto é, aqueles diretamente extraídos das comunidades. No caso deste estudo, serão utilizados os dados já trabalhos na representação de grafos, conforme descrito na Seção 3.2.3.

4.2.3 Definição dos Grupos

Assim como na Seção 3.1.3, foi definido que os 15% dos usuários com maiores pontuações são os de destaque (top 15). Aqueles não inseridos neste grupo dos usuários top 15, são considerados ordinários.

Entretanto, visando ampliar a visão sobre a eficácia dos métodos de *feature learning*, outro cenário foi avaliado. Assim, neste outro cenário, foram considerados os usuários de destaque como aqueles que estão no grupo com 20% maiores pontuações (top 20). Neste caso, os não inseridos no top 20 são considerados ordinários.

Resumindo, no estudo deste Capítulo observou-se dois cenários: (i) o primeiro considerando como de destaque os usuários top 15; (ii) e o segundo, considerando como de destaque os usuários top 20.

4.2.4 Ameaças à Validade

Neste momento, serão discutidas algumas ameaças a validade deste estudo. Além disso, são apresentadas formas para minimizá-las. Em geral, as ameaças a validade dos experimentos deste Capítulo são parecidas com as discutidas no Capítulo 3, tendo em vista que, o objetivo de ambos é o mesmo (detectar os tipos de usuários), porém, de modos distintos.

4.2.4.1 Validade Interna

Como discutido anteriormente (na Seção 3.1.4.1), a validade interna está relacionada com a capacidade de um estudo que use os mesmos dados replique os resultados. Para mitigar esta ameaça, os experimentos estão disponíveis no GitHub[8]. Além disso, os dados utilizados estão disponíveis na Web, como relatado na Seção 1.5.

4.2.4.2 Validade Externa

As ameaças a validade externa são as mesmas daquelas discutidas na Seção 3.1.4.2. Em síntese, giram em torno da possibilidade das pontuações oriundas das comunidades não representar alguém de destaque, bem como, a capacidade dos resultados dos experimentos não se reproduzirem com outros dados. Na Seção 3.1.4.2 são comentadas formas para minimizar tais ameaças sendo válidas também para este estudo.

4.2.4.3 Validade de Construção

As ameaças de construção, bem como, as maneiras de minimizá-las já foram comentadas na Seção 3.1.4.3. Reiterando, estas ameaças dizem respeito às definições dos grupos, como apresentado na Seção 4.2.3, que são passíveis de questionamentos. Na Seção 3.1.4.3, foi argumentado que tais definições não são arbitrárias, pois tiveram como base trabalhos anteriores.

[8] `https://bit.ly/2HAt7HN`

Desta forma, o mesmo argumento se aplica nos experimentos deste Capítulo.

4.2.4.4 Validade de Conclusão

Como também comentado na Seção 3.1.4.4, as ameaças de validade de conclusão têm relação com as variáveis utilizadas e sua conexão com os resultados obtidos. No Capítulo 3, foi argumentada a necessidade de se considerar mais de uma perspectiva, possibilitando ter uma visão mais completa sobre os comportamentos dos usuários, de forma a chegar a uma conclusão mais precisa.

No caso do estudo deste Capítulo não se lida com as perspectivas, mas com métodos de *feature learning* para 'caracterizar os usuários'. Assim, optou-se por considerar outros métodos de *feature learning* relacionados, aplicados ao mesmo problema. Assim, espera-se que métodos de *feature learning* parecidos apresentem resultados similares, de maneira a permitirem uma conclusão adequada sobre o problema tratado.

4.2.5 Mecanismos de Análises

Como amplamente discutido, o trabalho objetiva propor duas abordagens de *feature learning*, para endereçar o problema de detecção de tipos de usuários. Nesta linha, estas duas propostas serão comparadas com outras correlatas, aplicadas ao mesmo problema. Assim, serão estabelecidas hipóteses, visando entender se existem diferenças significativas entre as abordagens. Conforme na Seção 4.2.1, a hipótese que se deseja confirmar neste trabalho é que as abordagens propostas superam as demais relacionadas, aplicadas à ao problema específico de detecção de tipos de usuários.

Em síntese, os métodos de *feature learning* serão executados para gerar as características dos usuários. Depois, tais características serão submetidas a um classificador de *Machine Learning* que, além das classificações em si,

retornará a métrica *Area Under ROC curve* (*receiver operating characteristic curve*), usualmente referida como AUC [106]. A partir desta métrica, é possível comparar qual método de *feature learning* gerou as melhores características, de forma a permitir uma melhor classificação, isto é, a detecção do tipo de usuário.

Assim, a métrica AUC será usada para comprovar ou refutar as hipóteses (no caso, a nula e a alternativa), sendo estas submetidas a testes de inferência estatística. O teste de inferência estatística utilizado foi o teste de Welch, adequando para distribuições que apresentam sinais de normalidade (gaussianas). Sua escolha foi devido a aplicação do teste de Kolmogorov-Smirnov sobre as métricas AUC obtidas, que confirmou que tais distribuições apresentam fortes traços de gaussianas. Além disso, para obter o tamanho da diferença, caso o teste de Welch verifique isto, a medida de tamanho de efeito utilizada foi a Cohen's d. Esses métodos estatísticos são ideais quando deseja-se entender diferenças entre distribuições de dados, bem como, comprovar hipóteses, além de serem amplamente utilizados e aceitos na comunidade científica [9, 123, 39].

4.3 Execução do Estudo

Neste momento, inicia-se a execução dos estudo experimental. Deseja-se comparar quais métodos de *feature learning*, bem como suas variantes, são capazes de gerar as características que, quando submetidas a um modelo de classificação, têm maior índice de acertos.

Os métodos de *feature learning* analisados neste estudo foram:

- *Simple Walk*;
- *Go Ahead When Necessary*;

- *node2vec*, com o viés para busca em largura (*node2vec BFS-like*);
- *node2vec*, com o viés para busca em profundidade (*node2vec DFS-like*);
- *Deep Walk*;
- *LINE First Order*;
- *LINE Second Order*.

Para todos os métodos de *feature learning* considerados, foram executados de forma a gerar os exemplos, isto é, as características de cada nó. Cada um foi executado seis vezes. Cada execução gerou representações dos nós com diferentes dimensões (quantidade de números que representam cada nó). Detalhando, na primeira execução de cada método, cada nó foi representado por um vetor com cinco números reais (ou seja, a dimensão da representação é igual a cinco). Da mesma maneira, as demais execuções geraram represenções com 10, 15, 20, 25 e 30 dimensões para cada nó.

Para o método *Go Ahead When Necessary*, consideramos o parâmetro *maxLevel* igual a 8, tendo em vista que o diâmetro do grafo da comunidade BQA é 9 e da CQA é 10. Para clarificar, diâmetro é a maior distância entre qualquer par de nós. Assim, se optou pelo tamanho de profundidade máximo das voltas nos grafos por um valor próximo ao diâmetro das comunidades, visando possibilitar que nós poucos conectados sejam considerados sob a perspectiva de uma vizinhança mais ampla.

Depois, para cada conjunto de características geradas, estas foram submetidas ao classificador *Stochastic Gradient Boosting*, explicado na Seção 3.2.6. Similarmente, foram selecionados 60% dos dados disponíveis para ser parte do conjunto de treinamento e o restante o conjunto de teste. Também foi aplicada a validação cruzada *k-fold* (k = 5) objetivando evitar *over-fitting*.

Além disso, como comentado na Seção 4.2.3, há dois cenários possíveis para as classificações. O primeiro considera que os usuários de destaque são aqueles do top 15 e o segundo os que fazem parte do top 20. Assim, cada comunidade, método de *feature learning*, dimensão de representação e cenário corresponde a tentativas de classificação, sendo a qualidade das classificações medidas em termos da AUC. As Tabelas 4.1, 4.2, 4.3 e 4.4 mostram através da métrica AUC, o quão bem-sucedida foi cada classificação. Ademais, tais Tabelas mostram em qual comunidade a classificação ocorreu, a dimensão da representação usada como entrada ao classificador e, em seus respectivos títulos, o cenário utilizado.

4.3.1 Comparando as Classificações

Uma vez realizadas as classificações, obtém-se os indicadores AUC que permitem comparar a qualidade dos classificadores. Assim, para cada cenário, comunidade e método de *feature learning* usado, comparou-se os indicadores AUC.

Para a comparação dos indicadores AUC, como já comentado, foi utilizado o teste de Welch, que é um teste de inferência estatística usado para distribuições gaussianas ou bem parecidas com elas. Como relatado na Seção 4.2.5, foi verificada se as distribuições são ou não gaussianas através do método de Kolmogorov-Smirnov. Este confirmou que as distribuições dos indicadores AUC possuem fortes traços de normalidade, isto é, são muito provavelmente gaussianas.

As Figuras 4.8 e 4.9 sumarizam os resultados dos experimentos. Assim, é possível verificar a superioridade das abordagens propostas em relação as outras. Em especial, a abordagem *Simple Walk* foi a que obteve melhores resultados, isto é, as características geradas por este método possibilitaram melhores classificações nos termos da AUC. O outro método proposto, *Go*

Ahead When Necessary, também obteve resultados melhores que as demais abordagens, com exceção da *Simple Walk*. Um único cenário que não se pode afirmar qual método (*Simple Walk* vs. *Go Ahead When Necessary*) se saiu melhor foi a classificação dos top 20 na comunidade CQA.

A Tabela 4.5, mostra as comparações de cada método. Sua primeira coluna descreve o cenário de classificação proposto e a segunda a comunidade onde o teste foi realizado. A terceira e a quarta coluna, mostra as distribuições AUC obtidas pelos classificadores, que tiveram como entradas as características geradas por algum dos métodos propostos. Lembrando que, para cada método de *feature learning*, foram geradas 6 métricas AUC, ou seja, uma para cada dimensão. Estas 6 métricas compõem a distribuição AUC de cada método, aplicado a um cenário e comunidade. Continuando, a quinta coluna mostra o p-value (p) correspondente à comparação entre distribuições AUC. Caso p>0,05, a interpretação da diferença foi considerada inválida, isto é, sem possibilidade de conclusão. Nos demais casos (aqueles com p<0,05), a diferença foi a favor de alguma distribuição. Em outras palavras, quando se diz que a interpretação da diferença foi a favor de determinada distribuição associada a um método, significa que este obteve melhor resultado (maiores AUC). Voltando à Tabela 4.5, a sexta coluna mostra a interpretação da diferença entre as distribuições e, por fim, a sétima se a comparação da linha é válida ou não.

Figura 4.8: AUC - Feature Learning Top 15

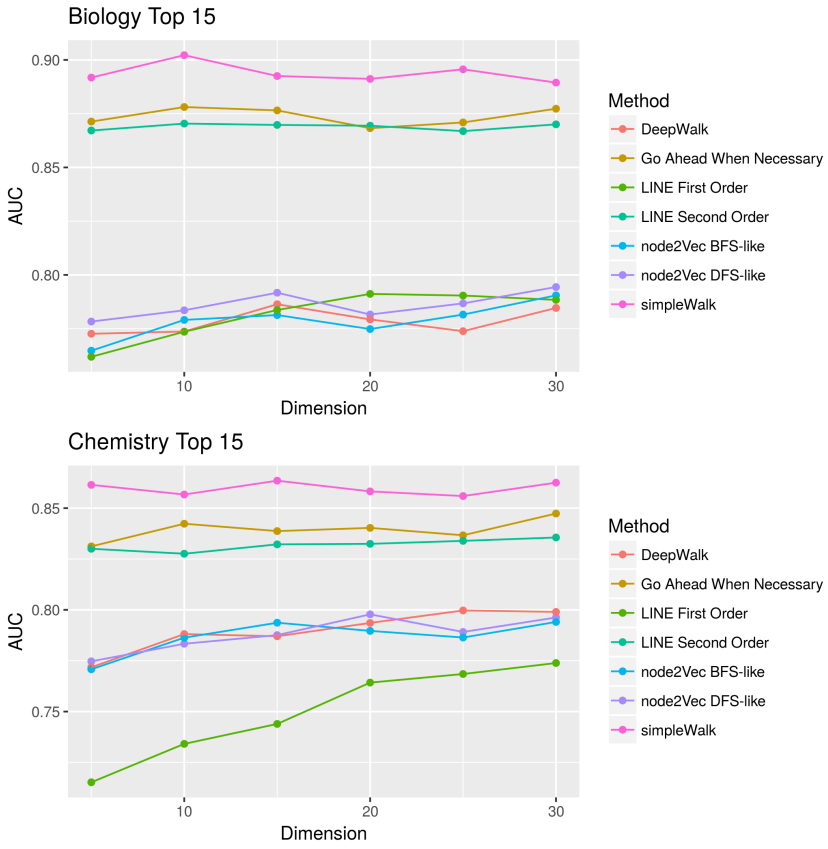


Figura 4.9: AUC - Feature Learning Top 20

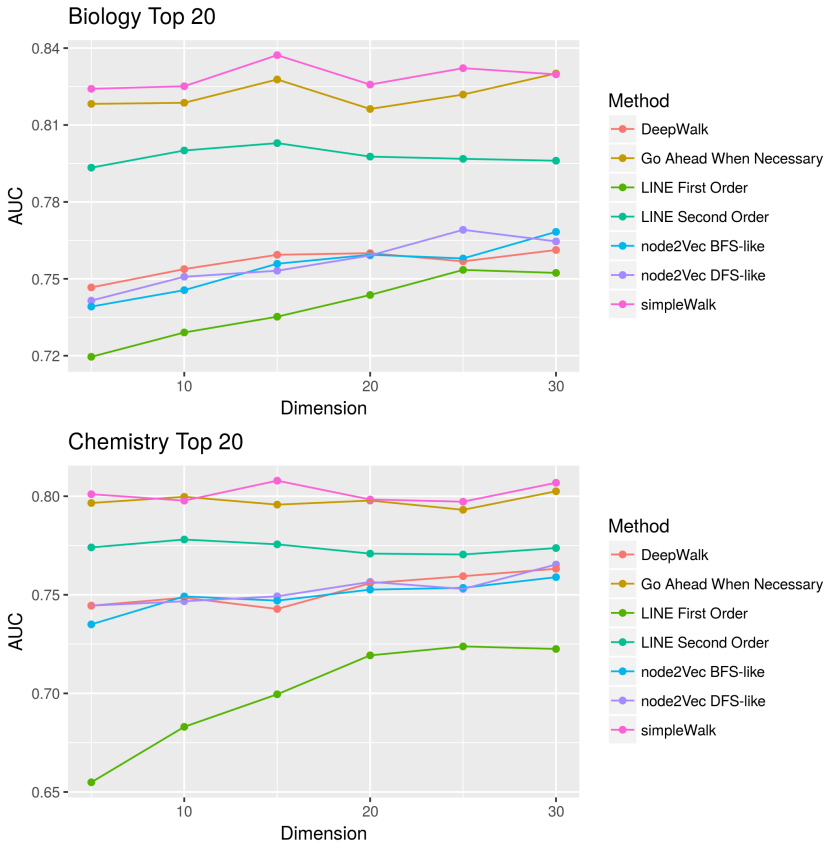


Tabela 4.1: BQA - Classificação top 15

Comunidade	Método	Dimensão	AUC
BQA	Go Ahead When Necessary	5	0,87
BQA	Simple Walk	5	0,89
BQA	node2Vec DFS-like	5	0,78
BQA	Deep Walk	5	0,77
BQA	node2Vec BFS-like	5	0,76
BQA	LINE First Order	5	0,76
BQA	LINE Second Order	5	0,87
BQA	Go Ahead When Necessary	10	0,88
BQA	Simple Walk	10	0,90
BQA	node2Vec DFS-like	10	0,78
BQA	Deep Walk	10	0,77
BQA	node2Vec BFS-like	10	0,78
BQA	LINE First Order	10	0,77
BQA	LINE Second Order	10	0,87
BQA	Go Ahead When Necessary	15	0,88
BQA	Simple Walk	15	0,89
BQA	node2Vec DFS-like	15	0,79
BQA	Deep Walk	15	0,79
BQA	node2Vec BFS-like	15	0,78
BQA	LINE First Order	15	0,78
BQA	LINE Second Order	15	0,87
BQA	Go Ahead When Necessary	20	0,87
BQA	Simple Walk	20	0,89
BQA	node2Vec DFS-like	20	0,78
BQA	Deep Walk	20	0,78
BQA	node2Vec BFS-like	20	0,77
BQA	LINE First Order	20	0,79
BQA	LINE Second Order	20	0,87
BQA	Go Ahead When Necessary	25	0,87
BQA	Simple Walk	25	0,90
BQA	node2Vec DFS-like	25	0,79
BQA	Deep Walk	25	0,77
BQA	node2Vec BFS-like	25	0,78
BQA	LINE First Order	25	0,79
BQA	LINE Second Order	25	0,87
BQA	Go Ahead When Necessary	30	0,88
BQA	Simple Walk	30	0,89
BQA	node2Vec DFS-like	30	0,79
BQA	Deep Walk	30	0,78
BQA	node2Vec BFS-like	30	0,79
BQA	LINE First Order	30	0,79
BQA	LINE Second Order	30	0,87

Tabela 4.2: CQA - Classificação top 15

Comunidade	Método	Dimensão	AUC
CQA	Go Ahead When Necessary	5	0,83
CQA	Simple Walk	5	0,86
CQA	node2Vec DFS-like	5	0,77
CQA	Deep Walk	5	0,77
CQA	node2Vec BFS-like	5	0,77
CQA	LINE First Order	5	0,72
CQA	LINE Second Order	5	0,83
CQA	Go Ahead When Necessary	10	0,84
CQA	Simple Walk	10	0,86
CQA	node2Vec DFS-like	10	0,78
CQA	Deep Walk	10	0,79
CQA	node2Vec BFS-like	10	0,79
CQA	LINE First Order	10	0,73
CQA	LINE Second Order	10	0,83
CQA	Go Ahead When Necessary	15	0,84
CQA	Simple Walk	15	0,86
CQA	node2Vec DFS-like	15	0,79
CQA	Deep Walk	15	0,79
CQA	node2Vec BFS-like	15	0,79
CQA	LINE First Order	15	0,74
CQA	LINE Second Order	15	0,83
CQA	Go Ahead When Necessary	20	0,84
CQA	Simple Walk	20	0,86
CQA	node2Vec DFS-like	20	0,80
CQA	Deep Walk	20	0,79
CQA	node2Vec BFS-like	20	0,79
CQA	LINE First Order	20	0,76
CQA	LINE Second Order	20	0,83
CQA	Go Ahead When Necessary	25	0,84
CQA	Simple Walk	25	0,86
CQA	node2Vec DFS-like	25	0,79
CQA	Deep Walk	25	0,80
CQA	node2Vec BFS-like	25	0,79
CQA	LINE First Order	25	0,77
CQA	LINE Second Order	25	0,83
CQA	Go Ahead When Necessary	30	0,85
CQA	Simple Walk	30	0,86
CQA	node2Vec DFS-like	30	0,80
CQA	Deep Walk	30	0,80
CQA	node2Vec BFS-like	30	0,79
CQA	LINE First Order	30	0,77
CQA	LINE Second Order	30	0,84

Tabela 4.3: BQA - Classificação top 20

Comunidade	Método	Dimensão	AUC
BQA	Go Ahead When Necessary	5	0,82
BQA	Simple Walk	5	0,82
BQA	node2Vec DFS-like	5	0,74
BQA	Deep Walk	5	0,75
BQA	node2Vec BFS-like	5	0,74
BQA	LINE First Order	5	0,72
BQA	LINE Second Order	5	0,79
BQA	Go Ahead When Necessary	10	0,82
BQA	Simple Walk	10	0,83
BQA	node2Vec DFS-like	10	0,75
BQA	Deep Walk	10	0,75
BQA	node2Vec BFS-like	10	0,75
BQA	LINE First Order	10	0,73
BQA	LINE Second Order	10	0,80
BQA	Go Ahead When Necessary	15	0,83
BQA	Simple Walk	15	0,84
BQA	node2Vec DFS-like	15	0,75
BQA	Deep Walk	15	0,76
BQA	node2Vec BFS-like	15	0,76
BQA	LINE First Order	15	0,74
BQA	LINE Second Order	15	0,80
BQA	Go Ahead When Necessary	20	0,82
BQA	Simple Walk	20	0,83
BQA	node2Vec DFS-like	20	0,76
BQA	Deep Walk	20	0,76
BQA	node2Vec BFS-like	20	0,76
BQA	LINE First Order	20	0,74
BQA	LINE Second Order	20	0,80
BQA	Go Ahead When Necessary	25	0,82
BQA	Simple Walk	25	0,83
BQA	node2Vec DFS-like	25	0,77
BQA	Deep Walk	25	0,76
BQA	node2Vec BFS-like	25	0,76
BQA	LINE First Order	25	0,75
BQA	LINE Second Order	25	0,80
BQA	Go Ahead When Necessary	30	0,83
BQA	Simple Walk	30	0,83
BQA	node2Vec DFS-like	30	0,76
BQA	Deep Walk	30	0,76
BQA	node2Vec BFS-like	30	0,77
BQA	LINE First Order	30	0,75
BQA	LINE Second Order	30	0,80

Tabela 4.4: CQA - Classificação top 20

Comunidade	Método	Dimensão	AUC
CQA	Go Ahead When Necessary	5	0,80
CQA	Simple Walk	5	0,80
CQA	node2Vec DFS-like	5	0,74
CQA	Deep Walk	5	0,74
CQA	node2Vec BFS-like	5	0,74
CQA	LINE First Order	5	0,65
CQA	LINE Second Order	5	0,77
CQA	Go Ahead When Necessary	10	0,80
CQA	Simple Walk	10	0,80
CQA	node2Vec DFS-like	10	0,75
CQA	Deep Walk	10	0,75
CQA	node2Vec BFS-like	10	0,75
CQA	LINE First Order	10	0,68
CQA	LINE Second Order	10	0,78
CQA	Go Ahead When Necessary	15	0,80
CQA	Simple Walk	15	0,81
CQA	node2Vec DFS-like	15	0,75
CQA	Deep Walk	15	0,74
CQA	node2Vec BFS-like	15	0,75
CQA	LINE First Order	15	0,70
CQA	LINE Second Order	15	0,78
CQA	Go Ahead When Necessary	20	0,80
CQA	Simple Walk	20	0,80
CQA	node2Vec DFS-like	20	0,76
CQA	Deep Walk	20	0,76
CQA	node2Vec BFS-like	20	0,75
CQA	LINE First Order	20	0,72
CQA	LINE Second Order	20	0,77
CQA	Go Ahead When Necessary	25	0,79
CQA	Simple Walk	25	0,80
CQA	node2Vec DFS-like	25	0,75
CQA	Deep Walk	25	0,76
CQA	node2Vec BFS-like	25	0,75
CQA	LINE First Order	25	0,72
CQA	LINE Second Order	25	0,77
CQA	Go Ahead When Necessary	30	0,80
CQA	Simple Walk	30	0,81
CQA	node2Vec DFS-like	30	0,77
CQA	Deep Walk	30	0,76
CQA	node2Vec BFS-like	30	0,76
CQA	LINE First Order	30	0,72
CQA	LINE Second Order	30	0,77

Tabela 4.5: Comparando AUC - Teste de Welch

Cenário	Comunidade	Distribuição AUC (d1)	Distribuição AUC (d2)	p	Interpretação Diferença	Válido
top 15	BQA	Go Ahead When Necessary	simpleWalk	$< 0{,}01$	Em favor de d2	Sim
top 15	BQA	Go Ahead When Necessary	node2Vec DFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	Go Ahead When Necessary	DeepWalk	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	Go Ahead When Necessary	node2Vec BFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	Go Ahead When Necessary	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	Go Ahead When Necessary	LINE Second Order	0,0337	Em favor de d1	Sim
top 15	BQA	simpleWalk	node2Vec DFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	simpleWalk	DeepWalk	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	simpleWalk	node2Vec BFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	simpleWalk	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	simpleWalk	LINE Second Order	$< 0{,}01$	Em favor de d1	Sim
top 15	BQA	node2Vec DFS-like	DeepWalk	0,0538	inconclusivo	Não
top 15	BQA	node2Vec DFS-like	node2Vec BFS-like	0,1190	inconclusivo	Não
top 15	BQA	node2Vec DFS-like	LINE First Order	0,4235	inconclusivo	Não
top 15	BQA	node2Vec DFS-like	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 15	BQA	DeepWalk	node2Vec BFS-like	0,9519	inconclusivo	Não
top 15	BQA	DeepWalk	LINE First Order	0,5752	inconclusivo	Não
top 15	BQA	DeepWalk	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 15	BQA	node2Vec BFS-like	LINE First Order	0,6383	inconclusivo	Não
top 15	BQA	node2Vec BFS-like	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 15	BQA	LINE First Order	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 15	CQA	Go Ahead When Necessary	simpleWalk	$< 0{,}01$	Em favor de d2	Sim
top 15	CQA	Go Ahead When Necessary	node2Vec DFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	Go Ahead When Necessary	DeepWalk	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	Go Ahead When Necessary	node2Vec BFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	Go Ahead When Necessary	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	Go Ahead When Necessary	LINE Second Order	0,0188	Em favor de d1	Sim
top 15	CQA	simpleWalk	node2Vec DFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	simpleWalk	DeepWalk	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	simpleWalk	node2Vec BFS-like	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	simpleWalk	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	simpleWalk	LINE Second Order	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	node2Vec DFS-like	DeepWalk	0,7581	inconclusivo	Não
top 15	CQA	node2Vec DFS-like	node2Vec BFS-like	0,7950	inconclusivo	Não
top 15	CQA	node2Vec DFS-like	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	node2Vec DFS-like	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 15	CQA	DeepWalk	node2Vec BFS-like	0,5901	inconclusivo	Não
top 15	CQA	DeepWalk	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	DeepWalk	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 15	CQA	node2Vec BFS-like	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 15	CQA	node2Vec BFS-like	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 15	CQA	LINE First Order	LINE Second Order	$< 0{,}01$	Em favor de d2	Sim
top 20	BQA	Go Ahead When Necessary	simpleWalk	0,0499	Em favor de d2	Sim
top 20	BQA	Go Ahead When Necessary	node2Vec DFS-like	$< 0{,}01$	Em favor de d1	Sim
top 20	BQA	Go Ahead When Necessary	DeepWalk	$< 0{,}01$	Em favor de d1	Sim
top 20	BQA	Go Ahead When Necessary	node2Vec BFS-like	$< 0{,}01$	Em favor de d1	Sim
top 20	BQA	Go Ahead When Necessary	LINE First Order	$< 0{,}01$	Em favor de d1	Sim
top 20	BQA	Go Ahead When Necessary	LINE Second Order	$< 0{,}01$	Em favor de d1	Sim
top 20	BQA	simpleWalk	node2Vec DFS-like	$< 0{,}01$	Em favor de d1	Sim
top 20	BQA	simpleWalk	DeepWalk	$< 0{,}01$	Em favor de d1	Sim
top 20	BQA	simpleWalk	node2Vec BFS-like	$< 0{,}01$	Em favor de d1	Sim

top 20	BQA	simpleWalk	LINE First Order	< 0,01	Em favor de d1	Sim
top 20	BQA	simpleWalk	LINE Second Order	< 0,01	Em favor de d1	Sim
top 20	BQA	node2Vec DFS-like	DeepWalk	0,9892	inconclusivo	Não
top 20	BQA	node2Vec DFS-like	node2Vec BFS-like	0,7417	inconclusivo	Não
top 20	BQA	node2Vec DFS-like	LINE First Order	0,0300	Em favor de d1	Sim
top 20	BQA	node2Vec DFS-like	LINE Second Order	< 0,01	Em favor de d2	Sim
top 20	BQA	DeepWalk	node2Vec BFS-like	0,6983	inconclusivo	Não
top 20	BQA	DeepWalk	LINE First Order	0,0229	Em favor de d1	Sim
top 20	BQA	DeepWalk	LINE Second Order	< 0,01	Em favor de d2	Sim
top 20	BQA	node2Vec BFS-like	LINE First Order	0,0511	inconclusivo	Não
top 20	BQA	node2Vec BFS-like	LINE Second Order	< 0,01	Em favor de d2	Sim
top 20	BQA	LINE First Order	LINE Second Order	< 0,01	Em favor de d2	Sim
top 20	CQA	Go Ahead When Necessary	simpleWalk	0,1279	inconclusivo	Não
top 20	CQA	Go Ahead When Necessary	node2Vec DFS-like	< 0,01	Em favor de d1	Sim
top 20	CQA	Go Ahead When Necessary	DeepWalk	< 0,01	Em favor de d1	Sim
top 20	CQA	Go Ahead When Necessary	node2Vec BFS-like	< 0,01	Em favor de d1	Sim
top 20	CQA	Go Ahead When Necessary	LINE First Order	< 0,01	Em favor de d1	Sim
top 20	CQA	Go Ahead When Necessary	LINE Second Order	< 0,01	Em favor de d1	Sim
top 20	CQA	simpleWalk	node2Vec DFS-like	< 0,01	Em favor de d1	Sim
top 20	CQA	simpleWalk	DeepWalk	< 0,01	Em favor de d1	Sim
top 20	CQA	simpleWalk	node2Vec BFS-like	< 0,01	Em favor de d1	Sim
top 20	CQA	simpleWalk	LINE First Order	< 0,01	Em favor de d1	Sim
top 20	CQA	simpleWalk	LINE Second Order	< 0,01	Em favor de d1	Sim
top 20	CQA	node2Vec DFS-like	DeepWalk	0,9724	inconclusivo	Não
top 20	CQA	node2Vec DFS-like	node2Vec BFS-like	0,5015	inconclusivo	Não
top 20	CQA	node2Vec DFS-like	LINE First Order	< 0,01	Em favor de d1	Sim
top 20	CQA	node2Vec DFS-like	LINE Second Order	< 0,01	Em favor de d2	Sim
top 20	CQA	DeepWalk	node2Vec BFS-like	0,5412	inconclusivo	Não
top 20	CQA	DeepWalk	LINE First Order	< 0,01	Em favor de d1	Sim
top 20	CQA	DeepWalk	LINE Second Order	< 0,01	Em favor de d2	Sim
top 20	CQA	node2Vec BFS-like	LINE First Order	< 0,01	Em favor de d1	Sim
top 20	CQA	node2Vec BFS-like	LINE Second Order	< 0,01	Em favor de d2	Sim
top 20	CQA	LINE First Order	LINE Second Order	< 0,01	Em favor de d2	Sim

4.4 Discussão

Neste momento, depois do estudo deste Capítulo, é possível discutir os resultados encontrados. Assim, a hipótese enunciada na Seção 4.2.1 é comentada, bem como os resultados que a suportam. Subsequentemente, são discutidas as contribuições, as limitações do estudo em questão, bem como, tentativas sem êxito que fizeram parte do percurso da pesquisa.

4.4.1 Comentando a Hipótese e os Resultados

Através do estudo, verificou-se que a hipótese alternativa H_1 foi comprovada, dentro do contexto e escopo da pesquisa. Conforme descrito na Seção 4.3, há diferenças significativas entre as duas abordagens propostas e os demais métodos da literatura analisados. Ademais, as propostas desta pesquisa se mostraram superiores para o problema enderaçado, como mostrado principalmente pela Tabela 4.5.

No entanto, alguns detalhes merecem maior atenção. O primeiro diz respeito ao número de dimensões geradas pelos métodos de *feature learning* e o indicador da qualidade das classificações (AUC). Em alguns métodos de *feature learning*, o número de dimensões parece exercer menos influência no resultado final da classificação. Por exemplo, o método *Simple Walk*, o *Go Ahead When Necessary* e o *LINE Second Order* se enquadram neste contexto. Analisando as Figuras 4.8 e 4.9, nota-se que a AUC pouco muda com o aumento das dimensões. Em outros contextos como, por exemplo, a classificação da comunidade CQA da Figura 4.9, os métodos *Deep Walk*, *node2vec BFS-like* e *node2vec DFS-like* permitem concluir o mesmo sobre a relação entre dimensão e classificação. Por outro lado, no método *LINE First Order*, o número de dimensões parece exercer forte influência nos resultados dos classificadores. Na maioria dos cenários analisados, as classificações relacionadas ao *LINE First Order* obtiveram melhoras significativas com o aumento do número de dimensões. É importante ressaltar que o *LINE First Order* desconsidera o sentido das arestas sendo esta uma das principais limitações e também diferença deste em relação as demais abordagens. Aparentemente, quando desconsideradas o sentido das arestas, o número de dimensões passa a fazer diferença nas classificações subsequentes.

O segundo ponto de atenção é com relação a coerência dos resultados. No

contexto do problema estudado, talvez o método da literatura mais similar aos propostos seja o *LINE Second Order*, que propõe uma busca em largura até o segundo nível de profundidade e considera o sentido das arestas. Em geral, o *LINE Second Order* pode ser considerado o terceiro melhor método para o problema endereçado. O *node2vec BFS-like* também caminha no grafo como uma busca em largura, porém, até níveis mais profundos. Assim, apesar de similar a ideia, para o problema desta tese, caminhadas profundas no grafo não melhoram os resultados.

4.4.2 Contribuições

Neste momento já é possível elencar algumas contribuições da pesquisa deste Capítulo.

Como principal contribuição da pesquisa apresentada neste Capítulo, cita-se uma nova perspectiva de solução para o problema de detecção de tipos de usuários, conforme amplamente discutido ao longo do trabalho. Através do estudo conduzido, mostrou-se que as abordagens de *feature learning* podem trazer bons resultados, sem a necessidade de se estudar profundamente uma comunidade, como feito no Capítulo 3. Por outro lado, perde-se o entendimento detalhado da comunidade, quando utilizado algum método de *feature learning*. Claramente, quando alguém se depara com este tipo de problema a pergunta que surge é: preciso do entendimento profundo da comunidade para detectar tipos de usuários? Caso a resposta seja não, é factível concluir que as abordagens de *feature learning* podem ser boas alternativas.

A segunda contribuição reside nas duas propostas de *feature learning* em si, bem como, a comparação destas com outras do estado da arte. Para o problema endereçado, as propostas desta tese foram superiores, sendo isto, uma importante contribuição do trabalho.

4.4.3 Limitações

A limitação principal deste trabalho está na estrutura das comunidades analisadas. De acordo com os dados expostos na Seção 3.2.3.4, se percebe que as comunidades estudadas apresentam uma estrutura do tipo 'rico fica mais rico', conforme observado na distribuição de grau do grafo de cada comunidade. Isto significa que, quanto mais interações uma pessoa tem, mais ela tende a ter. Desta forma, as conclusões deste trabalho estão inseridas neste contexto. Possivelmente, grafos com outras topologias, diferentes da 'rico fica mais rico', irão influenciar os resultados. Desta forma, testes futuros neste sentido devem ser considerados. Resumindo, outras comunidades de perguntas e respostas podem ser testadas objetivando confirmar os resultados. Assim, caso os resultados se confirmem, mais credibilidade terão as abordagens propostas.

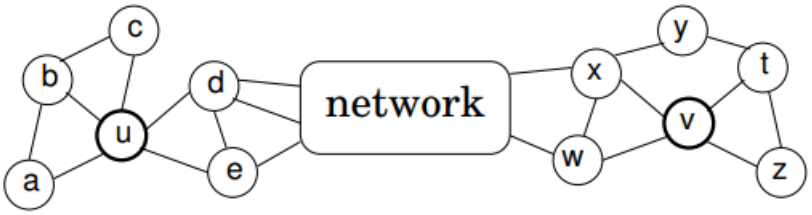
Além disso, os métodos propostos são de caráter geral, isto é, podem ser aplicados em qualquer grafo. No caso deste Capítulo, os métodos foram aplicados a um problema específico, atendendo as restrições do escopo definido. Entretanto, nada impede que estes sejam aplicados em outros grafos que, por sua vez, representam outras relações. Desta maneira, questões desta natureza devem ser apreciadas no futuro.

4.4.4 Outras Tentativas

Embora os resultados apresentados tenham sido animadores, houve tentativas de estudos nesta pesquisa que não obtiveram sucesso. Uma delas tem relação com o conceito de similaridade estrutural.

Como já comentado e feito neste trabalho, as interações entre pessoas podem ser representadas através de grafos. Assim, representando as interações através de um grafo, se percebe que alguns usuários (nós) são estruturalmente

Figura 4.10: u e v com graus e vizinhaça semelhantes, porém, distantes



semelhantes, mesmo estando distantes (Figura 4.10). Assim, na ocasião, foi considerada uma medida denominada distância estrutural, visando saber o quão estruturalmente parecidos são dois pares de nós. Esta distância estrutural pode ser calculada por meio de alguns passos.

Primeiro, são captadas as sequências ordenadas de grau dos nós a uma distância k (*hop-count*) de um nó de origem. Considerando a Figura 4.10 e tendo u como origem, quando k for igual a 0, a sequência de graus é [5], pois, o grau de u é 5. Quando k for igual a 1, a sequência será [2, 2, 3, 3, 4], pois, os vizinhos a e c têm grau 2, o nó b e o nó e têm grau 3 e d tem grau 4 (e todos estes nós estão em uma distância 1 de u, isto é, k = 1). Para k igual a 2, o mesmo procedimento é realizado, porém, considerando os vizinhos a uma distância 2 de u. Em síntese, devemos definir um valor para k e executar esse procedimento para todos os nós do grafo, isto é, todos deverão atuar como origem para se obter as sequências.

Uma vez tendo as sequências de grau de cada nó, se deve ser capaz de dizer se as sequências de dois nós são parecidas. Para isto, pode-se utilizar o algoritmo *Dynamic Time Warping* (DTW) [21] que, em síntese, recebe duas sequências e retornam um valor dizendo o quão similar elas são. Por exemplo, a distância estrutural entre os nós u e v, para k igual a 0, será calculada invocando o algoritmo DTW passando como parâmentro a sequência de u

(sequência igual a [5])[9] e v (sequência igual a [4]), conforme a Fórmula (a). De maneira similar, para calcular a distância estrutural entre u e v para k igual a 1, basta executar o algoritmo DTW tendo como parâmetros as sequências de grau a uma distância 1 e, em seguida, somar com a distância estrutural com k igual 0 destes nós, conforme a Fórmula (b)·

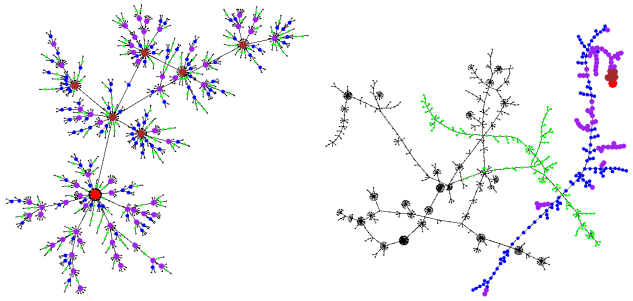
$$(a)\ D_0 = DTW(s_0(u), s_0(v)) \quad (b)\ D_k = DTW(s_k(u), s_k(v)) + D_{k-1}$$
$$s_0(u) = \textit{sequência de grau ordenada de u para k igual a } 0$$
$$s_k(u) = \textit{sequência de grau ordenada de u para qualquer } k$$

Tendo os valores das similaridades estrutural entre todos os pares de nó, utilizamos um algoritmo para gerar uma árvore de custo mínimo (novo grafo). Como entrada para este algoritmo, consideramos que cada par de nó tenha uma conexão com peso igual à similaridade estrutural calculada. Assim, o algoritmo irá gerar uma árvore, cujos nós estruturalmente semelhantes (no grafo original) tendem a ser vizinhos. Na Figura 4.11, é apresentada como seria a árvore gerada a partir do grafo original da comunidade (onde as cores destacam os nós estruturalmente semelhantes). Depois de tudo isto, este grafo (que é uma árvore) seria submetido aos métodos de *feature learning* da mesma maneira conforme explicado na Seção 4.3.

Como se percebe, esta abordagem tem como finalidade modificar o grafo original da comunidade de forma a tentar fazer com que o processo de *feature learning* gere melhores características e, consequentemente, obtenha melhores classificações. Apesar de interessante e tentadora a ideia, os resultados foram ruins para o problema endereçado nesta tese. Ademais, variantes do algoritmo que mede a similaridade entre sequências de grau, isto é, alterna-

[9]Conforme calculado, no exemplo da Figura 4.10.

Figura 4.11: À esquerda o grafo original e à direita a árvore gerada.



tivas ao DTW, foram consideradas. Dentre estas, pode-se citar por exemplo a distância de Levenshtein. No entanto, ainda assim, os resultados não obtiveram melhoras significativas.

Futuramente, se pretende analisar melhor os motivos para tal resultado, bem como explorar a abordagem em outros contextos.

4.5 Comentários Finais

Este Capítulo teve como objetivo apresentar um estudo sobre *feature learning* aplicado ao problema deste trabalho.

Primeiramente, se apresentou o conceito geral de *feature learning* e suas origens. Depois disso, se apresentou as abordagens relacionadas, bem como as propostas de *feature learning* desta tese. Em seguida, o estudo experimental foi descrito, expondo suas principais ameaças e também maneiras para minimizá-las. Por fim, a execução do estudo experimental foi apresentada de forma detalhada, comparando todas as abordagens mostradas e, posteriormente, discutidas.

Em geral, o Capítulo objetivou trazer uma nova visão sobre as soluções existentes para detecção de tipos de usuários. Ou seja, buscou-se desenvolver algo diferente de tudo na literatura de detecção de tipos de usuários. Além disso, os resultados encontrados foram satisfatórios e comparáveis às soluções

tradicionais aplicadas ao problema.

Por fim, mais detalhes técnicos da implementação dos experimentos estão nos Apêndices A e C.

5. Conclusão

Neste Capítulo de conclusão, são resumidas as principais contribuições, limitações e trabalhos futuros. Assim, este tem como finalidade revisitar pontos importantes mencionados ao longo do trabalho, de forma a fechar esta tese.

5.1 Síntese

Este trabalho realizou um estudo detalhado em comunidades de perguntas e respostas, em especial, endereçando um problema debatido na literatura: a detecção de tipos de usuários (de destaque ou ordinário), no contexto das comunidades de perguntas e respostas. No ínicio, foram apresentadas as motivações para a realização desta tese que passam por diversos fatores. Dentre os fatores motivacionais, focou-se no sucesso das comunidades online, sob o ponto de vista da popularidade e seu impacto na vida das pessoas. Além disso, argumentou-se sobre a possibilidade de aprender nestes ambientes, sendo o autor desta pesquisa, como comentado na Seção 1.2.1, uma evidência viva, embora pontual, disto. Entretanto, pesquisas de centros renomados corroboram a referida possibilidade.

Em seguida, introduzimos o problema da pesquisa. O escopo de atuação também foi delineado, mostrando os desdobramentos possíveis do problema.

Assim, se decidiu focar em uma parte específica dele argumentando sobre a dificuldade de se caracterizar e detectar tipos de usuários em ambientes online.

No Capítulo 2, vários trabalhos relacionados foram discutidos, confirmando a relevância do problema abordado. Ademais, conceitos fundamentais para o entendimento da tese também foram esclarecidos tais como o de comunidade online e prestígio social. Comentou-se também as interseções de teorias de aprendizagem com o objeto de estudo deste trabalho. Na sequência, apresentou-se o diferencial da pesquisa em relação os trabalhos relacionados: foi proposto um estudo detalhado dos comportamentos dos usuários sob vários perspectivas e também o uso de *feature learning*.

Nos Capítulos 3 e 4 as análises nas comunidades foram apresentadas. Os Capítulos tiveram o mesmo objetivo, porém, as propostas de soluções foram distintas. No Capítulo 3 apresentamos as análises multi-perspectiva para detectar os tipos de usuários e no Capítulo 4 as propostas de *feature learning*. Em ambos, o estudo foi devidamente descrito, expondo as ameaças envolvidas e formas para contorná-las. Por fim, os resultados foram amplamente discutidos, evidenciado as contribuições encontradas.

5.2 Principais Contribuições

As contribuições desta pesquisa foram discutidas nas Seções 3.3.2 e 4.4.2. Nesta Seção, será reiterado alguns pontos e acrescentados outros.

Em síntese, podemos definir duas categorias de contribuições deste trabalho: (i) de pesquisa; (ii) e a tecnológica. Considerando as contribuições de pesquisa, podemos citar:

1. O estudo empírico dos comportamentos dos usuários, demonstrados

através das análises de diversas perspectivas, dada a limitação do número de perspectivas nos trabalhos anteriores;

2. A proposição de modelos de classificação de tipos de usuários e melhores respostas, através do uso de pespectivas distintas. Tais propostas vão além de caraterísticas de perspectivas comumente usadas em trabalhos anteriores para tais classificações;

3. As duas soluções de *feature learning* propostas e sua comparação com outras abordagens correlatas, demonstrando a superioridade daquelas propostas nesta tese;

4. O conteúdo do conjunto de diversas publicações feitas ao longo do doutorado que deram origem a esta pesquisa [111, 110, 109, 113, 112, 8, 116, 114, 120, 115]. Destas, em especial, cinco têm QUALIS B1 e uma A1, demonstrando que a pesquisa já foi amplamente avaliada por pares de pesquisa científica de veículos de publicação respeitados[1];

5. A disponibilização de todos os artefatos necessários para a reprodução dos experimentos realizados.

Por outro lado, dentre as contribuições tecnológicas, pode-se citar o conjunto de implementações realizadas nas linguagens Java, Python e R. Maiores detalhes sobre estas contribuições estão nos Apêndices.

Dado o exposto, considerando especialmente as publicações, percebe-se que esta pesquisa percorreu um grande caminho de amadurecimento até chegar ao estágio atual. Através destas, *feedbacks* de grande importância foram recebidos, possibilitando a evolução gradual do trabalho.

[1]Os QUALIS citados foram conferidos no período da publicação

5.3 Principais Limitações

Assim como na Seção anterior, as limitações desta pesquisa também foram comentadas nas Seções 3.3.3 e 4.4.3. Neste momento, somente serão recapitulados as principais limitações.

A respeito das limitações das análises do Capítulo 3, estas giram em torno do aprofundamento das análises em cada perspectiva. Como demonstrado, o estudo focou nas diferenças entre os tipos de usuário. Assim, perguntas mais profundas relacionadas, por exemplo, com o grau de generosidade dos usuários podem ser apreciadas. Isto é, qual o interesse real de alguém na comunidade: se promover, ajudar genuinamente ou ambos? Possivelmente, tais questões passam pela área *Computacional Social Science*, que é uma área que busca entender fenômenos sociais através de dados disponíveis na Web.

Sobre os experimentos do Capítulo 4, a principal limitação das propostas de *feature learning* reside na necessidade de avaliar seus respectivos desempenhos em grafos com diferentes distribuições de grau, isto é, distintos dos analisados neste trabalho. Além disso, há também a possibilidade de se aplicar tais métodos em qualquer grafo. Ou seja, possivelmente as abordagens são passíveis de análise em qualquer problema modelado através de grafos. Entretanto, experimentos neste sentido não foram conduzidos, porém, esta pesquisa deixa esta possibilidade de verificação para um trabalho posterior.

5.4 Trabalhos Futuros

Como trabalhos futuros, primeiramente, se deseja atacar as limitações desta pesquisa, conforme relatado. Ademais, o autor desta pesquisa vê grandes possibilidades para a aplicação das análises das perspectivas do Capítulo 3 em ambientes de aprendizagem. Uma área da informática da educação

denominada *Learning Analytics* tem como finalidade, em geral, analisar dados sobre os rastros de aprendizagem de alunos em ambientes online. Desta forma, entende-se que as análises das perspectivas podem ser úteis para ajudar a identificar as competências e dificuldades de alunos ou mesmo predizer uma futura evasão. Há trabalhos que discutem formas para montar o perfil do usuário [79, 17] em uma rede. É possível que as perspectivas relacionadas aos usuários, discutidas nesta tese, enriqueça tais abordagens.

Na Seção 1.3, verifica-se que a recomendação de usuários que dominam determinados tópicos é um problema posterior a detecção dos tipos de usuários. Uma possibilidade futura reside na construção de sistemas de recomendação em comunidades online, que usam as propostas desta tese como base. Assim, uma avaliação qualitativa os métodos propostos torna-se possível.

Sobre as abordagens de *feature learning*, novas formas para caminhar no grafo podem ser consideradas com a finalidade de verificar seu desempenho no problema da tese.

Um outro trabalho futuro pode ser a evolução da ideia da similaridade estrutural, conforme descrito na Seção 4.4.4. Apesar de não ter sido bem-sucedida no contexto estudado, claramente a similaridade estrutural se aproxima da ideia de distância. Em outras palavras, a similaridade estrutural pode ser vista como uma medida de distância entre pessoas. Assim, explorar em outros contextos o significado desta distância, bem como, usá-la para agrupar pessoas pode ser interessante.

Uma questão de pesquisa futura interessante pode ser: como saber se um usuário realmente assimilou a informação de uma discussão? Possivelmente, avaliar a trajetória de um usuário após uma discussão (em que ele participou) seria um bom início para analisar se ele aprendeu determinado conteúdo.

Um atual debate é sobre a efemeridade da informação, conforme discutido

na Seção 2.5.2. Desta forma, investigar por quanto tempo uma postagem de uma comunidade constitui um conhecimento válido é uma possível questão futura de pesquisa.

Bibliografia

[1] Adamic, L. A. and Glance, N. (2005). The political blogosphere and the 2004 us election: divided they blog. In *Proceedings of the 3rd international workshop on Link discovery*, pages 36–43. ACM.

[2] Adamic, L. A., Zhang, J., Bakshy, E., and Ackerman, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pages 665–674. ACM.

[3] Agichtein, E., Liu, Y., and Bian, J. (2009). Modeling information-seeker satisfaction in community question answering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 3(2):10.

[4] Al-Garadi, M. A., Varathan, K. D., Ravana, S. D., Ahmed, E., Mujtaba, G., Khan, M. U. S., and Khan, S. U. (2018). Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Computing Surveys (CSUR)*, 51(1):16.

[5] Alstete, J. W. and Beutell, N. J. (2004). Performance indicators in online distance learning courses: a study of management education. *Quality Assurance in Education*, 12(1):6–14.

[6] Andy, A., Sekine, S., Rwebangira, M., and Dredze, M. (2016). Name variation in community question answering systems. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 51–60.

[7] Anttila, J. (2008). Advanced web 2.0 based interactive technology to support informal learning for enhancing quality of business management. *learning*, 18:5.

[8] Araujo, R. M., Procaci, T. B., Classe, T. M., and Chueri, L. O. V. (2017). Da pesquisa científica à inovação. *Pesquisa e Inovação: visões e interseções. editado por RM Araujo & LOV Chueri. Rio de Janeiro, RJ: Publit.*

[9] Arcuri, A. and Briand, L. (2011). A practical guide for using statistical tests to assess randomized algorithms in software engineering. In *Software Engineering (ICSE), 2011 33rd International Conference on*, pages 1–10. IEEE.

[10] Armstrong, A. C. (2008). The fragility of group flow: The experiences of two small groups in a middle school mathematics classroom. *The Journal of Mathematical Behavior*, 27(2):101–115.

[11] Bailey, M., Cao, R., Kuchler, T., and Stroebel, J. (2016). Social networks and housing markets. Technical report, National Bureau of Economic Research.

[12] Bakhtin, M. (2006). Volochinov. marxismo e filosofia da linguagem. *São Paulo.*

[13] Bakhtin, M. M. *Estética da criação verbal.*

[14] Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: quantifying influence on twitter. In *Proceedings*

of the fourth ACM international conference on Web search and data mining, pages 65–74. ACM.

[15] Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web*, pages 519–528. ACM.

[16] Bakshy, E., Simmons, M., Huffaker, D., Teng, C.-Y., and Adamic, L. (2010). The social dynamics of economic activity in a virtual world. *Ann Arbor*, 1001:48103.

[17] Balog, K., Azzopardi, L., and de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1–19.

[18] Banerjee, A. and Basu, S. (2008). A social query model for decentralized search. In *Proceedings of the 2nd Workshop on Social Network Mining and Analysiss. ACM, New York*, volume 124.

[19] Bauman, Z. (2013). *Liquid modernity*. John Wiley & Sons.

[20] Bengio, Y. and Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of machine learning research*, 5(Sep):1089–1105.

[21] Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA.

[22] Bittencourt, I. I. and Isotani, S. (2018). Evidence-based computers in education: A manifesto. *Brazilian Journal of Computers in Education*, 26(03):108.

[23] Bliuc, A.-M., Faulkner, N., Jakubowicz, A., and McGarty, C. (2018). Online networks of racial hate: A systematic review of 10 years of research on cyber-racism. *Computers in Human Behavior*, 87:75–86.

[24] Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

[25] Borondo, J., Borondo, F., Rodriguez-Sickert, C., and Hidalgo, C. A. (2014). To each according to its degree: The meritocracy and topocracy of embedded markets. *Scientific reports*, 4.

[26] Boroujeni, M. S. and Dillenbourg, P. (2018). Discovery and temporal analysis of latent study patterns in mooc interaction sequences. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, pages 206–215. ACM.

[27] Bosu, A., Corley, C. S., Heaton, D., Chatterji, D., Carver, J. C., and Kraft, N. A. (2013). Building reputation in stackoverflow: an empirical investigation. In *Mining Software Repositories (MSR), 2013 10th IEEE Working Conference on*, pages 89–92. IEEE.

[28] Buckley, P. and Doyle, E. (2017). Individualising gamification: An investigation of the impact of learning styles and personality traits on the efficacy of gamification using a prediction market. *Computers & Education*, 106:43–55.

[29] Burke, M. and Kraut, R. E. (2016). The relationship between facebook use and well-being depends on communication type and tie strength. *Journal of Computer-Mediated Communication*, 21(4):265–281.

[30] Caers, R. and Castelyns, V. (2011). Linkedin and facebook in belgium: The influences and biases of social network sites in recruitment and selection procedures. *Social Science Computer Review*, 29(4):437–448.

[31] Campbell, C. S., Maglio, P. P., Cozzi, A., and Dom, B. (2003). Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531. ACM.

[32] Carvalho, M. J. S. (2013). Proposições e controvérsias no conectivismo. *RIED. Revista Iberoamericana de Educación a Distancia*, 16(2):9–31.

[33] Carvalho, M. J. S., de Nevado, R. A., and de Menezes, C. S. (2005). Arquiteturas pedagógicas para educação à distância: concepções e suporte telemático. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 1, pages 351–360.

[34] Castells, M. (1999). A sociedade em rede, vol. 1. *São Paulo: Paz e Terra*.

[35] Castro, A. and Menezes, C. (2012). Capítulo 9 - aprendizagem colaborativa com suporte computacional. In Pimentel, M. and Fuks, H., editors, *Sistemas Colaborativos*, pages 135 – 153. Elsevier Editora Ltda.

[36] Cha, M., Haddadi, H., Benevenuto, F., Gummadi, P. K., et al. (2010). Measuring user influence in twitter: The million follower fallacy. *Icwsm*, 10(10-17):30.

[37] Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2015). Antisocial behavior in online discussion communities. In *ICWSM*, pages 61–70.

[38] Claes, W., Per, R., Martin, H., Magnus, C., Björn, R., and Wesslén, A. (2000). Experimentation in software engineering: an introduction. *Online Available: http://books. google. com/books*.

[39] Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1):155.

[40] da Costa, A. M. N. and Pimentel, M. (2012). Capítulo 1 - sistemas colaborativos para uma nova sociedade e um novo ser humano. In Pimentel, M. and Fuks, H., editors, *Sistemas Colaborativos*, pages 3 – 15. Elsevier Editora Ltda.

[41] Dom, B., Eiron, I., Cozzi, A., and Zhang, Y. (2003). Graph-based ranking algorithms for e-mail expertise analysis. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 42–48. ACM.

[42] Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

[43] Downes, S. (2012). Connectivism and connective knowledge: essays on meaning and learning networks. *Stephen Downes Web*.

[44] English, R. M. and Duncan-Howell, J. A. (2008). Facebook© goes to college: Using social networking tools to support students undertaking teaching practicum. *Journal of Online Learning and Teaching*, 4(4):596–601.

[45] Facebook (2016). Facebook reports second quarter 2016 results. Available at `https://s21.q4cdn.com/399680738/files/doc_financials/2016/Facebook-Reports-Second-Quarter-2016-Results.pdf`.

[46] Fagundes, L. d. C., Maçada, D., and Sato, L. (1999). Aprendizes do futuro: as inovações começaram. coleção informática para a mudança na educação–ministério da educação. *Brasília: Estação Palavra*.

[47] Fardouly, J., Diedrichs, P. C., Vartanian, L. R., and Halliwell, E. (2015). Social comparisons on social media: The impact of facebook on young women's body image concerns and mood. *Body Image*, 13:38–45.

[48] Fernandes, L. (2011). Redes sociais online e educação: contributo do facebook no contexto das comunidades virtuais de aprendentes. *Universidade Nova de Lisboa, Portugal*.

[49] Ferrara, E., Alipourfard, N., Burghardt, K., Gopal, C., and Lerman, K. (2017). Dynamics of content quality in collaborative knowledge production. *arXiv preprint arXiv:1706.03179*.

[50] Fire, M., Goldschmidt, R., and Elovici, Y. (2014). Online social networks: threats and solutions. *IEEE Communications Surveys & Tutorials*, 16(4):2019–2036.

[51] Freire, P. (1996). Ensinar não é transferir conhecimento. *Pedagogia da autonomia*, 11:52–101.

[52] Freitas, M. T. d. A. (1997). Nos textos de bakhtin e vygotsky: um encontro possível. *BRAIT, B.*

[53] Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

[54] Fritzen, E., Siqueira, S. W., and de Andrade, L. C. (2012). Recuperação contextual de informação na web para apoiar aprendizagem colaborativa

em redes sociais. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 23.

[55] Fu, M., Zhu, M., Su, Y., Zhu, Q., and Li, M. (2016). Modeling temporal behavior to identify potential experts in question answering communities. In *International Conference on Cooperative Design, Visualization and Engineering*, pages 51–58. Springer.

[56] Füller, J., Jawecki, G., and Mühlbacher, H. (2007). Innovation creation by online basketball communities. *Journal of business research*, 60(1):60–71.

[57] Gaggioli, A., Mazzoni, E., Milani, L., and Riva, G. (2015). The creative link: Investigating the relationship between social network indices, creative performance and flow in blended teams. *Computers in Human Behavior*, 42:157–166.

[58] Ghiassian, S. D., Menche, J., Chasman, D. I., Giulianini, F., Wang, R., Ricchiuto, P., Aikawa, M., Iwata, H., Müller, C., Zeller, T., et al. (2016). Endophenotype network models: Common core of complex diseases. *Scientific reports*, 6:27414.

[59] Gomes, A. S., Rolim, A., and Silva, W. (2012). Educar com o redu. *Recife: Redu Educational Technology.*

[60] Grover, A. and Leskovec, J. (2016a). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

[61] Grover, A. and Leskovec, J. (2016b). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864. ACM.

[62] Hamari, J., Koivisto, J., and Sarsa, H. (2014). Does gamification work?– a literature review of empirical studies on gamification. In *2014 47th Hawaii international conference on system sciences (HICSS)*, pages 3025–3034. IEEE.

[63] Hamid, S., Waycott, J., Kurnia, S., and Chang, S. (2015). Understanding students' perceptions of the benefits of online social networking use for teaching and learning. *The Internet and Higher Education*, 26:1–9.

[64] Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Random House.

[65] Herdağdelen, A., Adamic, L., Mason, W., et al. (2016). The social ties of immigrant communities in the united states. In *Proceedings of the 8th ACM Conference on Web Science*, pages 78–84. ACM.

[66] Hertel, G., Geister, S., and Konradt, U. (2005). Managing virtual teams: A review of current empirical research. *Human resource management review*, 15(1):69–95.

[67] Horowitz, D. and Kamvar, S. D. (2010). The anatomy of a large-scale social search engine. In *Proceedings of the 19th international conference on World wide web*, pages 431–440. ACM.

[68] Huberman, B. A., Romero, D. M., and Wu, F. (2008). Social networks that matter: Twitter under the microscope. *Available at SSRN 1313405*.

[69] Jadin, T., Gnambs, T., and Batinic, B. (2013). Personality traits and knowledge sharing in online communities. *Computers in Human Behavior*, 29(1):210–216.

[70] Jara-Figueroa, C., Yu, A. Z., and Hidalgo, C. A. (2015). Estimating technological breaks in the size and composition of human collective memory from biographical data. *arXiv preprint arXiv:1512.05020*.

[71] Katz, E. and Paul, F. (1955). Lazarsfeld. 1955. personal influence: The part played by people in the flow of mass communications. *Glencoe, Illinois: The Free Press. KatzPersonal Influence: The Part Played by People in the Flow of Mass Communication*.

[72] Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature physics*, 6(11):888.

[73] Krulwich, B., Burkey, C., and Consulting, A. (1996). The contactfinder agent: Answering bulletin board questions with referrals. In *AAAI/IAAI, Vol. 1*, pages 10–15.

[74] Kumar, S., Cheng, J., Leskovec, J., and Subrahmanian, V. (2017). An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, pages 857–866. International World Wide Web Conferences Steering Committee.

[75] Levy, P. (2010). *Cibercultura*. Editora 34.

[76] Li, Y.-M., Liao, T.-F., and Lai, C.-Y. (2012). A social recommender mechanism for improving knowledge sharing in online forums. *Information Processing & Management*, 48(5):978–994.

[77] Linked (2016). Linked reports second quarter 2016 results. Available at `https://press.linkedin.com/site-resources/news-releases/2016/linkedin-announces-second-quarter-2016-results`.

[78] Littlepage, G. E. and Mueller, A. L. (1997). Recognition and utilization of expertise in problem-solving groups: Expert characteristics and behavior. *Group Dynamics: Theory, Research, and Practice*, 1(4):324.

[79] Liu, X., Wang, G. A., Johri, A., Zhou, M., and Fan, W. (2014). Harnessing global expertise: A comparative study of expertise profiling methods for online communities. *Information Systems Frontiers*, 16(4):715–727.

[80] Livne, A., Simmons, M. P., Adar, E., and Adamic, L. A. (2011). The party is over here: Structure and content in the 2010 election. *ICWSM*, 11:17–21.

[81] Luckesi, C. C. (2000). O que é mesmo o ato de avaliar a aprendizagem. *Revista Pátio*, 12:6–11.

[82] Luckett, H. P. (1973). What's news: Electronic-mail delivery gets started. *Popular Science*, 202(3):85.

[83] Mason, W. and Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, 109(3):764–769.

[84] Mathieu, M., Couprie, C., and LeCun, Y. (2015). Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*.

[85] McCormick, C. (2016). Word2vec tutorial-the skip-gram model.

[86] Menezes, C. S., de Nevado, R. A., de Castro Jr, A. N., and Santos, L. N. (2008). Morfeu–multi-organizadorflexível de espaços virtuais para apoiar a inovação pedagógica em ead. In *Anais do Simpósio Brasileiro de Informática na Educação*, volume 1, pages 451–460.

[87] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[88] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

[89] Morris, M. R., Teevan, J., and Panovich, K. (2010a). A comparison of information seeking using search engines and social networks. *ICWSM*, 10:23–26.

[90] Morris, M. R., Teevan, J., and Panovich, K. (2010b). 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*, pages 1739–1748. ACM.

[91] Mui, Y. Q. and Whoriskey, P. (2010). Facebook passes google as most popular site on the internet, two measures show. *The Washington Post*.

[92] Mukherjee, A., Venkataraman, V., Liu, B., and Glance, N. S. (2013). What yelp fake review filter might be doing? In *ICWSM*.

[93] Mukherjee, S., Lamba, H., and Weikum, G. (2015). Experience-aware item recommendation in evolving review communities. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 925–930. IEEE.

[94] Nicolaci-da Costa, A. M. (2002). Revoluções tecnológicas e transformações subjetivas. *Psicologia: teoria e pesquisa*, 18(2):193–202.

[95] Odiete, O., Jain, T., Adaji, I., Vassileva, J., and Deters, R. (2017). Recommending programming languages by identifying skill gaps using analysis of experts. a study of stack overflow. In *Adjunct Publication of the 25th Conference on User Modeling, Adaptation and Personalization*, pages 159–164. ACM.

[96] Ottoni, R., Cunha, E., Magno, G., Bernadina, P., Meira Jr, W., and Almeida, V. (2018). Analyzing right-wing youtube channels: Hate, violence and discrimination. *arXiv preprint arXiv:1804.04096*.

[97] Pal, A., Chang, S., and Konstan, J. A. (2012). Evolution of experts in question answering communities. In *ICWSM*.

[98] Pal, A., Herdagdelen, A., Chatterji, S., Taank, S., and Chakrabarti, D. (2016). Discovery of topical authorities in instagram. In *Proceedings of the 25th International Conference on World Wide Web*, pages 1203–1213. International World Wide Web Conferences Steering Committee.

[99] Palmer, S., Holt, D., and Bray, S. (2008). Does the discussion help? the impact of a formally assessed online discussion on final student results. *British Journal of Educational Technology*, 39(5):847–858.

[100] Paul, S. A., Hong, L., and Chi, E. H. (2011). Is twitter a good place for asking questions? a characterization study. In *ICWSM*.

[101] Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

[102] Perrin, A. (2015). Social media usage: 2005-2015. *Pew Research Center. October 2015.* Avaliable at `http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/`.

[103] Pimentel, M. (2006). Comunicatec: Tecnologias de comunicação para educação e colaboração. *Anais do III Simpósio Brasileiro de Sistemas de Informação. Curitiba, PR.*

[104] Posnett, D., Warburg, E., Devanbu, P., and Filkov, V. (2012). Mining stack exchange: Expertise is evident from initial contributions. In *Social Informatics (SocialInformatics), 2012 International Conference on*, pages 199–204. IEEE.

[105] Potthast, M., Stein, B., and Gerling, R. (2008). Automatic vandalism detection in wikipedia. In *European Conference on Information Retrieval*, pages 663–668. Springer.

[106] Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

[107] Prates, J. C., Fritzen, E., Siqueira, S. W., Braz, M. H. L., and De Andrade, L. C. (2013). Contextual web searches in facebook using learning materials and discussion messages. *Computers in Human Behavior*, 29(2):386–394.

[108] Preece, J. (2000). Online communities: designing usability, supporting sociability. *Industrial Management & Data Systems*, 100(9):459–460.

[109] Procaci, T., Nunes, B. P., Júnior, F. P., and Siqueira, S. (2015a). Análise empírica de comunidades online baseadas em enriquecimento semântico para encontrar usuários especialistas. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 4, page 558.

[110] Procaci, T., Siqueira, S., and Andrade, L. (2015b). Identificando colegas para ajudar em minhas dúvidas: Um estudo empírico em comunidades de perguntas e respostas. In *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, volume 4, page 65.

[111] Procaci, T., Siqueira, S., Júnior, F. P., and Nunes, B. P. (2015c). Estudo exploratório das produçoes e colaboraçoes entre pesquisadores em informática na educaçao: uma análise de publicaçoes do simpósio brasileiro de informática na educaçao de 2001 a 2013. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 26, page 1323.

[112] Procaci, T. B., Araujo, R., Siqueira, S. W. M., and Nunes, B. P. (2016a). Prospecção tecnológica: Levantamento de patentes, atuação da academia e potenciais inovações em ambientes de aprendizagem no brasil de 2000 a 2015. *iSys: Revista Brasileira de Sistemas de Informação,*, 9:69–88.

[113] Procaci, T. B., Nunes, B. P., Nurmikko-Fuller, T., and Siqueira, S. W. (2016b). Finding topical experts in question & answer communities. In *Advanced Learning Technologies (ICALT), 2016 IEEE 16th International Conference*, pages 407–411. IEEE.

[114] Procaci, T. B., Siqueira, S. W., Nunes, and Pereira, B. (2018a). Learning in communities: How do outstanding users differ from other users? In *Advanced Learning Technologies (ICALT), 2018 IEEE 18th International Conference*. IEEE.

[115] Procaci, T. B., Siqueira, S. W., Nunes, and Pereira, B. (2019). Trust

investigation in communities using feature learning. In *Advanced Learning Technologies (ICALT), 2019 IEEE 19th International Conference*. IEEE.

[116] Procaci, T. B., Siqueira, S. W., Nunes, B. P., and Nurmikko-Fuller, T. (2017). Modelling experts behaviour in q&a communities to predict worthy discussions. In *Advanced Learning Technologies (ICALT), 2017 IEEE 17th International Conference*. IEEE.

[117] Procaci, T. B., Siqueira, S. W. M., Braz, M. H. L. B., and de Andrade, L. C. V. (2015d). How to find people who can help to answer a question? – analyses of metrics and machine learning in online communities. *Computers in Human Behavior*, 51, Part B:664 – 673. Computing for Human Learning, Behaviour and Collaboration in the Social and Mobile Networks Era.

[118] Procaci, T. B., Siqueira, S. W. M., and de Andrade, L. C. V. (2014a). Finding experts on facebook communities: Who knows more? *International Journal of Knowledge Society Research (IJKSR)*, 5(2):7–19.

[119] Procaci, T. B., Siqueira, S. W. M., and de Andrade, L. C. V. (2014b). Finding reliable people in online communities of questions and answers-analysis of metrics and scope reduction. In *ICEIS (2)*, pages 526–535.

[120] Procaci, T. B., Siqueira, S. W. M., Nunes, B. P., and Nurmikko-Fuller, T. (2018b). Experts and likely to be closed discussions in question and answer communities: An analytical overview. *Computers in Human Behavior*.

[121] Rabello, C. (2015). Interação e aprendizagem em sites de redes sociais: Uma análise a partir das concepções sócio-históricas de vygostky e bakhtin. *Revista Brasileira de Linguística Aplicada*, 15(3):735–760.

[122] Ratkiewicz, J., Conover, M., Meiss, M. R., Gonçalves, B., Flammini, A., and Menczer, F. (2011). Detecting and tracking political abuse in social media. *ICWSM*, 11:297–304.

[123] Razali, N. M., Wah, Y. B., et al. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of statistical modeling and analytics*, 2(1):21–33.

[124] RECUERO, R. Redes sociais na internet–porto alegre: Sulina, 2009. *Coleção Cibercultura*, 191.

[125] Rocha, E., Pimentel, M., and Diniz, M. (2017). Caracterização da participação dos usuários em sessões educacionais de bate-papo. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 1717.

[126] Rocha, E. B., Pimentel, M., and Diniz, M. C. (2015). Quantidade de participantes em bate-papo educacional: um modelo baseado em teoria de filas.

[127] Romero, C., López, M.-I., Luna, J.-M., and Ventura, S. (2013). Predicting students' final performance from participation in on-line discussion forums. *Computers & Education*, 68:458–472.

[128] Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., and Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, 111(52):E5616–E5622.

[129] Santos, E. and Silva, M. (2007). A pedagogia da transmissão e a sala de aula interativa. *Algumas vias para entretecer o pensar e o agir. Curitiba: Senar*, pages 17–35.

[130] Schroer, J. and Hertel, G. (2009). Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. *Media Psychology*, 12(1):96–120.

[131] Senter, R. and Smith, E. A. (1967). Automated readability index. Technical report, CINCINNATI UNIV OH.

[132] Shah, C., Oh, S., and Oh, J. S. (2009). Research agenda for social q&a. *Library & Information Science Research*, 31(4):205–209.

[133] Shaw, R.-S. (2012). A study of the relationships among learning styles, participation types, and performance in programming language learning supported by online forums. *Computers & Education*, 58(1):111–120.

[134] Shum, S. B. and Ferguson, R. (2012). Social learning analytics. *Educational technology & society*, 15(3):3–26.

[135] Siemens, G. (2006). Connectivism: Learning theory or pastime of the self-amused. *Manitoba, Canada: Learning Technologies Centre*.

[136] Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, page 0002764213498851.

[137] Siemens, G. (2015). Connectivism: A learning theory for the digital age.

[138] Siemens, G. and Weller, M. (2011). Higher education and the promises and perils of social network. *Revista de Universidad y Sociedad del Conocimiento (RUSC)*, 8(1):164–170.

[139] Silva, M. (2003). Pedagogia do parangolé: novo paradigma em educação presencial e online.

[140] Silva, M. (2008). Cibercultura e educação: a comunicação na sala de aula presencial e online. *Revista FAMECOS: mídia, cultura e tecnologia*, (37).

[141] Silva, M. and Santos, E. (2009). Conteúdos de aprendizagem na educação on-line: inspirar-se no hipertexto. *Educação & Linguagem*, 12(19):124–142.

[142] Simoes, A. J. G. and Hidalgo, C. A. (2011). The economic complexity observatory: An analytical tool for understanding the dynamics of economic development. In *Scalable Integration of Analytics and Visualization*.

[143] Smith, S. and Caruso, J. (2010). The ecar study of undergraduate students and information technology, 2010 (research study, vol. 6). boulder, co: Educause center for applied research.

[144] Souza, C., Magalhães, J., Costa, E., and Fechine, J. (2013). Social query: a query routing system for twitter. In *Proc. 8th International Conference on Internet and Web Applications and Services (ICIW)*, pages 147–153.

[145] Spyer, J. (2007). *Conectado: o que a internet fez com você e o que você pode fazer com ela*. Zahar.

[146] Srba, I., Grznar, M., and Bielikova, M. (2015). Utilizing non-qa data to improve questions routing for users with low qa activity in cqa. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 129–136. ACM.

[147] Streeter, L. A. and Lochbaum, K. E. (1988). Wo knows: a system based on automatic representation of semantic structure. In *RIAO 88:(Recherche d'Information Assistee par Ordinateur). Conference*, pages 380–388.

[148] Subramani, M. R. and Rajagopalan, B. (2003). Knowledge-sharing and influence in online social networks via viral marketing. *Communications of the ACM*, 46(12):300–307.

[149] Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee.

[150] Twitter (2016). Twitter reports second quarter 2016 results. Available at `https://investor.twitterinc.com/results.cfm?Quarter=2&Year=2016`.

[151] Ugulino, W., Marques, A. d. M., Pimentel, M., and Siqueira, S. W. (2009). Avaliação colaborativa: um estudo com a ferramenta moodle workshop. *XX Simpósio Brasileiro de Informática na Educação, Florianópolis-SC-2009, ISSN*, pages 2176–4301.

[152] Vasilescu, B., Capiluppi, A., and Serebrenik, A. (2012). Gender, representation and online participation: A quantitative study of stackoverflow. In *Social Informatics (SocialInformatics), 2012 International Conference on*, pages 332–338. IEEE.

[153] Vinayagam, A., Gibson, T. E., Lee, H.-J., Yilmazel, B., Roesel, C., Hu, Y., Kwon, Y., Sharma, A., Liu, Y.-Y., Perrimon, N., et al. (2016). Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets. *Proceedings of the National Academy of Sciences*, page 201603992.

[154] Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard university press.

[155] Wang, G., Gill, K., Mohanlal, M., Zheng, H., and Zhao, B. Y. (2013a). Wisdom in the social crowd: an analysis of quora. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1341–1352. ACM.

[156] Wang, G. A., Jiao, J., Abrahams, A. S., Fan, W., and Zhang, Z. (2013b). Expertrank: A topic-aware expert finding algorithm for online knowledge communities. *Decision Support Systems*, 54(3):1442–1451.

[157] Wasko, M. M., Faraj, S., and Teigland, R. (2004). Collective action and knowledge contribution in electronic networks of practice. *Journal of the Association for Information Systems*, 5(11):15.

[158] Wasserman, S. and Faust, K. (1994). *Social network analysis: Methods and applications*, volume 8. Cambridge university press.

[159] Wegner, D. M. and Ward, A. F. (2013). How google is changing your brain. *Scientific American*, 309(6):58–61.

[160] Wei, C.-T. and Young, S. S. (2011). Investigating the role and potentials of using web2. 0 in music education from student perspective. In *Advanced Learning Technologies (ICALT), 2011 11th IEEE International Conference on*, pages 344–346. IEEE.

[161] Wenger, E. C. and Snyder, W. M. (2000). Communities of practice: The organizational frontier. *Harvard business review*, 78(1):139–146.

[162] Yang, J., Peng, S., Wang, L., and Wu, B. (2016). Finding experts in community question answering based on topic-sensitive link analysis. In *Data Science in Cyberspace (DSC), IEEE International Conference on*, pages 54–60. IEEE.

[163] Yeniterzi, R. and Callan, J. (2015). Moving from static to dynamic modeling of expertise for question routing in cqa sites. In *ICWSM*, pages 702–705.

[164] Yu, T.-K., Lu, L.-C., and Liu, T.-F. (2010). Exploring factors that influence knowledge sharing behavior via weblogs. *Computers in human behavior*, 26(1):32–41.

[165] Zhang, J., Ackerman, M. S., and Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM.

[166] Zozzoli, R. M. D. (2012). A noção de compreensão responsiva ativa no ensino e na aprendizagem. *Bakhtiniana. Revista de Estudos do Discurso*, 7(1):253–269.

Apêndices

A. Dados Utilizados nos Experimentos

Para a realização dos experimentos mostrados nos Capítulos 3 e 4 um dump[1] do Stackexchange com dados das comunidades BQA e CQA foi extraído. No entanto, para as análises desta tese, a estruturação deste dump se fez necessário. Assim, se criou um programa escrito em Java, versão 8, para ler este dump e colocá-lo em um banco de dados relacional (PostgresSQL, versão 9.5). O autor desta pesquisa entendeu que seria mais fácil a realização das análises por meio de um banco de dados que suporte consultas SQL[2].

Em síntese, para estruturar o dump referido se deve: (i) Acessar o repositório GitHub[3]. Neste repositório se encontram todos os arquivos necessários para a estruturação do dump; (ii) Considerando o repositório GitHub citado, deve-se criar o schema do banco de dados PostgresSQL, executando o arquivo localizado em 'schema/schema.sql'; (iii) Depois de criado o schema do banco de dados, execute a aplicação Java, utilizando uma IDE[4]. No caso do autor desta pesquisa, foi utilizada a IDE IntelliJ[5]. Ademais, dentro do arquivo application.properties, se deve configurar o endereço do banco de dados PostgresSQL, bem como o usuário e a senha do mesmo. É importante mencionar

[1] `https://archive.org/download/stackexchange`
[2] Structured Query Language
[3] `https://bit.ly/2TGRPwj`
[4] Integrated Development Environment
[5] `https://www.jetbrains.com/idea/download/`

que o projeto Java teve todo seu processo de build realizado através do Maven[6] versão 3. Desta forma, é fundamental certificar se a IDE escolhida para da execução a aplicação suporta o Maven.

Por fim, se ressalta que o procedimento descrito foi configuração e executado no sistema operacional Ubuntu 16.04.4 LTS.

[6] `https://maven.apache.org/download.cgi`

B. Estudo Empírico do Capítulo 3

Os dados para a realização dos experimentos descritos no Capítulo 3 estão disponíveis em um repositório GitHub[1]. Este repositório consiste em um conjunto de scripts escritos na linguagem R, versão 3.4.2, que usam os dados estruturados descritos no Apêndice A. Assim, para as análises, os dados foram recuperados por meio de consultas SQLs e organizados em arquivos com extensão '.csv'. Desta forma, os scripts R realizam a leitura dos arquivos '.csv' e as análises estatísticas são feitas como descritas na Seção 3.1.5.

Além disso, o repositório GitHub referido neste Apêndice está estruturado em pastas, sendo cada uma responsável por um tipo de análise. Por fim, a síntese dos experimentos relatados neste Apêndice e no Capítulo 3 pode ser encontrada no artigo 'Learning in Communities: How Do Outstanding Users Differ From Other Users?', publicado e apresentado no evento IEEE International Conference on Advanced Learning Technologies - ICALT 2018 [114].

[1] `https://github.com/thiagoprocaci/diff-ourstanding-ordinary`

C. Métodos Feature Learning do Capítulo 4

No Capítulo 4, foram propostos dois métodos de *feature learning* (*Simple Walk* e *Go Ahead When Necessary*), bem como a comparação destes com outros da literatura. Assim, no repositório GitHub[1] tem as instruções para a execução dos experimentos.

Os experimentos foram feitos usando as linguagens Python versão 2.7, o R versão 3.4.2 e testado no sistema operacional Ubuntu 16.04.4 LTS. Em resumo, os métodos de *feature learning* propostos foram implementados em Python, devido à quantidade de bibliotecas que facilitam tal implementação. Depois da implementação, a execução destes foram em cima dos dados descritos no Apêndice A. Precisamente, os dados referentes as interações dos usuários foram extraídos do banco de dados relacional e transformados em um grafo, que são as arquivos 'biology.edgelist.csv' e 'chemistry.edgelist.cvs' do repositório citado. Isto é, os métodos de *feature learning* tem como entradas o grafo e cada nó será representado por um vetor de números.

Após a geração dos vetores, foi executado o classificador descrito na Seção 4.3, por meio de um script escrito em R. Ou seja, se usou o Python para gerar os vetores com as feature e o R para usar os vetores gerados e classificar o usuário. É importante mencionar que o encadeamento entre chamadas de

[1] `https://bit.ly/2HAt7HN`

scritps Python e R foram feitos através de um Shell script.

Ressalta-se que para os demais métodos de *feature learning* da literatura, primeiro, forneceu-se como entrada os grafos das comunidades e, depois, se chamou o script R para realizar as classificações. Neste caso, não foi implementado qualquer código em Python, isto é, somente se usou as implementações já disponibilizadas.

Por fim, a síntese dos experimentos relatados neste Apêndice e no Capítulo 4 pode ser encontrada no artigo 'Trust Investigation in Communities Using Feature Learning', do evento IEEE International Conference on Advanced Learning Technologies - ICALT 2019 [115].