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

Usuários Confiáveis em Comunidades Online: Um Estudo Empírico Envolvendo Análise de Métricas e Aprendizado de Máquina

Thiago Baesso Procaci

Orientadores:

Leila Cristina Vasconcelos de Andrade
Sean Wolfgand Matsui Siqueira

Rio de Janeiro – RJ, Brasil

AGOSTO, 2014

Usuários Confiáveis em Comunidades Online: Um Estudo Empírico Envolvendo Análise de Métricas e Aprendizado de Máquina

Thiago Baesso Procaci

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

Aprovada por:

Leila Cristina Vasconcelos de Andrade, D.Sc – UNIRIO

Sean Wolfgand Matsui Siqueira, D.Sc – UNIRIO

Kate Cerqueira Revoredo, D.Sc – UNIRIO

Jonice de Oliveira Sampaio, D.Sc – UFRJ

Bernardo Pereira Nunes, D.Sc – PUC-Rio

Rio de Janeiro – RJ, Brasil

AGOSTO, 2014

Procaci, Thiago Baesso.
P963 Usuários confiáveis em comunidades online: um estudo empírico envolvendo análise de métricas e aprendizado de máquina / Thiago Baesso Procaci , 2014.
101 f. ; 30 cm

Orientadora: Leila Cristina Vasconcelos de Andrade.
Coorientador: Sean Wolfgand Matsui Siqueira.
Dissertação (Mestrado em Informática) - Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, 2014.

1. Redes neurais (Computação). 2. Aprendizado do computador. 3. Grupos de discussão pela internet. 4. Algoritmos computacionais. I. Andrade, Leila Cristina Vasconcelos de. II. Siqueira, Sean Wolfgand Matsui. III. Universidade Federal do Estado do Rio de Janeiro. Centro de Ciências Exatas e Tecnológicas. Curso de Mestrado em Informática. IV. Título.

CDD – 006.32

Aos amigos presentes.

Agradecimentos

À minha família pelo apoio incondicional. Meus momentos de ausência na família foram constantes e longos durante os estudos. Contudo, sempre contei com a compressão familiar. No fundo acho que eles entenderam que a construção um trabalho como este requer dedicação, concentração e também um pouco de isolamento.

Agradeço ao professor Sean pelo acompanhamento constante do trabalho, generosidade e competência intelectual. Foram longas conversas e muitos e-mails trocados sobre diversos assuntos até chegar a concepção final do trabalho. Foi através de seus ensinamentos que entendi como se deve conduzir uma pesquisa.

Á professora Leila por acreditar em meu trabalho. Desde o início, sempre me recebeu com um belo sorriso e sempre acreditou que eu poderia fazer um bom trabalho. Em especial, agradeço às aulas de inteligência artificial que serviram de inspiração para a construção de parte desta dissertação.

Agradeço também a professora Maria Helena pela ajuda nas revisões dos artigos que compõem esta dissertação. Sem dúvidas, suas revisões serviram de orientação para boa parte deste trabalho.

Aos amigos do grupo de estudo do mestrado, Cristiane, Vanessa, Fernando, Michel, Gustavo, Rodrigo pelas ótimas ideias. É difícil achar um lugar com tanta gente boa reunida.

Aos professores do PPGI que tão bem me prepararam para a realização desta dissertação.

Aos funcionários da secretaria do PPGI que sempre muito bem me atenderam quando precisei.

A FAPERJ pelo apoio nas publicações em periódicos e na apresentação de artigos em eventos, através do projeto E-26-102.256/2013 – Associa: Explorando um Ambiente Semântico e Social de Ensino-Aprendizagem.

Aos membros da Banca Examinadora pelo trabalho de avaliação.

Aos amigos do trabalho pelas boas conversas nos almoços e nas confraternizações.

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

PROCACI, T. B. Usuários Confiáveis em Comunidades Online: Um Estudo Empírico Envolvendo Análise de Métricas e Aprendizado de Máquina. UNIRIO, 2014. Dissertação de Mestrado. Departamento de Informática Aplicada, UNIRIO.

Resumo

As comunidades online tornaram-se lugares importantes para usuários trocarem informações e construírem novos conhecimentos. Nessas comunidades, os participantes geralmente enviam e respondem perguntas, possibilitando aos usuários aprender uns com os outros. Contudo, alguns problemas podem ocorrer, tais como não obter respostas ou mesmo receber respostas erradas. Uma das alternativas para minimizar tais problemas é identificar pessoas que estão dispostas a ajudar e que são capazes de fornecer boas respostas, que neste trabalho as chamamos de usuários confiáveis. Foram investigados vários atributos (métricas) relacionados aos usuários de cinco comunidades online reais, com o objetivo de encontrar quais são as evidências ou dados que permitem dizer quais são os usuários mais confiáveis. Além disto, foi proposto o uso de um modelo e um algoritmo de aprendizado de máquina (uma rede neural artificial e um algoritmo de agrupamento) utilizados com os atributos dos usuários para encontrar os confiáveis das comunidades. Os resultados mostram que a utilização de uma rede neural artificial é uma boa abordagem, pois, cerca de 90% dos usuários foram corretamente identificados como confiáveis. Por outro lado, o algoritmo de agrupamento possibilita encontrar grupos de usuários confiáveis com mais facilidade.

Palavras-chave: Aprendizagem colaborativa; comunidades online; especialistas; reputação; estrutura Bow Tie; aprendizado de máquina.

Abstract

Online communities have become important places for users to exchange information and build knowledge. In these communities, people ask and answer questions, learn with each other, but some problems may occur such as not getting an answer or getting contradictory ones. In order to increase the responsiveness of the communities, it would be important to identify people who are willing to help and who provide good answers in such communities, whom we call reliable users. We investigated various components of online communities and users' attributes looking for a correlation between these characteristics and the users' reputation in these communities. After that, we proposed the usage of two machine learning approaches, artificial neural network and clustering algorithm, with the users' attributes for finding reliable sources. The results show that the usage of an artificial neural network is a good approach as around 90% of the users were correctly identified while the clustering algorithm makes to find groups of reliable users more easily.

Keywords: collaborative learning; online communities; experts; reputation; Bow Tie structure; machine learning.

Índice

Lista de Figuras

Lista de Tabelas

Lista de Fórmulas

1. Introdução da Pesquisa

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

1.1. Introdução

Vive-se nos dias de hoje uma revolução social, a revolução da Internet. Os computadores em rede, desenvolvidos a partir da metade do século XX, rapidamente se disseminaram por todo o sistema social e, desde então, vêm provocando profundas transformações em todos os setores da vida contemporânea (NICOLACI-DA-COSTA & PIMENTEL, 2011). Com o advento da Web na década de 1990 e a ampliação das capacidades e possibilidades de uso da Internet, criou-se um novo lugar para interações humanas denominado espaço digital. Esse novo lugar é o espaço da nova sociedade em rede, um espaço para interações humanas que possibilita vivenciar experiências e tem o grande poder de atrair e manter membros (LÉVY, 1999).

O acesso facilitado a tecnologias nas últimas décadas como, por exemplo, ao computador pessoal, promoveu mudanças no modo de pensar, de comunicar e até mesmo de viver das pessoas (WEI & YOUNG, 2011). 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 (CASTELLS, 1999) (NICOLACI-DA-COSTA, 2002). Desde então, muitas coisas mudaram, novos hábitos e atividades surgiram. Hoje em dia, pessoas parecem escrever menos cartas, comprar menos discos de músicas e, até mesmo, interagir menos pessoalmente. Em vez disto, pessoas passaram a enviar e-mails, escutar rádios online, assistir vídeos na Web, conversarem em chats etc. Além disso, novos vocábulos foram criados e antigos termos adquiriram novos significados. Nota-se também uma proliferação de expressões que antes eram inexistentes ou tinham significados diferentes. Dentre essas expressões pode-se citar: o copiar, colar, curtir, formatar, configurar, seguir etc.

O conceito de redes sociais foi também incorporado nesse novo espaço tecnológico. As redes sociais são tidas como as estruturas básicas que compõem uma sociedade e são formadas por pessoas e seus relacionamentos. Em outras palavras, todas as pessoas com quem nos relacionamos em algum momento da vida, fazem parte de nossa rede social. Contudo, com a expansão da Web, as

redes sociais passaram gradativamente a fazer parte desse espaço digital. As redes sociais online permitem conectar pessoas de forma descentralizada na Web. Nelas, as pessoas geralmente são descritas por um perfil, o qual contempla informações pessoais, hábitos e preferências, disponíveis e utilizadas como referência para a criação de novos laços sociais (LIEBOWITZ, 2007) (RHEINGOLD, 2000) (FRITZEN et al., 2013). Atualmente, as redes sociais online já fazem parte da rotina de uma parcela significativa da população mundial. Parece existir uma tendência mundial de pessoas se manterem conectadas a todo momento, seja através de computadores convencionais ou através de dispositivos móveis, para interagirem. Segundo NILSEN (2009), o uso das redes sociais e dos blogs já representam a quarta atividade mais realizada na Web, e já atinge 66,8% da população mundial. De acordo com COMSCORE (2010), houve um aumento de 51% no tráfego de Internet relacionado com o acesso às redes sociais online de 2009 para 2010 no Brasil. O Twitter[1], uma famosa rede social para microblogging[2], ilustra o fenômeno rápido e intenso de crescimento das redes sociais online. Em maio de 2008, o Twitter possuía apenas 1,2 milhão de usuários, alcançou 18,2 milhões em 2009, passando para 105 milhões em 2010 e chegou a 200 milhões de contas em janeiro de 2011 (MEIRA et al., 2011). Além disso, dados de 2011 (COMSCORE, 2011), apontam que o Brasil é o segundo país em volume de acesso nesta rede social, representando 21,8% do total de acessos ao Twitter.

Neste cenário, observa-se o surgimento de uma nova sociedade composta por indivíduos que nasceram e cresceram em meio a expansão do uso da tecnologia em suas atividades diárias. Tais indivíduos entendem a tecnologia em sua rotina de forma mais natural que gerações anteriores, utilizando tecnologias no lazer, no trabalho e também na educação (PINHATI, 2013). Alguns estudiosos denominam este novo indivíduo de "homo digitus" (NICOLACI-DA-COSTA & PIMENTEL, 2011) ou de Nativo Digital (PRENSKY et al., 2001). Essa nova classificação do indivíduo é oriunda de classificações de anteriores como a Geração de Baby Boomers (os nascidos entre o final da Segunda Guerra Mundial e o início dos anos 60), a Geração X (nascidos entre o início de 1960 e o final da década de 1970) a Geração Y (os que nasceram após 1980) e de Geração Z (os que nasceram após 1990) (GRAIL RESEARCH, 2010). Ainda existem discussões a respeito destas terminologias e classificações, bem como das próximas gerações, contudo, também existe um consenso em uma tendência cada vez maior da infiltração de tecnologias na vida das pessoas.

Com as constantes evoluções tecnológicas, em especial a expansão da Web, na área da

[1] Site: https://twitter.com/

[2] Forma de publicação de blog que permite aos usuários que façam atualizações breves de texto (geralmente com menos de 200 caracteres) e publicá-las para que sejam vistas publicamente ou apenas por um grupo restrito escolhido pelo usuário.

educação houve também consequências devido a esses avanços. Segundo RÓZEWSKI et al. (2011), por conta da evolução tecnológica, sistemas para apoiar o ensino e a aprendizagem surgiram e, por intermédio da Internet, foram fortemente disseminados. Tais sistemas propiciaram não apenas o apoio a aulas presenciais, mas também o crescimento e evolução da educação à distância, que foi uma das áreas da educação que mais rápido se desenvolveu no final do século XX (GILBERT et al., 2007). Assim, surgiu uma nova modalidade para educar denominada educação online, que SANTOS e SILVA (2009) afirmam que não é uma evolução do ensino a distância clássico ou do ensino presencial via suportes tradicionais tais como materiais impressos, rádio ou TV. A educação online é o conjunto de ações de ensino-aprendizagem ou atos de currículo mediados por interfaces digitais que potencializam práticas comunicacionais interativas e hipertextuais. Nessa nova educação, é desejável que a interação e aprendizagem ocorra de todos para todos. Em outras palavras, o professor deixar de ser somente o único polo de transmissão do conhecimento. De transmissor, o professor passar a ser um agente provocador de situações, arquiteto de percursos e mobilizador da inteligência coletiva (SANTOS & SILVA, 2009). Aparentemente, o conceito de educação online está fortemente ligado com o conceito de redes sociais ou comunidades online, uma vez que, essa modalidade de educação favorece a aprendizagem com base na troca, diálogo, participação e colaboração. Independente da modalidade de educação, não se pode negar que as redes sociais online começaram a fazer parte do processo de ensino e aprendizagem recentemente.

A Universidade Federal do Estado do Rio de Janeiro (UNIRIO), no início do ano de 2012, iniciou o uso de uma rede social online para apoiar algumas disciplinas presenciais do Programa de Pós-Graduação em Informática. No primeiro semestre de 2012, foi criado um grupo no Facebook[3] denominado Cibercultura UNIRIO 2012.1[4]. Esse grupo tinha como objetivo permitir uma maior interação entre professores e alunos através de um ambiente onde todos produziam e consumiam conteúdos. Nesse grupo, os professores geralmente atuavam como provocadores de situações, através do compartilhando conteúdos e ideias. Os alunos, em geral, comentavam sobre esses conteúdos deixando seu ponto de vista ou mesmo compartilhando novos conteúdos. Depois deste, outros grupos online apareceram no Programa de Pós-Graduação em Informática da UNIRIO para o mesmo fim. Além disso, várias universidades e grupos do mundo tem utilizado o Facebook como ferramenta de apoio a educação em vez de utilizarem plataformas de ensino tradicionais (ENGLISH & DUNCAN-HOWELL, 2008) (DENG & TAVARES, 2013). PINHATI (2013) propôs a construção um aplicativo no Facebook para apoiar o ensino e a aprendizagem de música. Este

[3] Site: https://www.facebook.com/
[4] Site: https://www.facebook.com/groups/384634338233143/

aplicativo era suportado por uma arquitetura formal para o desenvolvimento de ambientes virtuais e um modelo para criação facilitada de objetos de aprendizagem[5], ambos especializados para a área da música. Esse estudo envolveu a participação de 27 alunos do ensino médio de uma escola pública no Rio de Janeiro. A partir de análises realizadas, foi concluído que o oferecimento de recursos multimídia e sociais, garantidos pelo uso do aplicativo como guia, tanto na construção do ambiente quanto na dos objetos de aprendizagem, influenciam indiretamente na intenção de uso de ambientes virtuais de aprendizagem para educação musical pelos alunos. MARCON et al. (2012) refletem sobre a utilização da rede social Facebook como uma arquitetura pedagógica, conceito que pode ser definido como a união de software educacional e abordagem pedagógica em um só ecossistema a ser utilizado no processo de aprendizagem (CARVALHO et al., 2005). No trabalho, os autores relatam a experiência do uso do Facebook em uma disciplina de pós-graduação. Como pontos positivos, indicam a veloz manifestação dos integrantes do grupo através da comunicação pela ferramenta mural, a existência de ferramentas que podem ser utilizadas em abordagens pedagógicas oferecidas pelo ambiente e a utilização de um espaço (Facebook) já conhecido pelos participantes através de outras atividades, como o lazer. Como ponto negativo, os autores citam o excesso de informações a que os alunos ficam expostos. Isso faz com que os participantes necessitem estar muito cientes dos objetivos educacionais a serem alcançados e focados no seu atingimento.

Não se pode negar as consequências na área da educação devido as mudanças tecnológicas. Novos sistemas para apoiar a educação apareceram e promoveram uma abordagem mais colaborativa para o ensino e a aprendizagem. Contudo, neste cenário, emerge também uma forma mais informal de ensino e aprendizagem. Grupos e comunidades online começaram a aparecer na Web onde a presença de um tutor formal (professor) para conduzir o processo de ensino não necessariamente existe. Essas comunidades informais da Internet são lugares onde pessoas compartilham interesses comuns e voluntariamente trabalham para expandir a sua compreensão sobre um domínio do conhecimento (ALAN et al., 2013). Em geral, os membros dessas comunidades não se conhecem, podem ser identificados por pseudônimos e estão dispostos a ajudar uns aos outros por diversas razões: altruísmo, reputação, a reciprocidade esperada e os benefícios da aprendizagem (KOLLOCK, 1999) (LAKHANI & VON HIPPEL, 2000). É importante ressaltar que redes sociais online e comunidades online são conceitos diferentes. Redes sociais online são espaços digitais onde pessoas simplesmente podem interagir entre si. Já comunidades online, há também interação entre pessoas, contudo, geralmente, tais pessoas compartilham interesses comuns.

[5] Objeto de aprendizagem é geralmente definido como qualquer entidade, digital ou não, que pode ser usada para a aprendizagem, educação e treinamento.

Um fato interessante é que algumas redes sociais permitem a criação de comunidades como, por exemplo, os grupos (comunidades) do Facebook (rede social).

As comunidades online destinadas ao compartilhamento de conhecimento de maneira informal são fortemente dependentes de seus membros cooperantes. São através dos membros e de suas participações que a comunidade cresce e, como consequência, maiores são as chances de colaborações bem-sucedidas e construções de conhecimentos. Em outras palavras, através das colaborações dos membros torna-se possível a construção de um corpo de conhecimento na comunidade. Esse corpo de conhecimento geralmente são os registros escritos e publicados pelos usuários em uma comunidade. Todavia, YIMAM-SEID e KOBSA (2003) afirmam que o compartilhamento de conhecimento não é eficiente tendo somente o conhecimento exposto em algum ambiente. Segundo os autores, para que esse compartilhamento seja eficiente, é necessário ter exposto e acessível não somente o conhecimento produzido, mas também os especialistas que geraram esse conhecimento. A justificativa para isso é que, muitas vezes, um conhecimento escrito pode estar ambíguo ou mesmo incompleto. Desta maneira, um especialista pode ajudar a clarificar algum ponto duvidoso no assunto e indicar caminhos. Além disso, diferentemente das organizações tradicionais, onde aqueles com conhecimento único e específico sobre um assunto são considerados especialistas, a definição de especialistas em comunidades online é mais ampla, pois cada membro pode ter um grau de especialização em uma determinada área (ACKERMAN et al., 2002). Na educação tradicional esses especialistas são conhecidos como tutores ou professores.

Na educação tradicional o professor é uma peça fundamental no processo de aprendizagem. Ele é a pessoa que mostra caminhos, ajuda alunos a solucionarem problemas, cria situações para a construção de novos conhecimentos. A importância do professor na educação é indiscutível no mundo atual. Fatos que reforçam essa importância podem ser encontrados, por exemplo, nas medidas que o governo eventualmente toma em favor dos professores. O governo brasileiro vem aos poucos melhorando as condições de trabalho dos professores e incentivando o contínuo aperfeiçoamento. Por exemplo, em 2008 foi formulada a Lei 11.738/08 que instituiu o piso salarial nacional para professores do magistério (BRASIL, 2008). O Plano Nacional de Educação estabelecido pelo governo brasileiro tem como uma das linhas de atuação a valorização dos profissionais da educação, ressaltando a necessidade de uma atenção especial à formação inicial e continuada (MEC, 2000). Mesmo que, no Brasil de hoje, os professores ainda não estejam nas posições que realmente mereçam, esses tipos de medidas adotadas só reforçam a importância deste profissional. Relatos desde a antiguidade clássica já reforçam a importância do professor no processo de ensino e aprendizagem. Por exemplo, o filósofo e matemático grego Pitágoras, nascido

por volta de 571 A.C. e 570 A.C., fundou em vida a famosa escola Pitagórica com o objetivo de passar seus conhecimentos adquiridos aos seus discípulos e evoluí-los (SINGH, 1999). Pitágoras acreditava que educar as crianças através de tutorias era a melhor maneira para não ser preciso puni-las na idade adulta. Além disso, através de sua escola e sua tutoria, grandes descobertas sobre os números e suas propriedades foram realizadas.

Um pouco diferente da educação tradicional, nas comunidades online, onde a aprendizagem é informal, a forma como se aprende é mais interativa e colaborativa, pois, os participantes necessariamente aprendem uns com os outros. Contudo, nessas comunidades, existem membros que são mais experientes, cujas opiniões acabam norteando debates ou questões para rumos mais promissores, tornando assim o processo de aprendizagem mais eficaz. Esses membros especiais dessas comunidades são comumente chamados de especialistas, tutores ou mentores. Estes têm sido um grande alvo de pesquisas, uma vez que, a prévia identificação de tais membros especiais pode tornar possível, por exemplo, direcionar questões para pessoas que realmente são capazes de resolvê-las.

Além disso, inspiradas na importância do professor na educação tradicional, sugiram pesquisas que visam identificar usuários que faziam o papel de um tutor (ou especialista) em comunidades online. Dentre elas estão, por exemplo: abordagens baseadas em processamento de linguagem natural e mineração de texto (STREETER & LOCHBAUM, 1988) (KRULWICH & BURKEY, 1996), em métodos probabilísticos (DAVITZ et al., 2007) e em algoritmos de ranqueamento (DOM et al., 2003). A existência destas diversas abordagem demonstra, além do interesse da comunidade científica, a relevância que é dada por pesquisadores a este assunto.

É comum ver pessoas buscando por ajuda em comunidades online atualmente, principalmente quando elas se deparam com algum problema no qual uma simples busca por uma solução na Web não é suficiente. Nestes casos, tais pessoas geralmente procuram por alguém que possa orientá-las. Para ilustrar um cenário típico de uso de comunidades online, imagine que um aluno de um curso Sistemas de Informação queira iniciar um projeto utilizando a tecnologia Java. Entretanto, para esse aluno, o desenvolvimento Java é algo novo. Desta forma, ele encontra problemas ao tentar compilar sua primeira aplicação. Com objetivo de esclarecer suas dúvidas, o aluno tenta primeiramente fazer uma pesquisa rápida em um motor de busca da Web. Contudo, devido ao seu baixo nível de conhecimento sobre programação Java, ele não obtém resultados satisfatórios. Diante disso, ele decide procurar ajuda em uma comunidade online de perguntas e respostas com a finalidade de encontrar pessoas mais experientes que possam responder suas perguntas. Desta forma, o aluno posta sua pergunta e aguarda por respostas. É este o cenário em que

reside o contexto dessa pesquisa: a educação informal e a importância do tutor (ou alguém com mais experiência no assunto e que esteja disposto a colaborar) no processo de ensino e aprendizagem. A ideia é investigar abordagens neste contexto, com a finalidade de descobrir formas para identificar quem são os possíveis usuários que participam dessas comunidades online e podem fazer o papel de tutores, ou seja, apoiar a construção de conhecimento de outros participantes.

Diante do exposto, percebe-se a importância da elaboração de soluções que auxiliem na identificação desses usuários mais experientes em comunidades de ensino e aprendizagem informais.

1.2. Motivação

Além da relevância discutida na subseção anterior, destacam-se também como fatores motivadores deste trabalho a aprendizagem colaborativa, os benefícios da aprendizagem informal através de postagens de perguntas em comunidades online e a importância dada pela Ciência da Web às pesquisas relacionadas a análises de massa de dados da Internet.

1.2.1. Aprendizagem Colaborativa

Segundo FAGUNDES et al. (1999), o conhecimento não é um produto fixo e acabado, ele é construído num 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. Assim, a busca contínua pela adaptação ao meio físico e social leva ao aprendizado.

CASTRO et al. (2011) indicaram 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 em grupo.

A aprendizagem colaborativa tem sido defendida por educadores e praticadas por muitos professores nos diversos níveis escolares, do ensino fundamental à pós-graduação (MENEZES et al., 2008). Esta prática não é uma novidade e a disponibilidade das tecnologias de comunicação e de interação social tem contribuído para melhorias e adesão de novos interessados (CARVALHO et al., 2005). Além disso, os benefícios decorrentes das práticas pedagógicas baseadas na colaboração são inúmeros, dos quais se pode citar: a preparação para a vida em sociedade, o desenvolvimento do espírito crítico e a competência para resolver problemas de grande porte a partir de contribuições

individuais (CASTRO et al. 2011).

O surgimento da Web 2.0[6] 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 pedagógicas colaborativas. A popularização das mídias sociais como blogs, folksonomia, wiki, podcast[7] e redes sociais online, marcou um novo direcionamento para a geração de tecnologias Web, onde o foco central é a comunicação em pares, a troca de experiências, o compartilhamento e a construção coletiva (ANTTILA, 2006).

Um estudo realizado por SMITH et al. (2010) com 30.616 estudantes universitários (nativos digitais) dos Estados Unidos, constatou que 90,3% dos alunos dedicam tempo diariamente na utilização de redes sociais, que são importantes ambientes de colaboração online (SPYER, 2007). 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 (GOMES, 2012).

Dessa forma, diante da argumentação apresentada, entende-se como importante a realização de estudos em comunidades online utilizadas para fins de aprendizagem, onde pessoas se reúnem para debater ideias e buscarem soluções em conjunto para seus problemas.

1.2.2. Postagem de Perguntas em Comunidades Online (*Social Query*)

O processo de postagem de perguntas em uma comunidade online e espera por respostas é conhecido como *social query* (SOUZA et al., 2013) (MORRIS et al., 2010) (BANERJEE & BASU, 2008). *Social query* pode ser vista como uma alternativa aos motores de busca da Web. A justificativa disso está nos resultados dos motores de busca que, muitas vezes, são indesejados e incompletos. HUBERMAN et al. (2013) e MUI e WHORISKEY (2010) 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.

[6] 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.
[7] Podcast é o nome dado para comunicações em áudio, sem caráter formal, gravadas por usuários da Web e distribuídas em seus sites e blogs.

Motores de busca nem sempre são as melhores formas para buscar informações na Web, pois, seus resultados não necessariamente refletem o que se busca em um determinado momento (FRITZEN et al., 2013). Segundo HOROWITZ et al. (2010), alguns problemas são melhores resolvidos por pessoas: perguntas muito contextualizadas, pedidos de recomendação, pedidos de opiniões, conselhos etc. O motivo disto é que os sistemas computacionais podem desempenhar bem tarefas específicas em um ambiente conhecido e sem muitas mudanças. 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 (FRITZEN et al., 2013). Dada essa limitação, uma alternativa aos motores de busca para a resolução de problemas ou dúvidas são as comunidades online de perguntas e respostas como Stackoverflow[8], Quora[9] e Yahoo! Answers[10], onde os usuários perguntam e respondem de forma voluntária. 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 (MORRIS et al., 2010).

TEEVAN et al. (2010) apresentaram resultados confirmando que *social query* é um método viável para se obter respostas em uma comunidade online. Esse estudo foi realizado internamente na Microsoft, utilizando suas próprias ferramentas de comunicação, sendo 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. PAUL et al. (2013) fizeram estudos similares no Twitter, 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, diante dos dados expostos, considera-se importante o estudo de como identificar especialistas nas comunidades online de perguntas e respostas, devido a relevância e a eficiência do uso *social query* em alguns contextos para se obter soluções para problemas em comunidades online.

[8] Site: http://stackoverflow.com/
[9] Site: https://www.quora.com/
[10] Site: http://answers.yahoo.com/

1.2.3. Ciência da Web

Diante do cenário das comunidades online, um fato inevitável é a grande produção de conteúdo por parte dos usuários durante seus momentos de interação. O uso extenso da Web e as suas diversas interações tem chamado a atenção de estudiosos. Em 2006 surgiu uma nova área de pesquisa objetivando estudar a Web denominada Ciência da Web (BERNERS-LEE et al., 2006).

Neste novo domínio, a Web é o objeto a ser estudado e deixa de ser considerada somente uma tecnologia baseada em computadores destinada à comunicação e interação. LUCENA e MACULAN (2008) definem Ciência da Web como uma área que estuda todos os problemas associados a sistemas de informação descentralizados, englobando pessoas, *software*, *hardware* e suas múltiplas e complexas interações. O'HARA & HALL (2008) afirmam que, para poder entender a Web, é necessário estudar não apenas suas propriedades computacionais, mas também entender os contextos em que ela é usada.

Uma das grandes preocupações desta nova área reside em investigar formas de extrair informações da Web visando gerar um novo conhecimento. Esse novo conhecimento gerado tem como objetivo possibilitar um maior entendimento das interações entre usuários e apoiar algum tipo de decisão no mundo real. Evidentemente que, quando se fala em dados na Web, refere-se a uma grande quantidade de dados sendo, muitas vezes, inviável uma análise sem auxílio computacional.

Nesse contexto de análise de dados cuja relevância é exaltada pela Ciência da Web, se considera importante a realização da pesquisa desta dissertação, pois, seu foco reside em análise de massas de dados de comunidades online com a finalidade de gerar um novo conhecimento. Espera-se que esse novo conhecimento sirva de apoio para encontrar os usuários que estão dispostos a participarem e proverem boas respostas nessas comunidades.

1.3. Problema de Pesquisa

Apesar das vantagens do uso de *social query*, ela tem também suas limitações. Quando uma pergunta é postada em uma comunidade, alguns resultados não esperados podem ser encontrados, como: receber respostas erradas ou contraditórias; continuar recebendo respostas mesmo depois de o problema ser resolvido; e nunca receber uma resposta, uma vez que, algumas comunidades tendem a priorizar a visualização das postagens mais recentes (MORRIS et al., 2010).

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 (WASKO & FARAJ TEIGLAND, 2004). 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. Contudo, ainda assim, pessoas podem não encontrar soluções para seus problemas em fontes externas. Em comunidades online, por exemplo, usuários podem receber respostas incompletas ou contraditórias de pessoas inexperientes.

PAUL et al. (2013) demonstrou através de estudos que o uso de *social query* no Twitter sofre o "efeito da linha do tempo". Segundo esse estudo, utilizando o Twitter é possível se 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 do fato do Twitter priorizar a visualização de postagens mais recentes. Logo, é provável que alguns seguidores (usuários) nem fiquem sabendo da existência de uma determinada pergunta.

Uma das maneiras para minimizar algumas das limitações da *social query*, como respostas erradas ou nenhuma resposta, é encontrar as pessoas mais adequadas para responder a uma pergunta (especialistas ou possíveis tutores). Desta forma, a própria comunidade online poderá garantir que uma pergunta postada seja encaminha para um conjunto de especialistas previamente identificados. Assim, as chances de um usuário receber uma boa resposta podem aumentar.

Nesse contexto, as limitações da *social query* podem ser consideradas um problema indireto desta pesquisa. Este pode ser enunciado da seguinte forma: *O uso de social query para se obter respostas em comunidades online de perguntas e respostas tem limitações. Algumas dessas limitações são: receber respostas erradas ou não receber respostas.*

Contudo, a solução para o problema direto dessa dissertação consiste em descobrir uma maneira automática para encontrar os usuários que têm boas chances de fornecerem uma boa resposta em uma comunidade online. Essa solução, por sua vez, pode minimizar as limitações da *social query* (problema indireto). Algumas comunidades online até possuem mecanismos para identificarem os usuários com boas chances de fornecerem boas respostas, porém, esses mecanismos não são automáticos, uma vez que, dependem de avaliações constantes das participações dos usuários dessas comunidades. No capítulo 2 esse mecanismo não automático será discutido melhor.

Desta forma, embasado na motivação, relevância e no problema indireto apresentados, o problema direto deste trabalho pode ser enunciado como: *As comunidades online de hoje não fornecem mecanismos automáticos para identificarem quais usuários que têm maiores chance de responderem corretamente as perguntas postadas por outros usuários*.

1.4. Hipótese

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

SE forem extraídos atributos dos usuários de uma comunidade online de perguntas e respostas e estes atributos tiverem uma correlação no mínimo moderada com algum indicador de competência **ENTÃO s**erá possível identificar os usuários confiáveis da comunidade.

Detalhando a hipótese, quando se fala em extração de atributos dos usuários de uma comunidade online, refere-se ao conjunto de métricas que caracterizam um usuário ou a sua participação em um grupo. Por exemplo, em uma comunidade online, a quantidade de perguntas ou respostas são tipos de métricas que caracterizam a participação de um usuário. Entende-se também como atributos de um usuário as métricas que são derivadas através de métodos computacionais ou representações abstratas (estruturas de dados) que descrevem o usuário. Por exemplo, para representar as interações de usuários em uma rede, é comum utilizar grafos para este fim (ZHANG et al., 2007). Desta forma, a partir de um grafo, é possível extrair outras métricas como, por exemplo, o grau de entrada de um nó de um grafo que, por sua vez, pode ter um significado especial dependendo do contexto da representação. Também é entendido que, a escolha dos atributos extraídos dos usuários da comunidade, posteriormente analisados, não seja feito de forma aleatória para se alcançar o objetivo da hipótese. Tais atributos devem ter uma correlação estatística positiva e no mínimo moderada com algum indicador de competência. Ou seja, é preciso que o atributo seja uma possível evidência que o usuário é especialista ou não em algum assunto. Mais adiante, nos capítulos 2 e 3, será discutido melhor o indicador de competência utilizado nesta dissertação.

Além disso, entende-se que um usuário confiável como uma pessoa que tem condições de fornecer boas respostas em uma comunidade online de perguntas e respostas. Um usuário confiável pode ser um especialista em algum assunto, um possível professor ou mesmo um usuário mais experiente. A partir dessas premissas, será aplicada a solução proposta nessa dissertação com a finalidade de verificar a hipótese. Espera-se que com a solução proposta, seja possível encontrar com mais facilidade os usuários confiáveis de uma comunidade online destinada à aprendizagem informal.

1.5. Enfoque da Solução

Para o estudo e verificação da hipótese levantada, será proposta uma solução que consiste em um conjunto de procedimentos para identificação de usuários confiáveis em comunidades

online. Este conjunto de procedimentos é composto basicamente por três etapas: extração de dados, extração e análise de métricas e, por fim, identificação dos usuários confiáveis. Essas etapas foram aplicadas em comunidades online reais.

A etapa "extração de dados" consiste em extrair dados da comunidade online a ser analisada. Foi construído um *Web crawler* que tinha como objetivo extrair dados automaticamente das comunidades e salvá-los em arquivos para tratamentos posteriores. A etapa "extração e análise de métricas" consiste inicialmente na leitura dos dados salvos nos arquivos (oriundos da etapa anterior) e, em seguida, transformá-los em modelos de classes, com suas devidas instâncias, e em representações de grafos. A partir do modelo de classes e das representações de grafos construídas, será possível recuperar as métricas (atributos) que descrevem os usuários das comunidades. O modelo de classes e as representações de grafos são complementares. O modelo de classes irá conter todas as métricas relativas aos usuários das comunidades. Contudo, algumas dessas métricas dependem de uma representação de grafos para serem extraídas (por exemplo, os resultados de algoritmos que se aplicam somente em grafos). Desta forma, pode-se dizer que as instâncias do modelo de classes são também alimentadas por métricas extraídas da representação de grafos.

Além disso, ainda na etapa "extração e análise de métricas", cada comunidade analisada será dividida em partes (componentes), considerando os diferentes tipos de interações entre os usuários. A ideia dessa divisão é poder identificar lugares nas comunidades onde os usuários interagem em padrões similares. Uma vez divididas as comunidades, serão realizadas correlações estatísticas dos atributos (métricas dos usuários) considerando seu escopo geral (ou seja, considerando amostras aleatórias da comunidade) e seus escopos locais (considerando somente as partes das comunidades identificadas, de acordo com um padrão de interação). Todas essas correlações serão feitas com um indicador de competência previamente estabelecido. Em seguida, haverá uma análise das correlações gerais e locais obtidas. Desta forma, pretende-se identificar partes das comunidades onde as correlações locais mais se aproximam das correlações gerais. Uma vez feito isto, será possível identificar qual pedaço da comunidade (parte reduzida) melhor descreve a rede como um todo, considerando o contexto dos usuários confiáveis.

Por fim, a etapa "identificação" consiste em utilizar um método computacional que permita classificar os usuários em confiáveis ou não. No caso deste trabalho, foram escolhidas duas abordagens de aprendizado de máquina para este fim: uma rede neural artificial e um algoritmo de agrupamento. Basicamente, o funcionamento de uma rede neural artificial consiste em duas fases: treinamento e aplicação. Em síntese, a fase de treinamento tem como finalidade configurar a rede neural para, em seguida, utilizá-la na fase de aplicação. Na fase de treinamento dessa rede neural

serão utilizadas partes reduzidas das comunidades que melhor a representa como um todo. Depois disso, espera-se que essa rede neural artificial consiga automaticamente classificar qualquer usuário da comunidade. Já o algoritmo de agrupamento consiste na formação de grupos com objetos semelhantes (usuários). A ideia é verificar em quais grupos é possível encontrar os usuários confiáveis mais facilmente.

1.6. Objetivos

O objetivo principal desta pesquisa é definir um conjunto de procedimentos que permitam a identificação de usuários confiáveis em comunidade online destinadas à aprendizagem informal. Além deste objetivo, pretende-se nesta pesquisa explorar as seguintes questões:

- Quais são os atributos (métricas) de um usuário que permitem inferir que ele é um usuário confiável? Por quê?
- O que significam esses atributos dos usuários?
- É possível construir uma nova métrica para inferir que um usuário é confiável através de combinações de outras? Em quais contextos ela melhor se adéqua?
- Quais são as partes da comunidade onde os usuários interagem de forma parecida? O que isto significa?
- Os atributos dos usuários considerando partes específicas da comunidade permitem inferir que ele é um usuário confiável?
- Quais estratégias são úteis para conseguir classificar um usuário como confiável?

Esse trabalho também tem os seguintes objetivos secundários:

- Construção de um protótipo para a realização das análises necessárias nesta pesquisa.
- Comparar resultados obtidos neste trabalho com outros similares com o objetivo de verificar pontos em comum e diferenças.

1.7. Método

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

As cinco comunidades cujos dados foram coletados, possuem um mecanismo que

permitem aos usuários construírem a sua reputação na rede. A reputação de um usuário é construída por avaliações de outros usuários, baseadas nas participações na comunidade. Em outras palavras, um usuário tem suas perguntas e respostas avaliadas na comunidade. Desta forma, um usuário que responda ou faça perguntas de maneira coerente e que agregue valor a comunidade, é provável que ele tenha boas avaliações e, consequentemente, uma boa reputação. A reputação do usuário será o indicador de competência utilizado nessa pesquisa.

Depois de coletados os dados e extraídas as métricas das comunidades, estas serão correlacionadas estatisticamente com o indicador de competência. Uma vez feito, tanto a correlação geral quanto as correlações locais servirão de base para a escolha da parte das comunidades que será utilizada para configurar as técnicas de aprendizado de máquina escolhidas.

Por fim, para avaliar se as técnicas de aprendizado de máquina podem trazer bons resultados, seus valores de saída serão comparados com os indicadores de competência. Desta forma, espera-se que, por exemplo, caso um usuário seja classificado como confiável, o indicador de competência confirme esse resultado.

1.8. Organização da Dissertação

Essa dissertação está organizada em seis capítulos, sendo este o primeiro. O capítulo dois desta dissertação é destinado à apresentação de conceitos relacionados às comunidades online, compartilhamento de conhecimentos e usuários confiáveis. A ideia é que o capítulo dois dê uma visão geral e ampla sobre o tema abordado neste trabalho. Além disto, ainda serão apresentados os trabalhos relacionados a esta pesquisa e um estudo comparativo entre eles, mostrando suas similaridades e diferenças.

No capítulo três será conduzido um estudo empírico com o objetivo de descobrir quais métricas descrevem um usuário confiável de uma comunidade online. Para isto, serão realizadas medições, análises estatísticas e comparações com outros trabalhos visando colher evidências que reforçam que as métricas escolhidas podem de fato representar os usuários confiáveis.

O capítulo quatro será destinado à apresentação de uma estratégia que visa dividir as comunidades analisadas em partes, de acordo com os padrões de interação dos usuários. Busca-se investigar partes menores das comunidades onde pode ser mais fácil encontrar os usuários confiáveis.

O capítulo cinco mostrará como foi concebido o uso das técnicas de aprendizado de máquina para encontrar os usuários confiáveis de uma comunidade. Serão apresentados os

conceitos envolvidos e também as avaliações que foram feitas com a finalidade de averiguar se o uso de tais técnicas realmente traz bons resultados.

Por fim, o capítulo seis será destinado às conclusões e considerações finais.

2. Usuários Confiáveis em Comunidades Online

O objetivo deste capítulo é discorrer sobre comunidades online e as abordagens existentes para encontrar os seus usuários confiáveis. Serão apresentados conceitos e aspectos fundamentais que nortearam o desenvolvimento dessa pesquisa, dos quais se destacam: o conceito e objetivos de comunidades online, o compartilhamento de conhecimentos, a reputação de um usuário e os modelos computacionais já existentes para a identificação de usuários confiáveis em comunidades online.

2.1. Comunidades Online

As comunidades online são lugares onde os indivíduos se reúnem em um espaço na Web com o objetivo de discutir ideias, socializar ou mesmo pedir ajuda para outras pessoas (VASILESCU et al., 2012). Em geral, pessoas costumam se reunir para formarem grupos das mais diversas naturezas com a finalidade de promover debates sobre assuntos de seus interesses.

Hoje em dia, 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 diretamente por isso. Dentre essas atividades, se pode citar, por exemplo, a procura por novas maneiras para projetar ou refinar produtos (FÜLLER et al., 2007), a busca por ajuda para desenvolver ou depurar um novo software (HERTEL et al., 2003), a escrita de textos e espera críticas (SCHROER & HERTEL, 2009) ou mesmo a exposição ideias através de artes ou imagens (YU et al., 2010).

Atualmente, as comunidades online têm como público-alvo uma quantidade variada de tipos 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 (por exemplo, os usuários do LinkedIn[11]) ou mesmo um público específico, como profissionais de informática (por exemplo, os usuários da comunidade Stackoverflow). Apesar de grande parte dessas comunidades ser destinada 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[12]), outras são destinadas ao compartilhamento de projetos de software inteiros (como o

[11] Site: http://www.linkedin.com
[12] Site: http://snipplr.com

Github[13]e o Bitbucket[14]), algumas são usadas para compartilhar imagens ou fotografias (Flickr[15]). Outras comunidades têm como objetivo o compartilhamento de conhecimentos através da construção de conteúdos (como a Wikipédia[16]) ou através de perguntas e respostas (Yahoo! Answers, Quora, Stackoverflow).

Como se percebe, uma comunidade online pode ser utilizada para vários fins. Em especial, esta dissertação está interessada nas comunidades de perguntas e respostas utilizadas para o compartilhamento de conhecimentos.

2.1.1. Compartilhamento de Conhecimentos

O compartilhamento de conhecimento foi predominantemente estudado dentro do ambiente empresarial (DAVENPORT & PRUSAK, 1998) e, muitas vezes, com foco em equipes que trabalham geograficamente separadas (JADIN et al., 2012). Equipes com pessoas geograficamente dispersas, no caso das organizações empresariais, são tipicamente compostas por funcionários que trabalham em diferentes unidades organizacionais. Trabalhar em equipe significa trabalhar colaborativamente e, no caso das equipes geograficamente dispersas, seus membros acabam recorrendo a aplicativos baseados na Web, como e-mail ou chats, para se organizarem, comunicarem e compartilharem conhecimentos (HERTEL et al., 2005). Dentro do ambiente empresarial, essas equipes são criadas por um líder formal, com o objetivo de prover a organização necessária para viabilizar o trabalho colaborativo. Fora do contexto das organizações, os indivíduos também se organizam em grupos em comunidades online, porém, sem necessariamente ter a presença de um líder formal para isto.

ZIMMER (2001) realizou um estudo de caso com o objetivo de entender a criação e o compartilhamento de conhecimento dentro de uma empresa situada no Brasil cuja uma de suas equipes ficava geograficamente dispersas. A empresa desse estudo era da área de informática e a equipe do estudo de caso era composta por 23 pessoas, sendo que 19 pessoas dessa equipe se localizavam na cidade de Porto Alegre, no estado do Rio Grande do Sul, e 4 pessoas na cidade São Paulo, no estado de São Paulo. Neste trabalho foi concluído que o compartilhamento de conhecimentos é fundamental para a geração de vantagens competitivas. Contudo, apesar disto, as colaborações mediadas pelo computador não suprem a presença física das pessoas. Segundo ZIMMER (2001), a interação mediada pelo computador limita a riqueza da interação entre pessoas.

[13] Site: https://github.com/
[14] Site: https://bitbucket.org/
[15] Site: https://www.flickr.com/
[16] Site: http://www.wikipedia.org/

Apesar do compartilhamento de conhecimento ter sido amplamente estudado no contexto empresarial, atualmente, pesquisadores têm mudado um pouco o foco do estudo. Segundo HOLLOWAY et al. (2007), com o uso da Internet em larga escala, muitos estudos sobre compartilhamento de conhecimento têm como foco o contexto das redes sociais e das comunidades online. JADIN et al. (2012) afirmam que, nos últimos anos, as comunidades online se tornaram muito importantes 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 (dentro das organizações). Talvez, por essa razão, o foco do estudo sobre compartilhamento de conhecimento tenha mudado.

Com compartilhamento de conhecimentos em larga escala nas comunidades online, surgiram pesquisas para tentar entender como ele ocorre. Muitas dessas pesquisas são baseadas em técnicas de análises de redes sociais. Essas pesquisas buscam entender os aspectos da comunidade que estão relacionados ao compartilhamento de conhecimentos. Em geral, esses aspectos estão fortemente ligados com as interações dos usuários, uma vez que, compartilhar algo necessariamente envolve a participação de mais de uma pessoa.

Por exemplo, ADAMIC et al. (2008) realizam um estudo no Yahoo! Answers, uma comunidade online de perguntas e respostas com aproximadamente 23 milhões de questões resolvidas, objetivando entender aspectos relacionados ao compartilhamento de conhecimento. Nesse estudo feito no Yahoo! Answers um dos aspectos analisados foi o grau de reciprocidade entre usuários. O grau de reciprocidade consiste em uma medida que visa 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. Um outro aspecto interessante analisado nesse trabalho foi a profundidade do conhecimento dos usuários. Para analisar isto, postagens de usuários foram 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, conforme o trabalho de ADAMIC et al. (2008). No caso das comunidades online, diferentemente das pesquisas de dentro das organizações, o estudo quantitativo pode ser bem mais rico em detalhes e informações devido à grande quantidade de

usuários interagindo. Todavia, dentro das organizações, é possível a realização de estudos qualitativos que, por sua vez, pode-se extrair conclusões complementares ou diferentes dos estudos quantitativos.

2.1.2. Comunidades Online 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 que fazem parte das comunidades. Nessas comunidades, assim como outras similares, as discussões têm uma estrutura de trilhas (threads), ou seja, um usuário posta uma pergunta ou tópico e, logo após, outros usuários postam respostas ou comentários relativos à pergunta. Além disso, cada thread pertence a pelo menos uma categoria da comunidade (por exemplo: categoria Java, categoria banco de dados, categoria compiladores etc.) e cada usuário é avaliado por outros usuários baseado em suas perguntas ou respostas postadas.

As comunidades online para o compartilhamento de conhecimentos, como as de perguntas e respostas podem ser classificadas em duas categorias (FISCHER, 2001): comunidades de práticas e comunidade de interesses. Comunidades de práticas são aquelas compostas por pessoas que tem um conhecimento mais profundo em alguma área e estão interessados em melhorá-lo, através do aprendizado coletivo. Um exemplo de comunidades práticas são os fóruns de discussões específicos de alguns assuntos. No Brasil, o fórum GUJ [17] é um exemplo de uma comunidade de práticas. Nele, profissionais que trabalham com alguma linguagem de programação se reúnem para aprimorarem seus conhecimentos, resolver dúvidas, pedir orientações etc. Já as comunidades de interesses são aquelas cujos membros somente têm interesse sobre um determinado assunto, porém, não necessariamente possuem um conhecimento profundo. Em geral, essas comunidades não lidam com um assunto muito específico. Um exemplo desse tipo comunidade é o Yahoo! Answers, onde um usuário pode perguntar qualquer tipo de pergunta que, muito provavelmente, ele terá uma resposta (mesmo que superficial) devido à grande diversidade de assuntos presentes na comunidade.

A classificação de comunidades em práticas ou de interesses pode ser útil para nortear pesquisas e as conclusões extraídas. Por exemplo, nesta dissertação, será apresentado um processo para encontrar um usuário confiável em comunidades de perguntas e respostas. Contudo, um usuário confiável em uma comunidade de interesse, pode ser somente um usuário que é participativo e sempre faz uma breve pesquisa na Internet antes de expor algum ponto de vista em

[17] Site: http://www.guj.com.br/

alguma questão. Este usuário pode ter opiniões coerentes e valiosas, porém, superficiais. Por outro lado, um usuário confiável de uma comunidade de prática pode realmente ser um especialista em algum assunto.

2.2. Usuários Confiáveis

Comunidades online têm sido muito utilizadas para buscar e compartilhar conhecimentos (ZHANG et al., 2007). Nestas comunidades, é comum observar que alguns usuários se destacam mais quando comparados a outros, durante os momentos de interações. Esse destaque ocorre muitas vezes devido a participações ou interações consideradas importantes, ou seja, aquelas que de fato conseguem fazer com que outros usuários tirem alguma lição ou aprendizado. Esses usuários de destaque são também conhecidos como usuários confiáveis, especialistas ou tutores em potencial.

2.2.1. Reputação em Comunidades Online

Uma das formas para saber se alguém sabe alguma coisa é através de avaliações. Nas escolas e nas universidades, uma forma comum para avaliar se um aluno deve ser aprovado ou não é através das provas. Uma prova consiste em um conjunto de questões as quais cada aluno deve respondê-las. Baseado nas respostas fornecidas em uma prova, um professor as avalia, 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 são os melhores de uma turma, os medianos, ou quem precisa melhorar em algum ponto. É comum escutar nos corredores de escolas alunos dizendo frases do tipo "fulano é um CDF[18]", indicando que alguém obtém boas notas nas provas e é considerado um bom aluno. Em escolas, é comum também ver os melhores alunos atuando como monitores, ou seja, aqueles que ajudam os professores a passarem conteúdos para outros alunos, colaborando no processo de aprendizagem de uma turma.

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 pelos mesmos, pelo professor e/ou por avaliadores externos (UGULINO et al., 2009). 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.

[18] Expressão que significa cabeça de ferro. É utilizada para dizer que alguém é inteligente.

Algumas comunidades online de perguntas e respostas criaram um mecanismo similar ao de avaliações escolares, com o objetivo de descobrirem que são os seus melhores membros. Em geral, nestas comunidades, os usuários podem construir a sua reputação na rede, podendo ser positiva ou negativa. 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 especial na rede.

Segundo ZHANG et al. (2007), o prestígio de um usuário está relacionado com a sua quantidade de conhecimento exposto em uma comunidade online de perguntas e respostas. Ou seja, à medida que um usuário vai contribuindo com boas participações em uma comunidade, seu prestígio tende a aumentar. Nas relações sociais reais, e não virtuais, WASSERMAN & FAUST (1994) afirmam que pessoas que recebem elogios devido as suas capacidades tendem a ter um maior prestígio no meio em que vivem.

Em algumas comunidades online, como o Stackoverflow e a English Language and Usage[19], 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. (2013) fez um estudo no Stackoverflow e concluiu 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. Além disso, segundo BOSU et al. (2013), 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 e ser ativo fora do horário de pico da comunidade.

2.2.2. Alta Reputação, Confiabilidade e o Interesse da Academia

Reputação e participações importantes em uma comunidade são coisas que andam juntas, como argumentado na seção anterior. Essas participações importantes acabam se tornando um indicador de confiabilidade de um usuário na rede. Sob essa perspectiva, alta reputação e confiabilidade de um usuário tem uma relação direta.

Considerando o contexto dos usuários que se destacam em comunidades através de suas

[19] Site: http://english.stackexchange.com/

contribuições, pesquisadores iniciam estudos, propondo modelos e métodos, com o objetivo de encontrar de forma automática esses usuários confiáveis. Todavia, encontrar tais usuários, requer primeiramente conhecer detalhes de uma comunidade online bem como as características de cada usuário.

2.3. Trabalhos Relacionados

Estudos para encontrar os usuários confiáveis em uma comunidade já têm sido explorados na comunidade científica. Alguns destes têm foco em técnicas de recuperação de informações com processamento de linguagem natural (também conhecida como *document-based*) para identificar as competências de um usuário (STREETER & LOCHBAUM, 1988) (KRULWICH & BURKEY, 1996) (ACKERMAN & MCDONALD, 1996). Nessa abordagem, geralmente, os textos produzidos no ambiente virtual são representados através de um vetor de termos (palavras ou *tokens*) com a sua respectiva frequência. Desta forma, é possível inferir qual o tipo de competência que cada usuário tem, baseado em seus discursos. Todavia, o uso da abordagem com foco em recuperação de informação torna difícil elencar o nível de competência de cada usuário, uma vez que, é difícil julgar se um usuário fornece uma boa resposta somente fazendo um *parse* de seus textos produzidos na comunidade e, em seguida, processando-os (ZHANG et al., 2007). Segundo LITTLEPAGE & MUELLER (1997) essa abordagem tem se mostrado limitada.

BALOG et al. (2009) propuseram uma forma para identificar os usuários confiáveis baseado em consultas feitas em um ambiente e uma coleção de textos associados aos candidatos a especialistas. Este trabalho é baseado em técnicas de recuperação de informações e métodos probabilísticos e visa determinar a relevância entre uma consulta e os candidatos a usuários confiáveis. Outro trabalho baseado em recuperação de informações e métodos probabilísticos foi apresentado por LIU et al. (2012), no qual foi proposto um framework que gerava automaticamente os perfis especializados dos usuários da comunidade. Esses perfis continham informações sobre as competências dos usuários e eram construídos baseados na associação entre os tópicos da comunidade com o perfil comum do usuário.

Outra abordagem utilizada para isso é através de algoritmos de ranqueamento em grafos para encontrar os usuários confiáveis de uma rede. A ideia dessa abordagem é aplicar algoritmos na comunidade (representada através de um grafo) que atribui um número para cada usuário simbolizando seu grau de competência em algum assunto. CAMPBELL et al. (2003) e DOM et al. (2003) utilizaram o algoritmo de ranqueamento HITS em grafos para encontrar os usuários confiáveis 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: o tamanho da rede analisada. As redes eram relativamente pequenas e os resultados podiam não refletir a realidade. ZHANG et al. (2007) propuseram a construção de um algoritmo baseado em grafos para o mesmo fim, porém, aplicado em um fórum de discussão online tradicional. Apesar da abordagem de ZHANG et al. (2007) ter se mostrado interessante, os autores do trabalho concluíram, através de simulações, que comunidades com diferentes características deve ser analisadas separadamente, pois as características podem influenciar nos resultados obtidos, sendo necessário adaptações nas medidas ou nas técnicas utilizadas. ALAN et al. (2013) propuseram uma nova abordagem para identificar os usuários confiáveis, construindo um modelo híbrido da abordagem baseada em recuperação de informações com a baseada em algoritmos de ranqueamento em grafos.

BANERJEE & BASU (2008) apresentaram um algoritmo probabilístico que possibilitava direcionar perguntas para os usuários mais aptos a respondê-la. Esse algoritmo funcionava baseado em ações repetidas na rede no passado. DAVITZ et al. (2007) fez um trabalho similar, em que havia uma entidade global do sistema (agente) que monitorava a rede e decidia quais usuários receberiam (visualizariam) uma determinada questão postada através de uma análise probabilística. Todavia, essa solução baseada em agentes foi testada somente em uma comunidade pequena. SOUZA et al. (2013) propôs um algoritmo para encontrar os usuários confiáveis 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 o algoritmo proposto se mostrou eficaz para encontrar os usuários confiáveis no Twitter. Contudo, a avaliação deste algoritmo foi feita com poucos usuários.

A ideia desta dissertação é revisitar a abordagem baseada em grafos com algoritmo de ranqueamento mesclada com a abordagem recuperação de informações e com o uso de técnicas de aprendizado de máquina. A abordagem será baseada em grafos porque as comunidades serão representadas através de um grafo e baseada em recuperação de informação porque serão extraídos metadados (informações do usuário) das comunidades para realização de análises. Além disso, essa abordagem também é baseada em aprendizado de máquina porque será utilizada uma rede neural artificial e um algoritmo de agrupamento com a finalidade de encontrar os usuários de acordo com o seu grau de confiabilidade. Todavia, esta dissertação propõe uma forma diferenciada para encontrar os usuários confiáveis que será denominada de “análise por partes”. Essa análise consiste em dividir uma comunidade em vários componentes (partes) e analisá-los separadamente visando investigar lugares na rede onde os usuários confiáveis mais interagem ou participam. A partir dessa análise, é

que será possível classificar um usuário, de acordo com o seu grau de confiabilidade, através de uma rede neural artificial (RNA) ou de um algoritmo de agrupamento. A Tabela 1 mostra a relação de trabalhos relacionados onde a sigla AP se refere a análise por partes e a sigla AM se refere a aprendizado de máquina.

Tabela 1. Trabalhos Relacionados

Referência	Contexto	Técnica	Limitação	AP e AM
CAMPBELL et al. (2003)	Listas de e-mail	Algoritmo de ranqueamento.	Disponível somente para pequenas comunidades	Não
DOM et al. (2003)	Listas de e-mail	Algoritmo de ranqueamento.	Disponível somente para pequenas comunidades	Não
DAVITZ et al. (2007)	Pequenas comunidades (blogs, pequenos fóruns)	Métodos probabilísticos	Disponível somente para pequenas comunidades	Não
ZHANG et al. (2007)	Grande fórum de discussão	Algoritmo de ranqueamento	A variação das características da rede influencia os resultados	Não
BANEJEE & BASU (2008)	Grafos aleatórios e duas redes reais (California query network, Autonomous System network)	Métodos probabilísticos	Assume que cada usuário só tem uma especialidade	Não
BALOG et al. (2009)	W3C test collection (repositórios de documentos contendo: blogs, páginas pessoais, fórum etc)	Recuperação de informação e métodos probabilísticos	Assume premissas simplificadas sobre o modelo que associa os usuários aos textos	Não
LIU et al. (2012)	Comunidades: Sun forums (agora Oracle forums) e Apple Discussions	Recuperação de informação e métodos probabilísticos	Assume que o texto associado a um usuário reflete o seu conhecimento	Não
ALAN et al. (2013)	Microsoft Office Discussion Groups	Recuperação de informação e Algoritmo de ranqueamento	Assume premissas simplificadas sobre o modelo que associa os usuários aos textos	Não
SOUZA et. al. (2013)	Lista de seguidores do Twitter	Métodos de decisão multicritério	Avaliação com poucos usuários	Não
Esta dissertação	Comunidades: Biology Q& A; English Language and Usage; Physics Q&A; Mathematics Q&A; Travel Answers	Recuperação de informação, Algoritmo de ranqueamento e aprendizado de máquina	Assume que cada usuário só tem uma especialidade. Ou seja, o usuário é confiável na comunidade independente dos assuntos abordados	Sim

2.4. Comentários Finais

A ideia deste capítulo foi apresentar os conceitos básicos que foram utilizados nessa dissertação. Foi apresentado o conceito das comunidades online visando mostrar quais são seus objetivos, públicos e características gerais. Foi também mostrado resumidamente os estudos sobre o compartilhamento de conhecimentos em comunidades online.

Além disso, neste capítulo, foi dedicada uma parte para definir o que é um usuário

confiável em uma comunidade e como ele pode construir a sua reputação na rede. Por fim, foram mostrados trabalhos relacionados, apresentando suas características e diferenças e também a conexão deles com a proposta desta dissertação.

3. Métricas em Comunidades Online

Este capítulo tem como objetivo apresentar e analisar algumas métricas que podem ser extraídas de uma comunidade online de perguntas e respostas. Essas métricas serão analisadas com a finalidade de averiguar se elas podem indicar se um usuário é confiável. No decorrer do capítulo, serão apresentadas como foi realizada a extração dos dados das comunidades e a extração das métricas dos usuários. Em síntese, o objetivo deste capítulo é mostrar como foi conduzido parte do estudo empírico inicial nas comunidades online escolhidas para testar a hipótese desta dissertação. Este capítulo contém partes dos trabalhos apresentados em (PROCACI et al., 2014a), (PROCACI et al., 2014b) e (PROCACI et al., 2014c) que também foram escritos pelo autor desta dissertação.

3.1. Dataset e Características Gerais das Comunidades

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

- *Biology Q&A*[24]: uma comunidade destinada ao aprendizado de biologia;
- *English Language and Usage*[25]: uma comunidade voltada para o aprendizado da língua inglesa;
- *Physics Q&A*[26]: uma comunidade destinada ao aprendizado de física;
- *Mathematics Q&A*[27]: uma comunidade voltada para o aprendizado de matemática;
- *Travel Answers*[28]: uma comunidade destinada ao esclarecimento de dúvidas sobre viagens.

Em geral, as comunidades online de perguntas e respostas escolhidas para estudo apresentam algumas características estruturais parecidas. Em todas elas, as discussões têm uma estrutura de trilhas (*threads*) e os usuários podem responder ou comentar perguntas postadas. Além disso, nestas comunidades escolhidas existe um esquema de avaliação que permite aos usuários

[24] Site: http://biology.stackexchange.com/
[25] Site: http://english.stackexchange.com/
[26] Site: http://physics.stackexchange.com/
[27] Site: http://math.stackexchange.com/
[28] Site: http://travel.stackexchange.com/

construírem a sua reputação na rede. Ou seja, cada usuário é avaliado por outros usuários com base em suas perguntas ou respostas postadas. Esse esquema de avaliação existente foi o motivo da escolha das comunidades, pois, ele serviu como parâmetro de comparação com a proposta deste trabalho (conforme será mostrado mais adiante neste capítulo). Em síntese, dado o exposto, pode-se dizer que as cinco comunidades são parecidas, porém, frequentada por públicos diferentes.

A coleta dos dados das comunidades foi através de um *Web crawler* que consumia dados de cada comunidade. Um *crawler* é um programa de computador que navega sistematicamente em sistemas online, tipicamente com o propósito de coletar dados. O *crawler* foi desenvolvido utilizando a linguagem de programação Python e tinha a função de enviar várias requisições HTTP[29] para cada comunidade online e, em seguida, salvar os dados retornados em diversos arquivos.

A Tabela 2 mostra os dados coletados e algumas características gerais das comunidades como: número de usuários, número de mensagens, número de respostas, número de comentários, número de *threads* (que é número de postagens principais ou número de tópicos) etc.

Tabela 2. Características Gerais das Comunidades

Comunidade	**Número de mensagens**	**Número de threads**	**Número de respostas**	**Número de comentários**	**Tamanho médio de uma thread**	**Quantidade média de caracteres / postagens**	**Número de usuários**
Biology Q&A	25.828	4.549	5.734	15.545	3	314	2.317
English Language and Usage	326.915	30.044	79.978	216.893	6	236	20.408
Physics Q&A	250.337	31.678	51.838	166.821	4	328	15.753
Mathematics Q&A	1.035.275	149.948	215.346	669.981	4	222	51.245
Travel Answers	42.322	5.529	10.526	26.267	4	275	3.579

Na Tabela 2 se pode perceber que a comunidade *Mathematics Q&A* é a maior das cinco, enquanto a menor é a *Biology Q&A*. Esse fato pode ser percebido através do número de mensagens, respostas, comentários, *threads* e usuários. A quantidade média de caracteres escritos nas postagens é bem parecida nas cinco comunidades. Já considerando o tamanho médio de uma *thread*, a comunidade *English Language and Usage* apresenta o maior valor. Isso pode significar que nessa comunidade existem discussões mais longas quando comparada com as demais.

Sem entrar na discussão e classificação das comunidades em prática ou de interesse (como mostrado no capítulo 2), este trabalho visa estudar maneiras para encontrar os usuários confiáveis,

[29] Sigla de: protocolo de transferência de hipertexto.

independentemente de sua classificação.

3.2. Representação Abstrata e Métricas

Uma vez extraídos os dados das comunidades, foi necessário transformá-los em estruturas de dados (representação abstrata). Para isto, todos os dados gravados nos arquivos providos da fase anterior (extração de dados) foram lidos através de outro programa escrito na linguagem de programação Python. Todavia, esta pesquisa está somente interessada em dados relacionados com os usuários e suas participações. Logo, alguns dados desnecessários extraídos das comunidades foram descartados como, por exemplo, os relativos à paginação de cada requisição.

Em seguida, os dados lidos nesta etapa foram transformados em duas representações: modelo de classes e grafo. Através do modelo de classes e suas instâncias inicialmente foi possível extrair somente os atributos simples dos usuários (como número de perguntas ou respostas) e atributos derivados que não dependiam da representação baseada em grafo. O grafo permitiu que fossem extraídas medidas baseadas em algoritmos cuja execução depende desta representação. Uma vez extraídas as métricas do grafo, estas foram também colocadas nas instâncias do modelo de classes.

3.2.1. Representação das Comunidades Através de um Grafo

Para realizar as análises necessárias neste trabalho, foi preciso representar as comunidades através de um grafo. ZHANG et al. (2007) propõem o uso de grafo direcionado para representar o esquema de perguntas e respostas comum em comunidades online. Nessa representação, os nós do grafo representam os usuários e as arestas representam as interações entre usuários.

Tabela 3. Dados do Grafo das Comunidades

Comunidade	Número de Nós	Número de Arestas
Biology Q&A	2.317	10.316
English Language and Usage	20.408	149.993
Physics Q&A	15.753	96.938
Mathematics Q&A	51.245	418.995
Travel Answers	3.579	16.792

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 em direção ao B, simbolizando que B respondeu o A. Essa representação é mostrada na Figura 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.

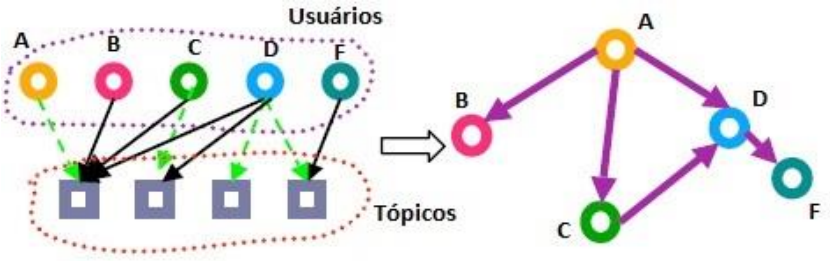

Figura 1. Exemplo de uma comunidade com seu respectivo grafo

Além do uso dessa representação proposta por ZHANG et al. (2007), este trabalho propõe uma pequena extensão desse modelo objetivando melhor representar as interações entre os usuários das cinco comunidades estudadas. Como nas comunidades analisadas é possível também comentar uma pergunta ou uma resposta, seguindo a mesma ideia do grafo proposto por ZHANG et al. (2007) mostrado na Figura 1, 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. Nessas comunidades, um comentário geralmente é usado para melhor elaborar uma pergunta ou uma resposta postada. Sobre essa perspectiva, os comentários podem ser considerados complementos às perguntas e respostas e tem como finalidade melhorar o entendimento do que foi escrito.

Desta forma, seguindo este modelo, as comunidades analisadas neste trabalho serão representadas através de um grafo que terão um determinado número de nós e número de arestas (conforme descrito na Tabela 3).

3.2.2. Distribuição de Grau

Com o objetivo de entender as interações entre os usuários e melhor caracterizar as comunidades analisadas, foi utilizada a distribuição de grau que é usualmente aplicada a grafos. Dado que as comunidades estudadas nessa dissertação serão representadas através de grafos, é perfeitamente possível elaborar as suas respectivas distribuições de grau. A distribuição de grau de

um grafo pode ser definida como uma função que descreve o número de nós que tem um determinado grau (número de vizinhos). Existem dois tipos de graus: o de entrada e o de saída. O grau entrada é o número de arestas que chegam a um nó (perguntas respondidas ou comentários fornecidos pelo usuário) e o de saída é o número de arestas que saem de um nó (perguntas postadas pelo usuário que teve pelo menos uma resposta recebida ou respostas que foram comentadas).

As Figuras 2, 3, 4, 5 e 6 mostram a distribuição de grau, tanto de entrada quando de saída, de cada comunidade. Através delas pode-se concluir que nas comunidades existem poucos usuários que são extremamente ativos e fazem várias perguntas (e recebem respostas) ou postam respostas incompletas que necessitam de complemento (comentários). Em outras palavras, existem poucos usuários que são muito ajudados (baixa frequência e alto grau de saída). Porém, a maioria dos usuários necessitam ou solicitam ajuda (alta frequência e baixo grau de saída). De forma similar, muitos usuários ajudam apenas a poucos (alta frequência e baixo grau de entrada) e, poucos, ajudam várias pessoas (baixa frequência e alto grau de entrada). Entende-se como ajuda o fornecimento de respostas ou comentários. Além disso, entende-se como frequência o número de usuários com determinado grau.

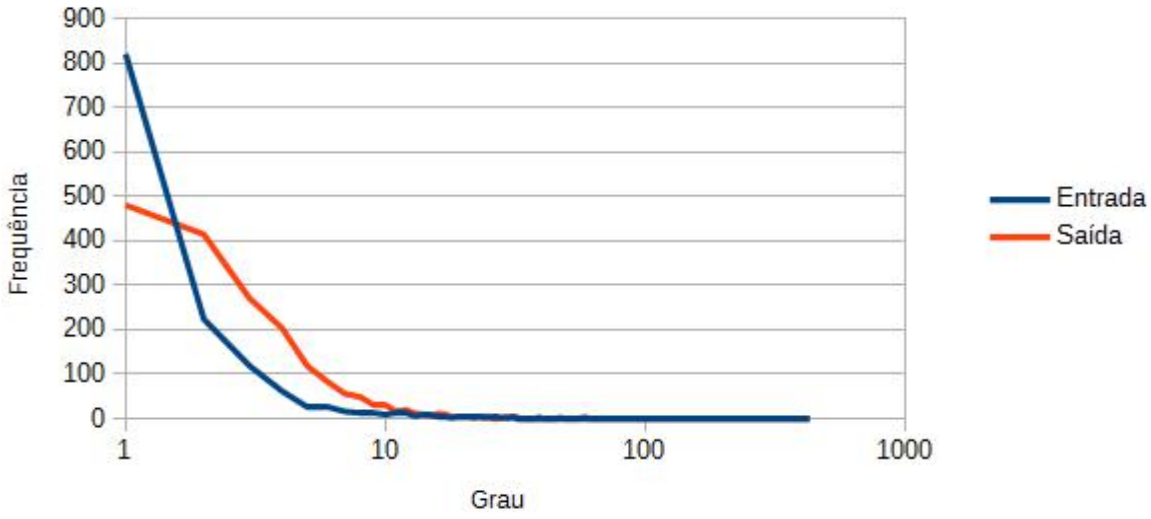

Figura 2. Distribuição de Grau – Biology Q&A

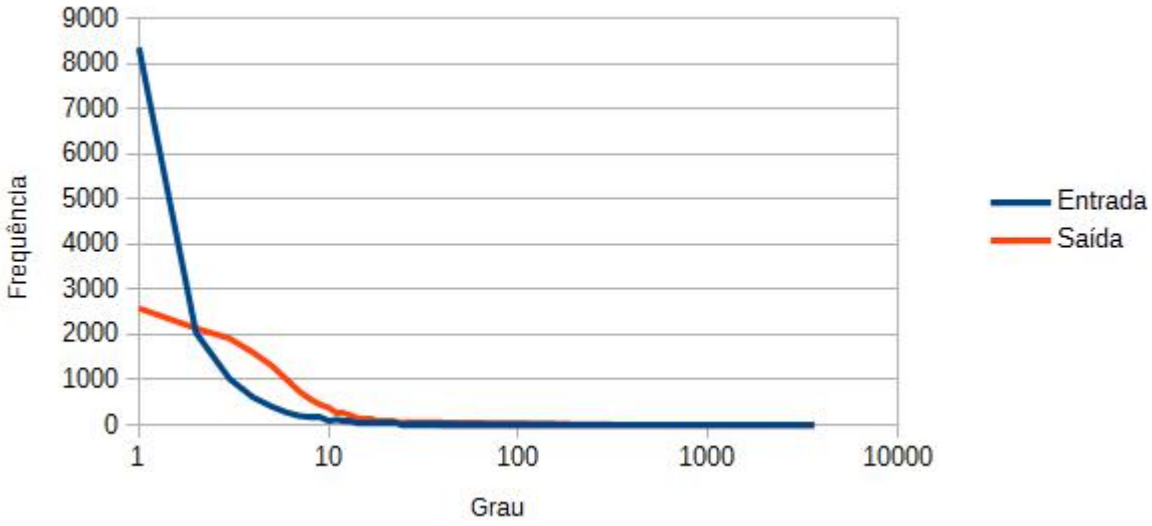

Figura 3. Distribuição de Grau – English Language and Usage

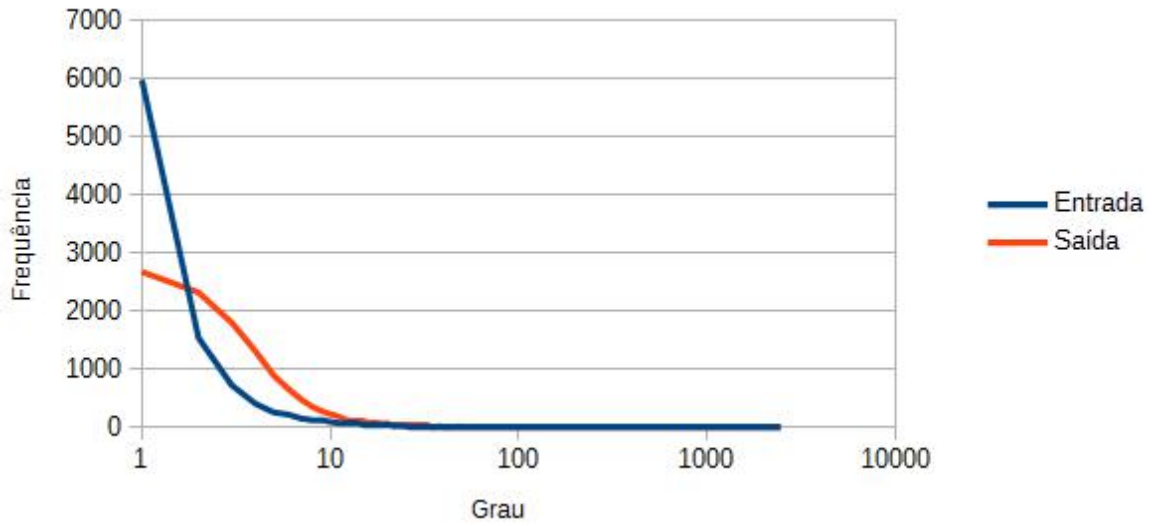

Figura 4. Distribuição de Grau – Physics Q&A

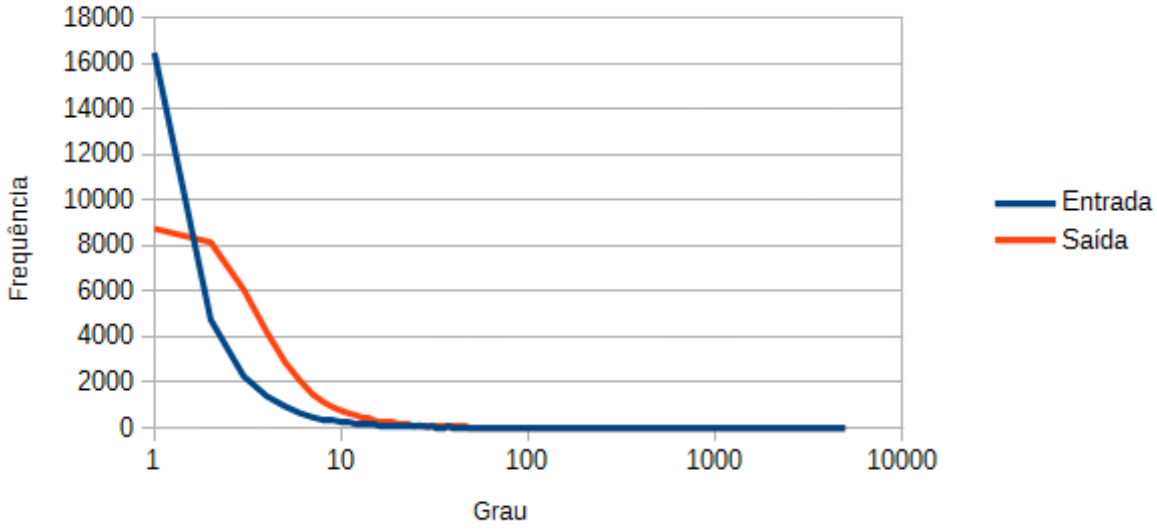

Figura 5. Distribuição de Grau – Mathematics Q&A

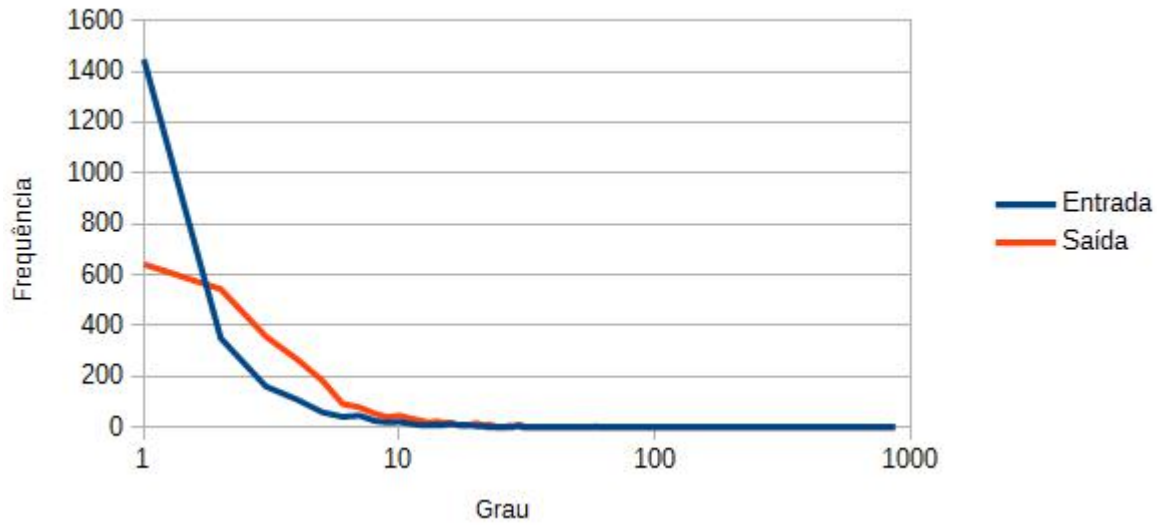

Figura 6. Distribuição de Grau – Travel Answers

3.2.3. Atributos dos Usuários - Métricas

Visando analisar métricas que podem indicar se um usuário é confiável, foram elencados alguns atributos para este fim. Além disso, foi proposta uma métrica denominada Índice de Confiança, objetivando também encontrar evidências de que um usuário é confiável. As métricas analisadas e comparadas com a métrica proposta foram:

- Entropia do usuário: a entropia é uma métrica que, no contexto desse trabalho, tem como objetivo estudar o foco de um usuário em determinados assuntos na comunidade.

O motivo da escolha desse atributo foi averiguar a relação do foco do usuário em assuntos específicos e sua reputação. A seção 3.2.5 explicará com mais detalhes a entropia do usuário.

- Número de respostas e número de comentários: Acredita-se que a reputação de um usuário em uma rede é construída através de suas boas respostas e seus bons comentários na rede. Partindo desse princípio, foi decidido analisar as relações entre o número de resposta e comentários com a reputação de um usuário.
- *z-score*: A ideia dessa medida é combinar o número de perguntas com o número de respostas de um usuário. Responder a muitas questões pode ser um indício que um usuário é confiável, mas perguntar muito pode ser um indício que esse mesmo usuário não seja confiável. A ideia dessa medida é buscar o equilíbrio entre o número de perguntas e respostas de um usuário. A seção 3.2.6 explica com mais detalhes esse atributo.
- Grau de entrada: como a comunidade será representada através de um grafo, conforme relatado na seção 3.2.1, o grau de entrada significa o número de pessoas que um usuário respondeu ou forneceu algum comentário. Acredita-se que a reputação de um usuário tenha relação com a quantidade de pessoas que ele ajuda e, por esse motivo, essa métrica foi escolhida para análise.
- *Page Rank*: Foi selecionado um algoritmo de ranqueamento visando averiguar se é uma boa escolha o seu uso para encontrar os usuários confiáveis em uma rede. Como já relatado, existem trabalhos que usam algoritmos similares ao *Page Rank* para o mesmo fim (conforme na seção 2.3), porém, se deseja comparar essa abordagem com a proposta. A representação através do grafo é necessária para o seu uso. É importante ressaltar que a variação do *Page Rank* usada nesta dissertação foi a que utiliza um grafo com arestas direcionadas sem considerar os pesos das arestas. Optou-se por essa variação do *Page Rank* pelo fato de ser mais simples daquela que considera os pesos das arestas. Além disso, estudos similares sustentam que considerar ou não os pesos podem trazer resultados similares considerando o contexto deste trabalho, como mostrado adiante na seção 3.3.3. Existem trabalhos relacionados que usam o algoritmo de ranqueamento HITS para o mesmo fim. Nesta dissertação optou-se pelo uso do *Page Rank*, pelo fato de mais recente e sucessor do HITS.
- Índice de Confiança: Essa é a medida proposta neste trabalho. A ideia desta medida é

combinar o grau de participação de um usuário, o seu foco em determinados assuntos da comunidade e também quão recentes são as participações.

É importante ressaltar que o número de perguntas não foi utilizado diretamente neste trabalho como uma possível métrica para indicar se um usuário é confiável. A razão disto é que, através dos trabalhos relacionados, se pode concluir que um usuário que somente faz perguntas tende a não ser confiável em uma comunidade (ZHANG et al, 2007). Contudo, o número de respostas é utilizado para o cálculo da métrica *z-score*.

3.2.4. Modelo de Classes

A partir dos atributos previamente escolhidos para análise, foi possível construir um modelo de classes cujas instâncias armazenam informações sobre os usuários e suas interações nas comunidades. A ideia do modelo é prover um acesso fácil aos atributos dos usuários de forma rápida e direta para as análises. Desta maneira, uma vez extraídos os dados das comunidades através do *crawler* e calculados os atributos derivados (inclusive aqueles oriundos da representação de grafos), estes foram colocados dentro das instâncias do modelo de classes. A Figura 7 mostra de forma esquemática o modelo de classes utilizado nesta dissertação.

Em síntese, o modelo contém seis classes: usuário, postagem, comentário, pergunta resposta e categoria. A classe usuário representa um membro de uma comunidade. Um usuário possui um identificador, a reputação (provida pelo esquema de avaliação das comunidades) e todos os atributos descritos na seção 3.2.3. Além disso, um usuário pode escrever zero ou várias postagens e zero ou vários comentários. A classe postagem é uma generalização da classe pergunta e da classe resposta. Em outras palavras, a classe postagem contém um conjunto de atributos que são comuns à classe pergunta e resposta. Os atributos da classe postagem são: o identificador, a data de criação, a data da última edição, a data da última atividade, o contador de avaliações positivas (contadorUpVote), o contador de avaliações negativas (contadorDownVote), o número de visualizações, a diferença entre as avaliações positivas e as avaliações negativas (*score*), o título e o texto relativo a postagem. A classe pergunta, que é subclasse da classe postagem, tem como atributo o tamanho da discussão (tamanhoThread). Além disso, uma pergunta pode ter zero ou várias respostas. Cada pergunta pertence a uma categoria e cada categoria tem como atributo um identificador e um nome. Por fim, cada postagem pode ter zero ou vários comentários associados.

Como já argumentado, a ideia desse modelo é somente prover uma maneira simples e rápida para extrair características das comunidades facilmente, bem como, facilitar as análises que

esta dissertação se propõe a fazer. Desta forma, os valores das métricas previamente calculadas no modelo de representação de grafos (como o grau de entrada de um usuário e o resultado do algoritmo *Page Rank*) ou medidas derivadas da combinação de outros atributos (como a entropia, o *z-score* e o índice de confiança) ficam armazenados nas instâncias desse modelo de classe. Sobre essa perspectiva, simplificadamente, as instâncias do modelo de classes podem ser vistas como uma espécie de memória cache para as análises posteriores.

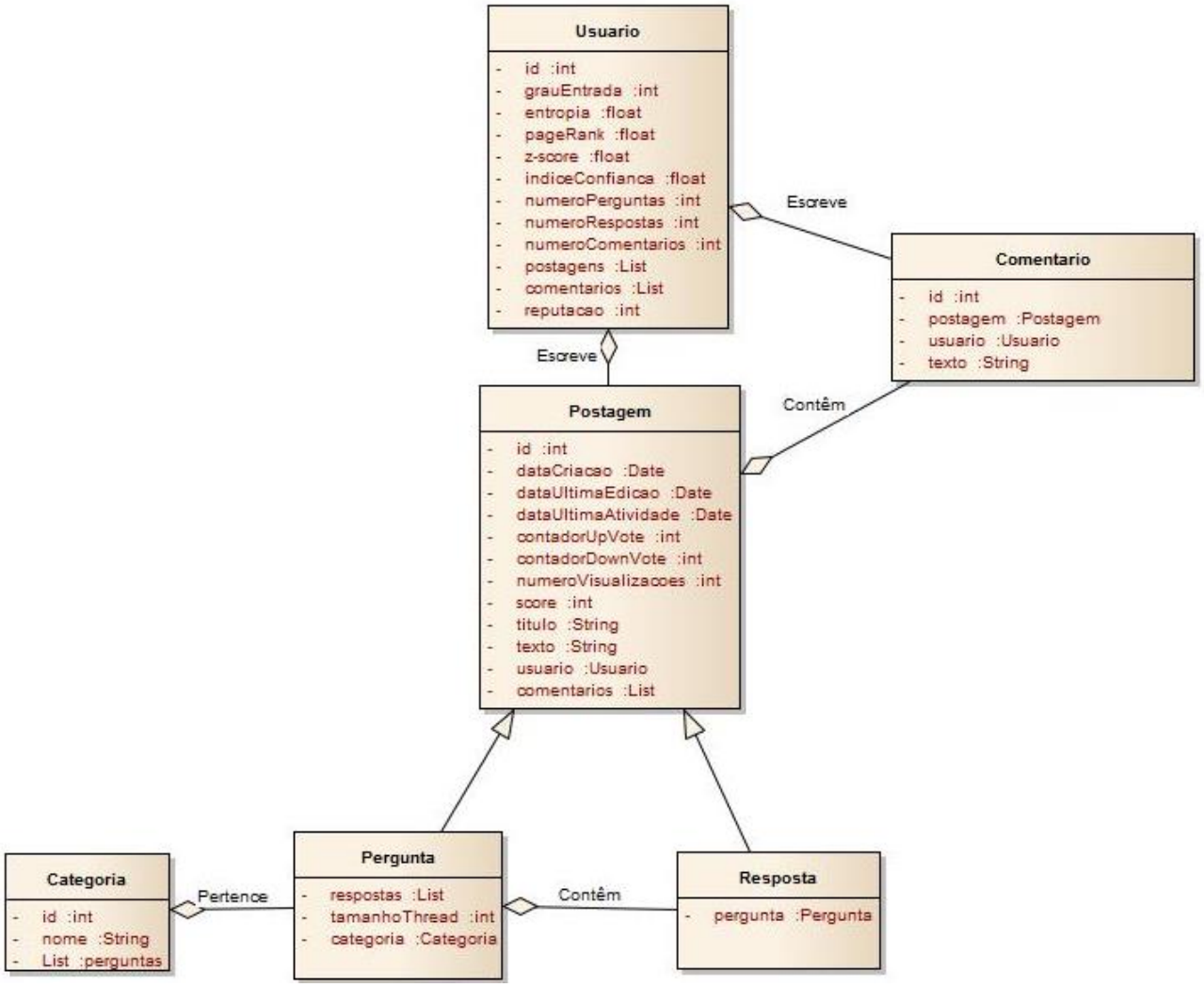

Figura 7. Modelo de Classes

3.2.5. Entropia do Usuário e sua Relação com a Reputação

Nesse trabalho, foi estudada uma medida que visa capturar o grau de concentração das respostas e dos comentários de um usuário em determinadas categorias das comunidades. A entropia é uma medida que permite capturar esse grau de concentração. Quanto mais concentradas forem as respostas ou comentários de uma pessoa em uma determinada categoria, menor é a

entropia e maior o foco. Já uma pessoa que possui alta entropia, significa que ela geralmente responde ou comenta tópicos de várias categorias, ou seja, ela tem um foco menor em assuntos específicos. ADAMIC et al. (2008) mostra em seu trabalho que a entropia de um usuário pode ser descrita através da seguinte fórmula:

$$\text{entropia} = -\sum_{i} \text{Pi} * \log_2(\text{Pi})$$

$$Condição\ de\ Existência: Pi > 0$$

Fórmula 1. Cálculo da Entropia

A variável "P" da fórmula da entropia é utilizada para determinar a capacidade de um usuário para transmitir informações. Nessa dissertação essa capacidade é participação do usuário na comunidade. Para melhor explicar a fórmula da entropia, no contexto desse trabalho, imagine que um usuário tenha postado dez respostas em uma comunidade de perguntas e respostas cujas discussões são relacionadas à matemática. Porém, três das dez respostas foram relacionadas à categoria cálculo, três relacionadas à categoria álgebra e quatro relacionadas à categoria probabilidade. Assim, para calcular a entropia desse usuário, antes, deve-se calcular o "P" de cada categoria. O valor de "P", conforme descrito na fórmula, nada mais é que um indicador de participação de um usuário em uma categoria, levando em consideração a sua participação geral na rede. Por exemplo, o "P" desse usuário na categoria cálculo é 0,3, pois 3 das 10 respostas postadas foram para a categoria cálculo (é a divisão 3/10). Desta forma, para calcular a entropia do usuário basta realizar o seguinte cálculo:

$$-((\text{Pcalc} * \log_2(\text{Pcalc})) + (\text{Palg} * \log_2(\text{Palg})) + (\text{Pprob} * \log_2(\text{Pprob})))$$

$$-((0{,}3 * \log_2(0{,}3)) + (0{,}3 * \log_2(0{,}3)) + (0{,}4 * \log_2(0{,}4))) = 1{,}57$$

Fórmula 2. Exemplo Cálculo Entropia

Uma vez entendido como se calcula a entropia e tendo a reputação dos usuários fornecida pela própria rede e sabendo que ela foi construída por avaliações realizadas por outros usuários da rede, estas foram correlacionadas estatisticamente com a entropia de cada usuário. Em outras palavras, a reputação de cada usuário foi correlacionada com a sua entropia.

Tabela 4. Coeficiente de Correlação de Pearson (entropia vs. reputação)

	Biology Q&A	English Language and Usage	Physics Q&A	Mathematics Q&A	Travel Answers
Correlação	0,59	0,36	0,33	0,36	0,44

A Tabela 4 mostra as correlações da entropia com a reputação dos usuários das comunidades. Através dela, pode-se concluir que a entropia se correlaciona moderadamente (valores entre 0,3 e 0,7) com a reputação do usuário, quando se analisa as redes. Diante disso, após as análises nas cinco comunidades, se pode concluir que um usuário com mais alta entropia (menos foco em alguns assuntos), pode ser um indicador moderado de que um usuário tem alta reputação na rede. Ou seja, um usuário que participa de várias categorias da comunidade tem maiores chances de ser confiável.

Para as correlações mostradas na Tabela 4, foram utilizados todos os usuários extraídos das comunidades Biology Q&A e Travel Answers pelo fato delas serem as menores. Amostras de comunidades menores podem não refletir o todo adequadamente, logo, foi decidido utilizar todos os usuários que foram extraídos dessas duas comunidades. Nas demais comunidades, foram utilizadas amostras aleatórias dos usuários. Na comunidade English Language and Usage foram utilizados cerca de 50% do total de usuários extraídos, na Physics Q&A foram usados por volta de 63% dos usuários e, por fim, na Mathematics Q&A foram utilizados cerca de 20% dos usuários. A ideia é que essas amostras aleatórias contemplem o máximo de tipos de usuários possíveis (os que participam muito, pouco, em várias categorias etc.). Além disso, as populações (usuários) são dinâmicas, ou seja, estão em constante mudança pelo fato das comunidades serem online. Isto resulta na impossibilidade de sempre analisar todos os elementos das populações, sendo necessário recorrer ao uso de amostras. Dado este fato, por mais que se use todos os dados extraídos de uma comunidade em um determinado momento, estes provavelmente não representam o todo, sendo portanto, sob essa visão, amostras também. Dado o que foi exposto, quanto maior for uma comunidade, menor pode ser a porcentagem de usuários utilizada na amostra, pois, a probabilidade de conter boa parte dos tipos usuário é maior.

3.2.6. Relação da Reputação com os Demais Atributos do Usuário

Com a finalidade de analisar mais atributos de um usuário (além da entropia) que possam indicar que ele tem alta reputação (usuário confiável), foram extraídas algumas medidas dos

usuários (bem como do nó do grafo que os representa) para serem correlacionadas com a reputação dos usuários. Nesta seção, foram analisadas os seguintes atributos: o número de respostas postadas, o número de comentários postados, o somatório do número de respostas com o número de comentários, o grau de entrada, o valor *z-score* e o valor atribuído pelo algoritmo *Page Rank* a cada nó (usuário) da rede. Optou-se por não ter seções separadas para cada atributo desta seção, como foi feita na seção anterior com a entropia. A relação entropia e foco nem sempre é intuitiva e, por isto, houve a necessidade de uma explicação mais detalhada.

O *z-score* é uma medida proposta por ZHANG et al. (2007) que objetiva atribuir um valor para um usuário indicando sua reputação ou *expertise* na rede. A ideia dessa medida é combinar o número de perguntas com o número de respostas de um usuário, uma vez que responder a muitas questões pode ser um indício que um usuário é confiável, porém perguntar muito pode ser um indício que esse mesmo usuário não seja confiável. No trabalho de ZHANG et al. (2007), é demostrado como foi elaborado o cálculo do z-score e chegaram na seguinte fórmula:

$$\mathrm{z-score} = \frac{(\mathrm{P}-\mathrm{R})}{\sqrt{(\mathrm{P}+\mathrm{R})}}$$

$$Condição\ de\ Existência: P + R > 0$$

Fórmula 3. Cálculo do z-score

Desta maneira, o *z-score* de cada usuário pode ser calculado considerando o número de perguntas (variável "P") e o número de respostas (variável "R") que ele postou.

O algoritmo *Page Rank*, proposto por PAGE et al. (1998), atribui um valor a todos os nós de um grafo indicando sua importância na rede. O *Page Rank* foi usado nesse trabalho para identificar os usuários mais relevantes na rede. A ideia do *Page Rank* é atribuir pontuações para cada nó de um grafo. Esse algoritmo tende a atribuir pontuações maiores para nós cujo o número de arestas que chegam a ele é maior e pontuações menores para nós cujo o número de arestas que chegam a ele é menor.

Uma vez extraídas todas as medidas das redes, estas foram correlacionadas com a reputação do usuário (indicador de competência). A Tabela 5 mostra as correlações dos atributos dos usuários de cada comunidade com a reputação adquirida na rede, onde a legenda "Num Resp" significa número de respostas, a legenda "Num Com" significa número de comentários e a legenda

“R + C” significa o somatório do número de respostas com o número de comentários.

Analisando os resultados da Tabela 5, se pode concluir que as correlações obtidas nas comunidades foram fortes (acima de 0,7). Isto significa que os atributos escolhidos podem ser fortes indícios de que um usuário é confiável. Em outras palavras, quanto maior for o valor dos atributos, maior a chance de um usuário ser confiável.

É importante ressaltar que, para as correlações mostradas na Tabela 5, foram utilizadas as mesmas amostras de usuários relatadas na seção 3.2.5.

Tabela 5. Coeficiente de Correlação de Pearson (atributos vs. reputação)

	Num Resp	**Num Com**	**R + C**	**z-score**	**Grau Entrada**	**Page Rank**
Biology Q&A	0,92	0,84	0,89	0,79	0,91	0,90
English Language and Usage	0,92	0,76	0,82	0,81	0,88	0,86
Physics Q&A	0,91	0,73	0,81	0,70	0,82	0,81
Mathematics Q&A	0,89	0,88	0,90	0,81	0,90	0,89
Travel Answers	0,97	0,83	0,91	0,76	0,93	0,91

3.2.7. Índice de Confiança

Dado o cenário das análises realizadas, este trabalho buscou uma medida que combinasse diferentes fatores presentes nas comunidades de forma a construir um novo indicador que pudesse também representar um usuário confiável em uma rede. Esse novo indicador foi denominado de Índice de Confiança e ele considera o grau de participação de um usuário na rede, o seu foco (entropia) e há quanto tempo o usuário participa da comunidade.

O grau de participação mede a interação do usuário na rede. A participação pode ser definida, por exemplo, através de medidas com o número de comentários, de respostas ou o grau de entrada. Neste trabalho foi escolhido o número de respostas como o indicador de participação pelo fato dessa medida ter obtido as melhores correlações, com exceção da comunidade Mathematics Q&A (conforme mostrado na seção 3.2.6). Contudo, no geral, pode-se considerar que o número de respostas obteve os melhores resultados.

Para medir o foco do usuário em determinados assuntos, foi utilizada a entropia. Desta forma, como o número de respostas e a entropia se correlacionam positivamente com a reputação

provida pela rede, julgou-se factível neste trabalho realizar o produto dessas duas medidas. Desta maneira, pode-se obter o equilíbrio entre um usuário que participa muito em um somente assunto e um usuário que participa muito em vários assuntos. A ideia é que seja privilegiado aquele que participa muito (alto número de respostas) em vários assuntos (alta entropia e menos foco) em relação àquele que participa muito em poucos assuntos (baixa entropia e mais foco).

Além disso, durante a elaboração do índice de confiança, buscou-se considerar o tempo de vida do usuário na rede. A ideia disso é privilegiar um usuário que participa muito em menos tempo a um usuário que participa muito, porém, em um intervalo de tempo maior. Para isso, o produto da participação pela entropia foi dividido pelo intervalo de tempo que descreve o tempo de vida do usuário na rede. A fórmula do Índice de Confiança é descrita abaixo:

$$\text{índice de confiança } = \frac{participação * entropia}{tempo}$$

$$= \frac{\text{num resp} * \text{entropia}}{data\ atual - data\ primeira\ participação}$$

$$Condição\ Existência: tempo > 0$$

Fórmula 4. Cálculo do Índice de Confiança

Uma vez calculado o Índice de Confiança para os usuários da rede, estes foram correlacionados estatisticamente com a reputação (indicador de competência) provida pela rede. Através da Tabela 6, percebe-se que o Índice de Confiança pode ser um indicador útil em comunidades online. Todavia, quando se compara as correlações do número de respostas com as do Índice de Confiança pode-se perceber que não há melhorias substanciais nas correlações. Isto significa que mesmo sendo o Índice de Confiança uma métrica mais complexa, quando se deseja um indicador de confiabilidade de um usuário, pode ser mais interessante utilizar uma métrica mais simples como o número de respostas.

De qualquer forma, pode-se concluir que o Índice de Confiança pode ser um bom indicador de usuários confiáveis em uma comunidade. Em outras palavras, quanto maior o Índice de Confiança, mais confiável pode ser um usuário para responder perguntas.

Tabela 6. Coeficiente de Correlação de Pearson (ind. confiança vs. reputação)

	Biology Q&A	English Language and Usage	Physics Q&A	Mathematics Q&A	Travel Answers
Correlação	0,90	0,92	0,89	0,88	0,96

Para as correlações mostradas na Tabela 6, foram utilizadas as mesmas amostras de usuários conforme relatado na seção 3.2.5.

3.3. Comparando Resultados com Métricas de um Grupo do Facebook

Com o objetivo de buscar mais evidências e averiguar se os atributos analisados na seção 3.2 deste trabalho podem ser, de fato, indícios da confiabilidade de um usuário, foi realizado um estudo comparativo com estudo similar feito em grupo do Facebook. Esse estudo feito no grupo do Facebook foi apresentado em (PROCACI et al., 2014a). Neste estudo se buscou obter métricas que descrevam um usuário confiável, considerando o contexto do Facebook.

O grupo do Facebook desse estudo era uma comunidade destinada a debates relacionados à tecnologia Java. De forma similar feita nesta dissertação, as interações do grupo foram representadas através de um grafo, considerando os pesos das arestas tornando possível contabilizar o número de interações entre os usuários. Também, de maneira similar a esse trabalho, foram extraídas algumas métricas dos usuários objetivando identificar aquelas que podem ser um indício que um usuário é confiável.

3.3.1. O Grupo Java do Facebook

O grupo Java do Facebook que foi analisado no trabalho referenciado é um local destinado para debates, discussões sobre a tecnologia Java. Geralmente, nessa comunidade, as pessoas entram e fazem alguma pergunta sobre tópicos de programação Java. Além disso, a comunidade conta com diversos tipos de usuários, variando desde iniciantes até profissionais com vasta experiência na tecnologia. Devido ao grande número de membros, um usuário pode ter uma pergunta respondida rapidamente. Nesse estudo, foram utilizados 7.598 usuários da comunidade e 97.867 mensagens postadas. Desse total de mensagens, 15.357 são postagens principais, ou seja, a primeira postagem de uma discussão que geralmente contêm uma pergunta.

3.3.2. Distribuição de Grau do Grupo Java do Facebook

Nesse estudo realizado no Grupo do Facebook, também foi elaborada a sua distribuição de grau com o objetivo de caracterizar melhor a comunidade. A Figura 8 mostra a distribuição de grau.

De acordo com a Figura 8, apesar do Facebook e as comunidades estudadas nesta dissertação serem diferentes, as distribuições de grau são similares. Desta forma, a conclusão tirada foi similar a distribuição de grau das comunidades estudadas nesta dissertação: em vez de todos se ajudarem igualmente no grupo do Facebook, existem poucos usuários que são extremamente ativos e fazem várias perguntas (e recebem respostas), porém, a maioria dos usuários faz poucas perguntas. De forma semelhante, muitos usuários respondem apenas a poucas questões e, poucos, respondem a várias.

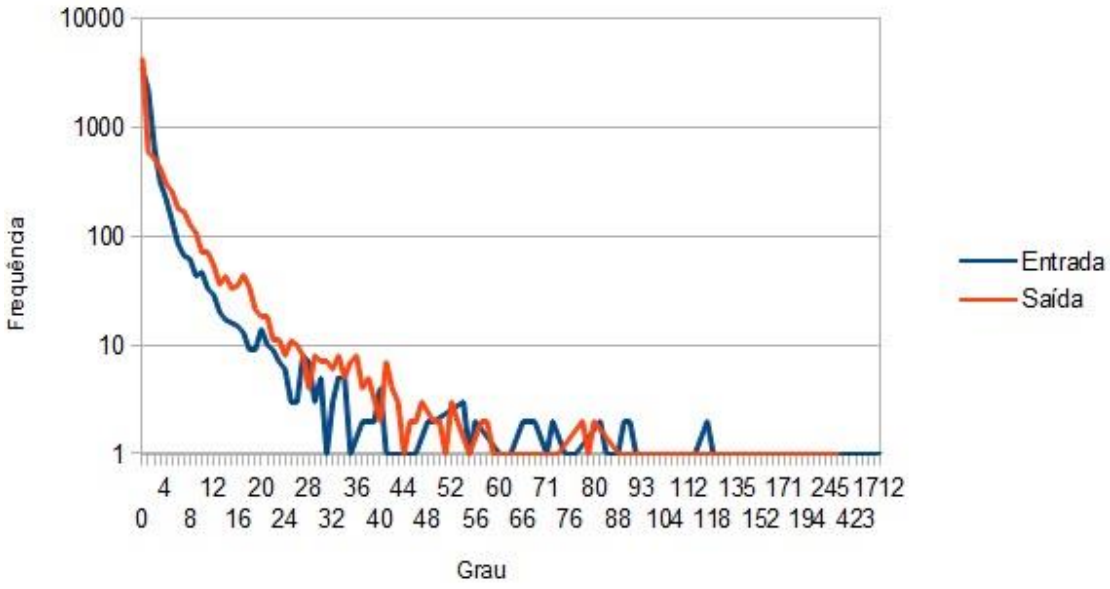

Figura 8. Distribuição de Grau – Grupo Facebook

3.3.3. Experimentos Realizados no Grupo Java do Facebook

Nesta seção será descrito como foram realizados os experimentos para encontrar os indícios dos usuários confiáveis na comunidade do Facebook. Inicialmente, foram escolhidas algumas medidas para serem extraídas da comunidade ou do grafo que a representa. As medidas escolhidas foram: o grau de entrada de cada nó, o número de respostas que cada usuário postou, o número de "*Likes*" [31] que cada usuário recebeu e, por fim, resultados atribuídos a cada nó pelas

[31] Recurso que permite um usuário expressar sua aprovação por uma mensagem postada por ele mesmo ou por outro usuário do Facebook.

quatro variações analisadas do algoritmo *Page Rank*. As variações do *Page Rank* foram: *Page Rank* em um grafo direcionado não utilizando os pesos das arestas (PRDSP), *Page Rank* em um grafo direcionado utilizando os pesos das arestas (PRDCP), *Page Rank* em um grafo não direcionado não utilizando os pesos das arestas (PRNDSP), *Page Rank* em um grafo não direcionado utilizando os pesos das arestas (PRNDCP). É importante ressaltar que o número de *"Likes"* é um recurso exclusivo do Facebook e é diferente do esquema de avaliações das cinco comunidades analisadas anteriormente.

Além disso, foram escolhidos aleatoriamente 50 usuários da comunidade que tiveram suas postagens na comunidade avaliadas por um profissional Sênior na tecnologia Java. Esse profissional não fazia parte da equipe da pesquisa e era integrante de um time de arquitetos Java em uma empresa estatal brasileira. Esse profissional ficou responsável por atribuir uma nota de 1 a 5 a cada usuário escolhido, de acordo com a Tabela 7, objetivando classificar o nível de competência de cada um na tecnologia Java. O objetivo dessa classificação feita pelo profissional foi servir de base para a comparação com as métricas extraídas da comunidade do Facebook.

Tabela 7. Níveis de Competência em Java

Nota	**Classificação**	**Descrição**
1	Iniciante	Iniciando no mundo Java.
2	Aprendiz Java	Sabe conceitos básicos e consegue programar algo.
3	Usuário Java	Sabe conceitos avançados e programa bem.
4	Profissional	Pode responder quase todas as perguntas sobre Java e seus conceitos.
5	Sênior	Domina a tecnologia Java e sabe profundamente conceitos avançados.

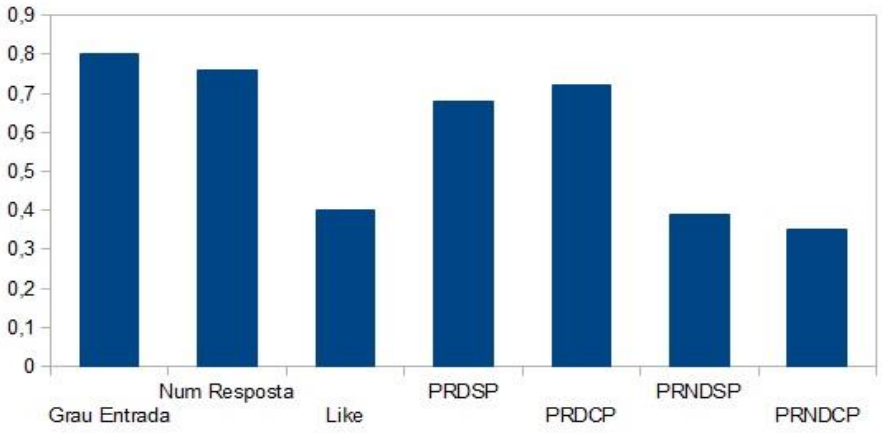

Figura 9. Correlações com as Análises Humana

As medidas extraídas da comunidade do Facebook foram correlacionadas estatisticamente com a análise humana do profissional objetivando descobrir quais eram as que mais se aproximam da classificação humana. O gráfico com as correlações é mostrado na Figura 9.

Após as correlações, verificou-se que o grau de entrada, o número de respostas, o *Page Rank* direcionado utilizando o peso (PRDCP) e o *Page Rank* direcionado não utilizando o peso (PRDSP) obtiveram as melhores correlações (por volta de 0,8). Um fato interessante é que o uso do peso no *Page Rank* direcionado não influenciou substancialmente as correlações. Contudo, as demais variações do *Page Rank* e o número de *Likes* obtiveram correlações piores, mostrando-se que não são as melhores métricas para identificarem os usuários confiáveis na comunidade. A conclusão que se pode tirar desse trabalho é que medidas simples como o grau de entrada e o número de respostas são as melhores alternativas quando se deseja identificar os usuários confiáveis nas comunidades online.

3.3.4. Comparando Resultados

Analisando os resultados obtidos nas cinco comunidades investigadas nessa dissertação e o trabalho realizado no grupo do Facebook, se pode concluir que os resultados foram similares quando comparadas as medidas que foram comuns a ambos. Por exemplo, nas análises das cinco comunidades e do grupo do Facebook, o número de respostas, o grau de entrada e o *Page Rank* em um grafo direcionado não utilizando os pesos das arestas foram concluídos como bons indícios de um usuário confiável. Em outras palavras, quanto maior for o número de respostas, o grau de entrada e o valor do *Page Rank* de um usuário, maiores são as chances de ser confiável.

Todavia, o trabalho realizado no grupo do Facebook tem um ponto que pode colocar em

dúvida a sua credibilidade: a sua avaliação. A avaliação deste trabalho foi realizada com um grupo de apenas 50 usuários que, por sua vez, pode não ser suficiente para a extração de conclusões sobre toda a comunidade. Contudo, os resultados das métricas comuns a aquelas analisadas nessa dissertação convergiram. Desta forma, os dois trabalhos quando analisados em conjunto só reforçam as evidências que as métricas analisadas em ambos podem ser indícios de confiabilidade.

3.4. Comentários Finais

Neste capítulo foi apresentado um estudo em cinco comunidades distintas visando explorar métricas que possam indicar que um usuário é confiável. As cinco comunidades analisadas foram inicialmente descritas de acordo com suas características como número de usuários, mensagens etc. Além disso, foi mostrado como foram extraídos os dados das comunidades e como estes foram representados através de um modelo de classes e de grafo.

Foi estudada a relação entre a entropia de um usuário (foco em assuntos específicos) com a sua reputação na rede. Foi concluído que a entropia se correlaciona moderadamente com a reputação do usuário. Isto significa que, um usuário que não foca sua participação na rede (alta entropia) em categorias específicas, pode ser um indicador moderado que ele tem alta reputação. Foram analisados e correlacionados também vários atributos dos usuários com suas respectivas reputações. Além disso, foi proposta uma métrica denominada Índice de Confiança com o objetivo de verificar se essa medida pode ser útil para encontrar os usuários confiáveis de uma rede. A métrica proposta se mostrou uma boa alternativa para este fim.

Por fim, os resultados obtidos nas cinco comunidades foram comparados aos resultados obtidos de um estudo similar realizado em um grupo de discussões Java do Facebook. Algumas métricas estudadas neste capítulo foram também estudadas no trabalho feito no grupo do Facebook. A partir dessa interseção, os resultados obtidos nos dois trabalhos foram comparados e foi constatado que eles convergem para uma mesma conclusão.

Em síntese, através do exposto neste capítulo, pode-se concluir que as métricas apresentadas e analisadas podem ser um indicador moderado ou forte que um usuário é confiável. Em outras palavras, quanto maior for o valor das métricas, maiores são as chances de um usuário ser confiável.

4. Particionando Comunidades

Este capítulo tem como objetivo analisar métricas, como aquelas apresentadas no capítulo 3, porém, considerando somente partes das comunidades escolhidas para análise nessa dissertação. A ideia é dividir a comunidade em partes menores e depois verificar em quais lugares serão obtidas as melhores correlações dos atributos dos usuários com o indicador de competência (reputação). Desta forma, espera-se encontrar partes das comunidades onde os atributos indiquem os usuários confiáveis. Este capítulo contém partes do trabalho apresentado em (PROCACI et al., 2014b), que foi também escrito pelo autor desta dissertação.

4.1. Estrutura Bow Tie

Nessa dissertação, foi estudada uma forma de dividir as comunidades em partes considerando alguns padrões de interação. Existem vários perfis de usuários variando desde usuários que entram em comunidades e somente fazem perguntas até aquele que somente respondem questões na comunidade. Utilizando a estrutura Bow Tie, conforme descrita por BRODER et al. (2000), foi examinada a estrutura geral das comunidades. Em síntese, a estrutura Bow Tie objetiva descrever a organização de uma rede baseada em tipos de interação. Basicamente, a ideia principal dessa estrutura é classificar uma rede em seis componentes distintos: *Core*, *IN*, *OUT*, *Tendrils*, *Tubes* e *Disconnected*. Contudo, para que essa classificação seja possível, uma rede deve ser representada através de um grafo.

No contexto desta dissertação, o componente *Core* das comunidades analisadas contêm os usuários que frequentemente se ajudam mutuamente. De acordo com a representação de grafos proposta para a comunidade, o *Core* pode ser identificado através de um algoritmo que identifica os componentes fortemente conexos de um grafo. BRODER et al. (2000) afirma que o *Core* é composto pelos nós do maior componente fortemente conexo de um grafo. Contudo, neste trabalho, decidiu-se por considerar como parte do *Core* todos os nós de todos os componentes fortemente do grafo, pois, julgou-se factível agrupar pessoas com comportamentos similares, mesmo que, todas elas não estejam conectadas entre si. No contexto das cinco comunidades estudadas, se deseja saber quem oferece ajuda mútua através do *Core*.

O componente *IN* é aquele que contêm usuários que somente fazem perguntas e obtêm respostas ou comentários de algum membro do *Core* (ou seja, um nó que têm grau de entrada igual

a zero, grau de saída maior que zero e tem uma aresta que sai de si e chega em algum membro do *Core*). Em outras palavras, os nós que compõem o *IN* são aqueles que conseguem alcançar algum membro do *Core*, mas não podem ser alcançados por nenhum membro do *Core*.

O componente *OUT* é aquele que contêm os usuários que respondem ou comentam mensagens postadas por algum membro do componente *Core* (um membro do *OUT* é o nó cujo grau de saída é igual a zero, o grau de entrada é maior que zero e existe uma aresta que sai de um membro do *Core* e incide em si). Ou seja, os membros do *OUT* podem ser alcançados pelos membros do *Core*, porém, não podem alcançar o *Core*.

Os componentes *Tendrils* e *Tubes* se conectam no componente *IN* ou no componente *OUT*, porém, não se conectam no *Core*. Os *Tendrils* são aqueles que podem ser alcançados por algum membro do *IN* e não alcançam o *Core* **ou** aqueles que podem alcançar algum membro do *OUT* e não podem ser alcançados por alguém do *Core*. Já os *Tubes*, são aqueles que podem ser alcançados por algum membro do *IN* e não alcançam o *Core* **e** aqueles que podem alcançar algum membro do *OUT* e não podem ser alcançados por alguém do *Core*. Os *Disconnected* são aqueles que não se enquadram em nenhum dos componentes anteriores.

A estrutura Bow Tie foi proposta por BRODER et al. (2000) com a finalidade de entender a organização da Web. BRODER et al. (2000) utilizaram dois *Web crawlers* e cada um percorreu cerca de 200 milhões de páginas e 1,5 bilhão de *links*. A partir disso, construíram o grafo da Web onde cada página era um nó do grafo e os links entre as páginas eram as arestas. Uma vez tendo o grafo da Web, BRODER et al. (2000) a classificou de acordo com a estrutura Bow Tie (Figura 10) e obtiveram resultados interessantes como: páginas que usualmente se conectam umas às outras, por exemplo, uma página A que tem um *link* para uma página B e vice-versa (membros do *Core*), ou mesmo páginas que somente têm *links* para outras e nunca são referenciadas por nenhuma outra. ZHANG et al. (2007) analisaram mensagens de um fórum de discussão tradicional e as representou através de um grafo, que era composto por 13.789 nós e 55.761 arestas e, depois disso, o classificaram de acordo com a estrutura Bow Tie.

Como as comunidades analisadas neste trabalho foram representadas através de um grafo, foi possível extrair quais nós (usuários) pertencem a cada componente da estrutura Bow Tie. A Tabela 8 mostra dados da estrutura Bow Tie das comunidades analisadas, dados da estrutura da Web, relatada em (BRODER et al., 2000) e também da estrutura de um fórum tradicional, descrita em (ZHANG et al., 2007).

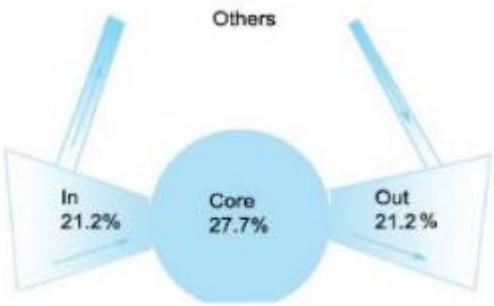

Figura 10. Estrutura Bow Tie da Web (BRODER et al., 2000)

Através da Tabela 8, percebe-se que o componente *Core* das cinco comunidades analisadas são bem grandes. Logo, se pode inferir que as comunidades analisadas são lugares onde grande parte das pessoas desejam ajudar e serem ajudadas, como a *Mathematics Q&A*, por exemplo, em que 50,54% estão dispostos a ajudar e ser ajudada. Nesta mesma comunidade, 33,93% dos usuários pertencem ao componente *IN*, significando que uma parcela menor dos usuários quando comparado com o *Core*, porém considerável, fazem somente perguntas na rede. Por outro lado, 8,48% fazem parte do *OUT* e somente respondem ou comentam os tópicos. A comunidade *English Language and Usage* tem 48,03% de seus usuários no *Core* que, por sua vez, é bem semelhante ao *Core* da comunidade *Biology Q&A* (47,48%).

Tabela 8. Dados da Estrutura Bow Tie

	Core	**IN**	**OUT**	**Tendrils**	**Tubes**	**Disconnected**
Web	27,7%	21,2%	21,2%	21,5%	0,4%	8,0%
Fórum Tradicional	12,3%	54,9%	13,0%	17,5%	0,4%	1,9%
Biology Q&A	47,48%	32,59 %	11,52%	2,67%	0,0%	5,74 %
English Language and Usage	48,03%	25,91%	18,89%	3,10%	0,015%	4,05%
Physics Q&A	49,12%	29,81%	13,01%	3,08%	0,006%	4,97%
Mathematics Q&A	50,54%	33,93%	8,48%	1,74%	0,005%	5,30%
Travel Answers	41,32%	28,95%	22,60%	3,10%	0,03%	4,00%

O componente *IN* da comunidade *Biology Q&A* (32,59 %) e da comunidade *Mathematics Q&A* (33,93%) são relativamente maiores quando comparados o *IN* das comunidades *English Language and Usage* (25,91%), *Physics Q&A* (29,81%) e *Travel Answers* (28,95%). Isso significa que, nessas duas comunidades (*Biology Q&A* e *Mathematics Q&A*), há mais pessoas interessadas

em apenas fazer perguntas e não responder (ou ajudar) alguém quando comparado com as demais (*English Language and Usage*, *Physics Q&A* e *Travel Answers*). O componente *OUT* é semelhante nas comunidades *Physics Q&A* (13,01%) e *Biology Q&A* (11,52%).

4.2. Análise das Métricas Considerando a Estrutura Bow Tie

Com o objetivo de analisar as métricas dos usuários e sua relação com o indicador de competência em partes das comunidades, foram realizadas correlações dessas métricas com a reputação do usuário considerando os componentes da estrutura Bow Tie. Esta estrutura é muito importante no contexto desta dissertação, pois, através de seus componentes, será possível investigar se existem partes menores das comunidades onde os usuários confiáveis podem ser encontrados com mais facilidade. É esperado que a análise de partes das comunidades possa trazer resultados mais relevantes quanto a análise realizada com amostras aleatórias das comunidades durante buscas por usuários confiáveis.

Seguindo um esquema similar ao apresentado no capítulo 3 para avaliação das métricas, correlações foram feitas dos atributos dos usuários com o indicador de competência escolhido (reputação), porém, agora considerando os componentes da estrutura Bow Tie. As Tabelas 9, 10, 11, 12 e 13 mostram as correlações dos atributos dos usuários com o indicador de competência, ora considerando as amostras das comunidades (Geral – apresentado no capítulo 3) ora considerando os componentes da estrutura Bow Tie.

Tabela 9. Coeficiente de Correlação de Pearson – Biology Q&A

Atributo/Componente	Geral	Core	IN	OUT	Tendrils	Tubes
Num Resp	0,92	0,93	0,05	-0,14	0,03	-
Num Com	0,84	0,83	0,11	0,43	0,47	-
R + C	0,89	0,90	0,12	0,34	0,40	-
z-score	0,79	0,83	0,008	0,36	0,26	-
Grau de Entrada	0,91	0,92	-	0,32	0,15	-
Entropia	0,59	0,58	0,11	0,3	0,25	-
Índice de Confiança	0,90	0,91	0,03	0,13	0,30	
Page Rank	0,90	0,91	-	0,29	0,14	-

Tabela 10. Coeficiente de Correlação de Pearson – English Language and Usage

Atributo/Componente	Geral	Core	IN	OUT	Tendrils	Tubes
Num Resp	0,92	0,93	0,07	-0,06	-0,10	-
Num Com	0,76	0,76	0,16	0,39	0,28	-
R + C	0,82	0,83	0,17	0,35	0,19	-
z-score	0,81	0,81	0,05	0,40	0,14	-
Grau de Entrada	0,88	0,86	-	0,36	0,20	-
Entropia	0,36	0,34	0,12	0,4	0,2	-
Índice de Confiança	0,92	0,92	0,07	0,28	0,12	
Page Rank	0,86	0,84	-	0,32	0,21	-

Tabela 11. Coeficiente de Correlação de Pearson – Physics Q&A

Atributo/Componente	Geral	Core	IN	OUT	Tendrils	Tubes
Num Resp	0,91	0,90	0,07	-0,29	-0,20	-
Num Com	0,73	0,74	0,10	0,38	0,26	-
R + C	0,81	0,80	0,11	0,28	0,17	-
z-score	0,70	0,69	0,02	0,36	0,15	-
Grau de Entrada	0,82	0,81	-	0,25	0,16	-
Entropia	0,33	0,32	0,09	0,37	0,22	-
Índice de Confiança	0,89	0,90	0,08	0,19	0,18	
Page Rank	0,81	0,80	-	0,31	0,13	-

Tabela 12. Coeficiente de Correlação de Pearson – Mathematics Q&A

Atributo/Componente	Geral	Core	IN	OUT	Tendrils	Tubes
Num Resp	0,89	0,87	0,07	-0,02	0,02	-
Num Com	0,88	0,85	0,10	0,38	0,18	-
R + C	0,90	0,87	0,11	0,37	0,17	-
z-score	0,81	0,78	0,02	0,41	0,13	-
Grau de Entrada	0,90	0,86	-	0,34	0,16	-
Entropia	0,36	0,33	0,11	0,41	0,33	-
Índice de Confiança	0,88	0,86	0,04	0,24	0,27	
Page Rank	0,89	0,87	-	0,32	0,12	-

Tabela 13. Coeficiente de Correlação de Pearson – Travel Answers

Atributo/Componente	Geral	Core	IN	OUT	Tendrils	Tubes
Num Resp	0,94	0,97	0,21	0,36	0,14	-
Num Com	0,83	0,85	0,28	0,18	0,02	-
R + C	0,91	0,89	0,31	0,38	0,16	-
z-score	0,76	0,81	0,12	0,41	0,09	-
Grau de Entrada	0,93	0,92	-	0,39	0,09	-
Entropia	0,44	0,45	0,25	0,43	0,16	-
Índice de Confiança	0,96	0,97	0,15	0,28	0,13	
Page Rank	0,91	0,90	-	0,28	-0,05	-

Através das correlações, se percebe que o componente *Core* é o que obteve resultados mais próximos das correlações gerais (considerando amostras das comunidades). Analisando a entropia, as melhores correlações entre as partes das comunidades estudadas foram obtidas pelo componente *Core* e *OUT*. Todavia, as correlações relativas às entropias são moderadas.

Analisando os demais atributos, a maioria deles apresentou uma correlação forte (valores acima de 0,7) no componente *Core*. Além disso, os valores obtidos no componente *Core* são bem similares quando comparados com os resultados que considera amostras gerais das comunidades. Os demais componentes (*IN*, *OUT*, *Tendrils* e *Tubes*) obtiveram correlações piores quando comparados com o *Core*. Uma das possíveis razões para o componente *IN* não apresentar boas correlações pode estar relacionada ao seguinte fato: os usuários que compõem este componente só fazem perguntas. Logo, isto pode ser um fator que impossibilite um usuário a construir sua reputação na rede, indicando ausência de *expertise*. Talvez, por essa razão, trabalhos similares a este, raramente consideram o número de perguntas como indicador de competência (no máximo utilizam o número de respostas para compor uma nova métrica, como o caso do *z-score*).

Comparando as correlações entre o número de respostas e o grau de entrada no componente *Core* das comunidades, se pode perceber que o número de pessoas que um usuário responde (grau de entrada) traz piores correlações quando comparada com o número de resposta que um usuário fornece. Isto pode indicar que o número de vezes que um usuário interage é mais importante que o número de pessoas com quem ele interage neste tipo de problema.

4.3. Outras Métricas do Usuário

Geralmente, quando comunidades online são estudadas e representadas através de grafos, é comum ver outras métricas clássicas de análises de redes sociais sendo utilizadas, tais como:

betweenness, *closeness* e *eigenvector*. Diante disto, buscou-se neste trabalho verificar se essas métricas podem ser evidências que um usuário é confiável. Todavia, os algoritmos utilizados para calcular essas métricas requerem um tempo de processamento muito longo para trazerem resultados. Isso se deve à alta complexidade dos algoritmos envolvidos. Devido a isso, a utilização de tais métricas pode ser um gargalo em análises envolvendo grafos com muitos nós e muitas arestas.

A métrica *betweenness* é um valor atribuído a cada nó de um grafo. Ela representa o número de caminhos mínimos de todos os vértices para quaisquer outros vértices que passam por um determinado nó (ABASSI et al., 2012). Existe também a métrica *betweenness* para arestas (CUZZOCREA et al., 2012), porém, esta não será discutida neste trabalho. Esta dissertação está interessada em investigar métricas que descrevem características próprias de cada usuário. Desta forma, utilizar a métrica *betweenness* para arestas com a finalidade de averiguar se um usuário é confiável, aparentemente, não é muito coerente, pois, uma aresta envolve necessariamente dois nós (usuários). Provavelmente, existem formas para utilizar a métrica *betweenness* para arestas para encontrar os usuários confiáveis, contudo, esta pesquisa se limita a aplicação a métrica *betweenness* em nós.

Em grafos conexos, existe uma distância entre todos os pares de nós, definido pelo comprimento de seus caminhos mínimo. *Farness* é uma métrica de um nó que é definida como a soma das distâncias de si para todos os outros nós. O inverso da métrica *farness* é definida como *closeness* (DANGALCHEV, 2006). Quanto mais central for um nó de um grafo, menor é a sua distância para todos os demais nós. A métrica *closeness* é muito útil quando se quer estimar o tempo requerido para espalhar informações de um nó para todos os demais.

A complexidade de tempo para calcular a métrica *betweenness* e *closeness* para todos os nós de um grafo é $O(n^3)$ (ZHUNGE & ZHANG, 2010) (BADER & MADDURI, 2006). Existem estratégias para melhorar (diminuir) a complexidade para o cálculo dessas métricas, como apresentada no trabalho de (BRANDES, 2001), porém, analisar e avaliar essas estratégias está fora do escopo dessa dissertação.

A métrica *eigenvector* mede a influência de um nó em uma rede (BONACICH & LLOYD, 2001). O cálculo dessa métrica envolve atribuir pontuações a todos os nós da rede. A ideia é que os nós que se conectem com nós de alta pontuação, tendam a ter alta pontuação. Por outro lado, os nós que se conectem com nós de baixa pontuação, tendam a ter baixa pontuação. O algoritmo *Page Rank* é uma variação mais recente da métrica *eigenvector*, o que justifica o fato desta métrica não ser calculada para todos os nós das comunidades analisadas neste trabalho.

A partir do grafo de cada comunidade, foi extraído um subgrafo para facilitar o cálculo das métricas *betweenness*, *closeness* e *eigenvector*. Cada subgrafo extraído contém somente nós do componente *Core* de cada comunidade, uma vez que este componente aparenta representar melhor as interações nas comunidades (além das boas correlações obtidas). Cada nó escolhido para compor o subgrafo foi aleatoriamente selecionado do componente *Core* de cada comunidade. A Tabela 14 mostra os dados de cada subgrafo.

Tabela 14. Dados dos subgrafos

Comunidade	**Número de Nós**	**Número de Arestas**
Biology Q&A	1 100	7 711
English Language and Usage	3 001	31 010
Physics Q&A	3 001	29 811
Mathematics Q&A	3 001	13 652
Travel Answers	1 479	12 302

Depois disto, foram calculadas as correlações entre a reputação (indicador de competência) dos usuários com as três métricas descritas nesta seção (*betweenness*, *closeness* e *eigenvector*). Na Tabela 15 se pode verificar que as correlações envolvendo a métrica *betweenness* obtiveram os melhores resultados. Um nó (usuário) com *betweeness* alta significa que ele conecta diferentes partes de uma rede, evitando criar um grafo desconexo. A métrica *eigenvector* também obteve boas correlações que, por sua vez, são similares às correlações do *Page Rank* apresentadas na seção anterior. A métrica *closeness*, apesar de em alguns casos ter correlações moderadas, na maioria as correlações foram fracas. Contudo, se pode concluir que a métrica *closeness* obteve as piores correlações. Isto significa que, uma métrica que mede a velocidade com que uma informação é espalhada pode não ser útil quando se deseja encontrar os usuários confiáveis de uma comunidade.

Tabela 15.Coeficiente de Correlação de Pearson - Subgrafos

Comunidade	**Betweenness**	**Closeness**	**Eigenvector**
Biology Q&A	0.92	0.45	0.88
English Language and Usage	0.94	0.19	0.80
Physics Q&A	0.84	0.17	0.67
Mathematics Q&A	0.87	0.18	0.80
Travel Answers	0.94	0.39	0.78

Em síntese, baseado nos resultados apresentados nesta seção se pode concluir que quanto maior o valor da métrica *betweenness* e *eigenvector*, maiores são as chances de um usuário ser confiável. Porém, considerando os resultados e o tempo de processamento que essas métricas

podem levar, as métricas mostradas na seção 4.2 podem ser mais interessantes, pois, os resultados das correlações foram similares.

4.4. Comentários Finais

Neste capítulo foi apresentada uma forma de dividir uma comunidade em partes através de seus padrões de interação denominada estrutura Bow Tie. As cinco comunidades analisadas neste trabalho foram divididas de acordo com os componentes (partes) da estrutura Bow Tie. Em seguida, os componentes da estrutura Bow Tie de cada uma das cinco comunidades foram comparados entre si, bem como, comparados com componentes Bow Tie de outras redes analisadas em trabalhos previamente feitos. A partir destas comparações, se pode concluir que as cinco comunidades analisadas nessa dissertação são lugares onde grande parte dos usuários estão dispostos a ajudar e serem ajudados, devido ao tamanho do componente *Core* de cada uma delas.

Além disso, foram analisados e correlacionados alguns atributos dos usuários das cinco comunidades com os seus respectivos indicadores de competência (reputação). As melhores correlações foram obtidas no componente *Core* de cada comunidade. Isto pode significar que, para encontrar os usuários confiáveis das comunidades, talvez seja interessante considerar somente os atributos dos usuários no componente *Core* em vez de considerar amostras aleatórias da comunidade.

Foram analisadas também três métricas clássicas geralmente utilizadas em problemas de análises de redes sociais. Essas métricas foram: *betweenness*, *closeness* e *eigenvector*. Foi mostrado que o cálculo da métrica *betweenness* e *closeness* é computacionalmente muito custoso. Devido a este fato, o cálculo de tais medidas pode ser um gargalo durante as construções de análises. Foi visto também que a métrica *eigenvector* se trata de uma variação mais antiga do algoritmo *Page Rank*. Dado isto, foi decidido somente calcular essas métricas para alguns usuários do componente *Core* de cada comunidade analisada e, logo após, verificar a sua relação com a reputação (indicador de competência) dos usuários. Os resultados mostraram que os usuários que apresentem valores altos das métricas *eigenvector* e *betweenness* tendem a ter maior reputação nas comunidades, de acordo com as correlações apresentadas. Já a métrica *closeness* apresentou as piores correlações com o indicador de competência.

5. Aprendizado de Máquina para Encontrar Usuários

Este capítulo tem como finalidade mostrar a proposta de uso de dois métodos de aprendizado de máquina para classificar um usuário de acordo com o seu grau de confiabilidade. Os métodos de aprendizado de máquina utilizado nesta dissertação foram uma rede neural artificial, denominada de *Perceptron* Multicamadas, e um algoritmo de agrupamento, denominado *K-means*. O objetivo do uso dessas técnicas é prover uma forma eficaz de encontrar os usuários confiáveis de uma comunidade online.

Espera-se que a rede neural artificial utilizada neste trabalho possibilite a classificação de qualquer usuário das comunidades analisadas de acordo com uma escala de confiabilidade. Essa escala tem como finalidade definir níveis de confiabilidade de um usuário, variando desde o nível menos confiável até o mais confiável. Já o algoritmo de agrupamento possibilitará encontrar grupos de usuários com maiores probabilidades de se encontrar um usuário confiável.

Este capítulo contém partes do trabalho apresentado em (PROCACI et al., 2014d), que também foi escrito pelo autor desta dissertação.

5.1. Aprendizado de Máquina

Aprendizado de máquina é a ciência que estuda formas para fazerem com que computadores executem alguma tarefa sem que esta seja explicitamente programada. As técnicas de aprendizado de máquina são modelos computacionais de caráter geral, ou seja, algoritmos que podem ser aplicados nos mais diversos domínios. Contudo, para fazer com que esses modelos computacionais executem alguma tarefa específica, é necessário alimentá-los com dados. São através desses dados que esse modelo executa suas funções e fornece resultados. A ideia dessas técnicas é fazer com que uma máquina aprenda a executar uma determinada tarefa somente considerando dados que lhes são fornecidos não sendo, portanto, necessário programar tal tarefa explicitamente.

Na década passada e na atual, as técnicas de aprendizado de máquina foram e têm sido amplamente aplicadas nas mais diversas áreas trazendo grandes inovações tecnológicas e facilidades para o dia-a-dia das pessoas. Dentre essas inovações se podem citar: carros que são conduzidos sem intervenção humana, softwares que reconhecem falas de seres humanos, motores de buscas na Web mais eficientes, softwares anti-spam etc. As técnicas de aprendizado de máquina estão tão presentes no mundo atual, que é muito provável que pessoas as usem várias vezes ao dia

mesmo sem saber de sua existência. Duas das técnicas de aprendizado de máquina amplamente utilizada são as redes neurais artificiais e os algoritmos de agrupamento.

5.2. Redes Neurais Artificiais

Redes neurais artificiais são técnicas computacionais que apresentam um modelo matemático inspirado na estrutura neural de organismos inteligentes e que adquirem conhecimento através da experiência. Uma rede neural artificial é composta por várias unidades de processamento. Essas unidades geralmente são conectadas por canais de comunicação que estão associados a determinado peso. As unidades fazem operações apenas sobre dados que lhe são fornecidos O comportamento inteligente de uma rede neural artificial vem das interações entre as unidades de processamento da rede.

As unidades de processamento de uma rede neural artificial foram propostas por MCCULLOCH & PITTS (1943). Seu funcionamento é ilustrado na Figura 11 e pode ser descrito simplificadamente da seguinte forma:

- Os sinais de entradas (X_1, X_2, ..., X_p) são os dados de entrada da unidade. Esses dados de entrada são dados numéricos (números inteiros ou reais);
- Cada dado de entrada é associado a um peso (W_1, W_2, ..., W_p). Esse peso é um número que indica a influência de cada entrada. Desta forma, cada dado de entrada é multiplicado pelo seu peso;
- Em seguida, é feita a soma ponderada ($\sum$) dos dados de entrada que, por sua vez, produz um resultado;
- Se este resultado exceder certo limite, a unidade produz uma determinada resposta de saída (y). A resposta de saída é produzida através de uma função matemática.

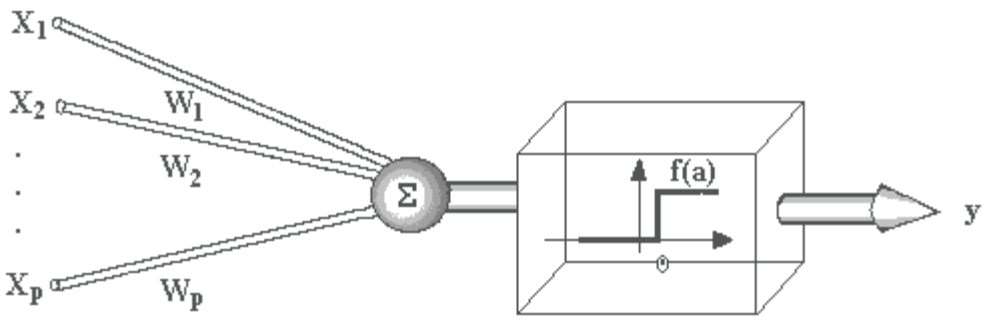

Figura 11. Unidade Processamento de uma Rede Neural Artificial

Como já relatado, as redes neurais artificiais são tipicamente compostas por várias unidades de processamentos que são conectadas entre si. A propriedade mais importante das redes neurais artificiais é a habilidade de aprender com base nos seus dados de entrada e com isso melhorar seu desempenho para executar uma determinada tarefa. Isso é feito através de um processo iterativo de ajustes aplicado a seus pesos. Esse processo de ajustes dos pesos é denominado fase de treinamento. Em geral, os pesos são ajustados até o momento em que a rede neural pode executar a tarefa que lhe é solicitada com um grau de assertividade desejado.

Para os ajustes dos pesos de uma rede neural, é utilizado um algoritmo de aprendizado. Denomina-se algoritmo de aprendizado a um conjunto de regras bem definidas que descrevem o modo como os pesos são modificados.

5.2.1. *Perceptron* Multicamadas

Para alcançar o objetivo desta dissertação, descobrir maneiras para encontrar os usuários confiáveis de uma comunidade, foi utilizada uma rede neural artificial denominada *Perceptron* Multicamadas para este fim. Neste trabalho, foi usada a ferramenta Weka[32] para que fosse possível fazer os testes com o *Peceptron*. Além disso, foi escrito um script na linguagem de programação Python, com a finalidade de gerar os arquivos necessários para configurar o *Perceptron* no Weka, com os dados dos usuários das cinco comunidades.

Os *Perceptrons* têm sido amplamente utilizados em problemas de classificação como no caso deste trabalho, onde se deseja classificar um usuário de acordo com o seu grau de confiança. A partir dessa classificação, será possível encontrar os usuários confiáveis.

O *Perceptron* foi inicialmente proposto por ROSENBLATT (1957) onde, em sua concepção inicial, só era possível classificar problemas cujo conjunto de dados de entrada são linearmente separáveis. Após evoluções no *Perceptron*, concluíram que com algumas adaptações seria possível tratar também problemas linearmente não separáveis. Desta forma, surgiu o *Perceptron* multicamadas.

Neste trabalho, os dados de entrada para o *Perceptron* Multicamadas serão os atributos dos usuários das comunidades (descritos no capítulo 3). Infelizmente, em comunidades online, não há como garantir que os valores dos atributos dos usuários serão sempre linearmente separáveis. Desta forma, foi decidido utilizar o *Perceptron* Multicamadas devido a sua capacidade de classificar conjuntos de dados que são linearmente não separáveis.

[32] Site: http://www.cs.waikato.ac.nz/ml/weka/

Em síntese, o funcionamento de uma rede neural artificial se resume a duas fases: a fase de treinamento e a fase de aplicação. A fase de treinamento objetiva configurar a rede neural para, em seguida, utilizá-la na fase de aplicação. Uma vez configurada a rede (treinada), é esperado que ela classifique um usuário em confiável ou não na fase de aplicação.

Além dos dados de entrada, o *Perceptron* Multicamadas exige que sejam indicados os resultados esperados (as classificações desejadas) para cada dado de entrada na fase de treinamento. São através dos dados de entradas e as saídas desejadas que o *Perceptron* poderá ajustar os pesos de cada entrada. Desta forma, o *Perceptron* aprende como classificar um usuário de acordo com os dados de entrada e as saídas desejadas. Para definir os parâmetros de saída da rede neural, ou seja, os resultados desejados, foi utilizada uma forma de classificação provida por um esquema de médias aritméticas.

5.2.2. Esquema de Médias Aritméticas

O esquema de médias aritméticas proposto neste trabalho tem como objetivo definir as possíveis classificações de um usuário das comunidades. Essas classificações variam desde não confiável até extremamente confiável.

O esquema de médias aritméticas é baseado na reputação (indicador de competência) fornecida pelas comunidades. Como já relatado, a reputação de um usuário é construída com base em avaliações feitas por outros usuários das comunidades. Em outras palavras, o usuário tem suas respostas, comentários e perguntas avaliadas por outros usuários.

Baseados nisto, médias aritméticas sucessivas das reputações dos usuários de cada uma das cinco comunidades foram feitas, com o objetivo de identificar quem são os usuários confiáveis da comunidade. Primeiramente, foi preciso encontrar a reputação maior e a menor na comunidade. A Figura 12 mostra um exemplo deste esquema considerando a maior reputação é igual a cem (MAX) e o menor igual a zero (MIN). Uma vez conhecido a maior e a menor reputação, foi calculada a média aritmética entre elas. O resultado desta operação foi chamado de M1. Em seguida, foi calculada a média aritmética entre M1 e MAX e o resultado foi denominado M3. Posteriormente, calculou-se a média aritmética entre M1 e M3 cujo resultado foi chamado de M2. Depois disso, foi calculada a média aritmética entre M3 e MAX e o resultado foi chamado de M4. Por fim, foi calculada a média aritmética entre M4 e MAX e o resultado foi denominado M5.

É importante mencionar que a decisão da escolha pelo uso de médias aritméticas neste trabalho foi porque, nesse tipo de média, os valores extremos influenciam os resultados

intermediários. Em outras palavras, a classificação de qualquer usuário das comunidades depende do usuário mais confiável (maior reputação) e do menos confiável (menor reputação). Depois destes cálculos, a confiabilidade de um usuário foi definida de acordo com a Tabela 16. Em síntese, buscou-se classificar como não confiável os usuários que ficaram abaixo da média aritmética entre a maior e a menor reputação de cada comunidade. Para os usuários que ficaram acima desta média, buscou-se definir faixas de confiabilidades, variando desde pouco confiável até extremamente confiável.

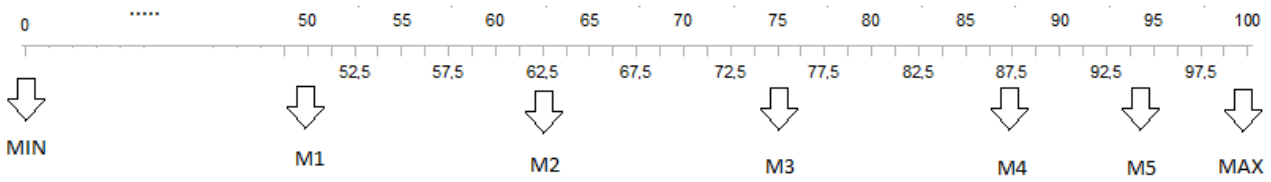

Figura 12. Esquema de Médias Aritméticas

Tabela 16. Classificação dos Usuários

Condição	**Classificação**
Reputação <= M1	Não confiável
M1 < Reputação <= M2	Pouco confiável
M2 < Reputação <= M3	Razoavelmente confiável
M3 < Reputação <= M4	Confiável
M4 < Reputação <= M5	Muito confiável
M5 < Reputação <= MAX	Extremamente confiável

A ideia essa classificação dos usuários baseada na reputação é a mesma daquela apresentada no estudo comparativo do grupo do Facebook no capítulo 3. Pretende-se saber previamente qual é a classificação de um usuário, considerando a sua reputação construída na comunidade, para que, através disto, possa ser possível treinar e testar se a rede neural proposta para classificar os usuários pode trazer bons resultados. Porém, no caso do estudo comparativo apresentado no capítulo 3, essa classificação prévia do usuário foi feita por um profissional, através de uma análise humana. Neste capítulo, a classificação prévia será através da reputação, ou seja, o indicador de competência escolhido.

Neste cenário, uma vez sabendo previamente qual é a classificação do usuário, será utilizado o conjunto de atributos dos usuários para treinar a rede neural e, em seguida, verificar se

através deles, a rede neural pode classificar corretamente um usuário. Em outras palavras, a classificação prévia proposta na Tabela 16 servirá de base para a validação das classificações fornecidas pela rede neural.

5.2.1. *Perceptron* Multicamadas como Classificador de Usuários

Basicamente, a rede neural utilizada irá classificar um usuário de acordo com a classificação proposta na Tabela 16 e terá como parâmetros de entrada os atributos dos usuários analisados. Através dessa classificação, será possível encontrar os usuários mais confiáveis. A Figura 13 mostra como é a rede neural (*Perceptron* Multicamadas) usada neste trabalho para classificar os usuários. A rede neural usada é composta por treze neurônios e função que produz a saída de cada neurônio foi a Sigmoide. É importante ressaltar que chegou-se a essa configuração da rede neural através de testes realizados. Uma vez tendo a rede neural pré-configurada, o próximo passo é escolher quais serão os dados que farão parte do conjunto de treinamento.

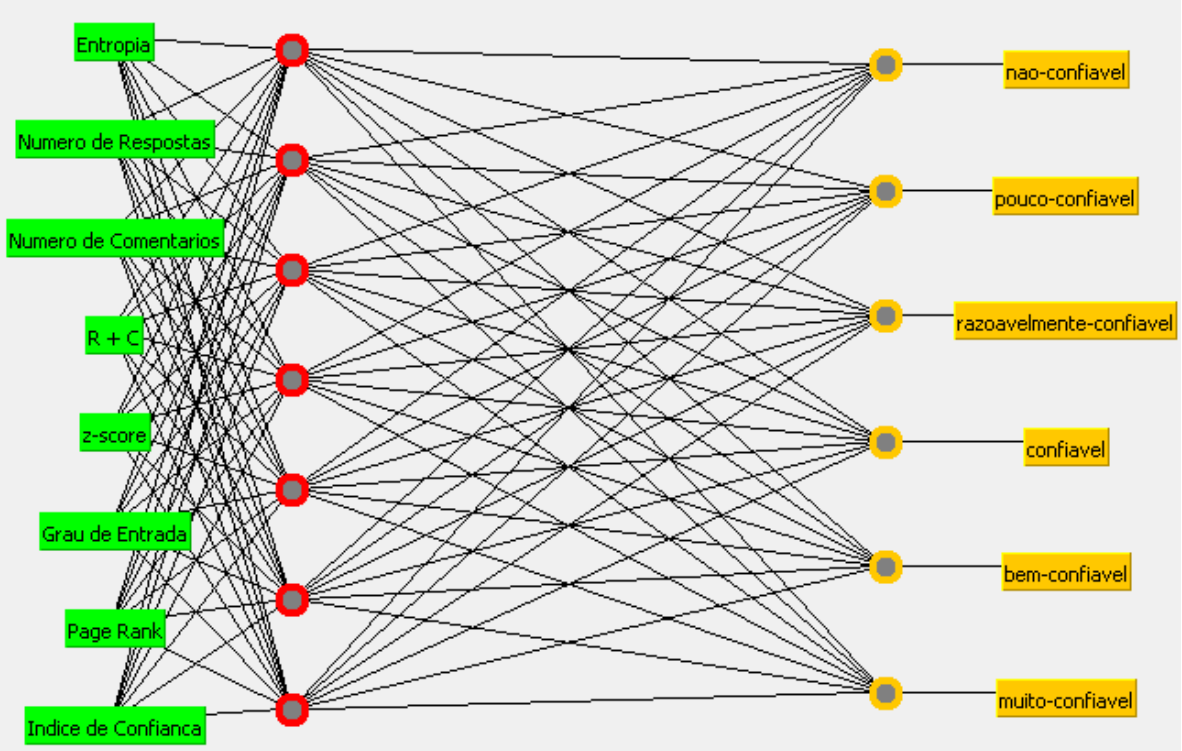

Figura 13. Perceptron Multicamadas

5.2.2. Dados de Treinamento

Geralmente, os dados de treinamento são um subconjunto dos dados do problema

analisado. No caso deste trabalho, os dados de entrada da rede serão os atributos dos usuários. Todavia, para a fase de treinamento da rede neural, é desejável definir um subconjunto de dados, que represente todas as possíveis classificações dos usuários. Desta forma, é possível encontrar uma configuração para a rede neural onde ela possa classificar com um bom grau de assertividade qualquer usuário (inclusive aqueles que não fizeram parte dos dados de treinamento).

Para o conjunto de dados de treinamento, foi escolhido o componente *Core* de cada uma das comunidades analisadas. Os motivos da escolha do *Core* como conjunto de dados de treinamento foram as boas correlações obtidas dos atributos com a reputação do usuário, conforme descritas no capítulo 4. Além disso, a fim de buscar evidências que o componente *Core* realmente representa todos os tipos de usuários das comunidades, foram elaboradas as Tabelas 17, 18, 19, 20 e 21 que mostra a quantidade de usuários de cada tipo (segundo a classificação da Tabela 16) em cada um dos componentes da estrutura Bow Tie.

Tabela 17. Classificação dos Usuários Biology Q&A – Quantidade por Componente

Classificação	Core	IN	OUT	Tendrils	Tubes	Disconnected
Não confiável	792	744	264	60	0	130
Pouco confiável	139	11	3	2	0	2
Razoavelmente confiável	90	0	0	0	0	0
Confiável	50	0	0	0	0	1
Muito confiável	21	0	0	0	0	0
Extremamente confiável	8	0	0	0	0	0

Tabela 18. Classif. dos Usuários English Lang. and Usage – Quantidade por Componente

Classificação	Core	IN	OUT	Tendrils	Tubes	Disconnected
Não confiável	7635	5245	3822	628	3	815
Pouco confiável	1189	41	33	5	0	11
Razoavelmente confiável	665	2	0	1	0	0
Confiável	232	0	0	0	0	0
Muito confiável	55	0	0	0	0	0
Extremamente confiável	26	0	0	0	0	0

Tabela 19. Classificação dos Usuários Physics Q&A – Quantidade por Componente

Classificação	Core	IN	OUT	Tendrils	Tubes	Disconnected
Não confiável	5884	4628	2034	482	1	764
Pouco confiável	908	61	17	3	0	16
Razoavelmente confiável	585	7	0	0	0	3
Confiável	284	0	0	0	0	0
Muito confiável	55	0	0	0	0	0
Extremamente confiável	21	0	0	0	0	0

Tabela 20. Classificação dos Usuários Mathematics Q&A – Quantidade por Componente

Classificação	Core	IN	OUT	Tendrils	Tubes	Disconnected
Não confiável	21239	17265	4202	862	3	2687
Pouco confiável	2476	113	124	30	0	29
Razoavelmente confiável	1406	9	20	2	0	2
Confiável	608	0	0	0	0	0
Muito confiável	119	0	0	0	0	0
Extremamente confiável	49	0	0	0	0	0

Tabela 21. Classificação dos Usuários Travel Answers – Quantidade por Componente

Classificação	Core	IN	OUT	Tendrils	Tubes	Disconnected
Não confiável	1087	1028	797	110	1	143
Pouco confiável	222	7	12	1	0	0
Razoavelmente confiável	99	1	0	0	0	0
Confiável	53	0	0	0	0	0
Muito confiável	12	0	0	0	0	0
Extremamente confiável	6	0	0	0	0	0

A partir das Tabelas 17, 18, 19, 20 e 21 se percebe que o componente *Core* é o único que contempla todos os tipos de usuários, de acordo com a classificação proposta na Tabela 16. Diante disto, além das boas correlações obtidas no capítulo 4, foi escolhido o componente *Core* como conjunto de dados de treinamento da rede neural utilizada neste trabalho. Uma vez definido o conjunto de dados de treinamento, iniciou-se a fase de treino da rede neural e, após isso, obteve-se a rede configurada para o uso.

5.2.3. Testes com o *Perceptron* Multicamadas

Depois de treinar a rede neural, a etapa de aplicação foi iniciada. A ideia dessa etapa é utilizar a rede neural já configurada para classificar todos os usuários de cada comunidade. Desta forma, a etapa de aplicação foi executada e, em seguida, as classificações dos usuários providas pela rede neural foram comparadas com as classificações providas pelo esquema de médias aritméticas.

A Tabela 22 mostra os resultados depois de executar a rede neural treinada em cenários distintos (colunas). Analisando os resultados da Tabela 22, se percebe que a rede neural classificou corretamente a maioria dos usuários. Contudo, somente verificar a porcentagem de usuários classificados corretamente não é suficiente para se tirar alguma conclusão sobre a eficácia da rede neural (qualidade do classificador). A porcentagem de usuários corretamente classificados não leva em consideração as classificações feitas ao acaso. Dado este problema, foi utilizado o coeficiente

kappa com a finalidade de verificar se, executando a rede neural diversas vezes com o mesmo conjunto de dados, pode trazer classificações similares. Em outras palavras, se buscou saber a concordância entre as várias classificações (se a classificação de um usuário persiste depois de executar a rede neural várias vezes que, neste caso, foram 500 iterações). É importante ressaltar que, para o teste do coeficiente kappa, a cada vez que a rede neural foi executada, ela era reconfigurada (os pesos eram modificados) e isto pode alterar os resultados das futuras classificações. Por este motivo, utilizou-se o coeficiente kappa, objetivando saber se a rede neural está suficiente treinada.

Na Tabela 22, as siglas BQA refere a comunidade *Biology Q&A*, ELU a comunidade *English Language and Usage*, PQA a comunidade *Physics Q&A*, MQA a comunidade *Mathematics Q&A* e TA a comunidade *Travel Answers*.

Tabela 22. Comparação: Rede Neural X Reputação

	BQA	**ELU**	**PQA**	**MQA**	**TA**	**BQA**	**ELU**	**PQA**	**MQA**	**TA**
Usuários classificados corretamente	89,2 %	91,0%	90,5%	92,1 %	91,5%	88,7%	91,5%	90,0%	92,2%	90,2%
Usuários classificados não corretamente	10,8%	9,0%	9,6%	7,9%	8,5%	11,3%	8,5%	10,0%	7,8%	9,8%
Kappa	0,476	0,551	0,4707	0,4491	0,5401	0,4718	0,5294	0,5099	0,4339	0,582
Dados de Treinamento	Core BQA	Core ELU	Core PQA	Core MQA	Core TA	20% Core BQA	20% Core ELU	20% Core PQA	20% Core MQA	20% Core TA

LANDIS & KOCH (1977) caracterizaram o coeficiente kappa como: menor que 0 indicando ausência de concordância entre os classificadores; de 0 até 0,20 como concordância leve; acima de 0,20 até 0,40 como concordância razoável; acima de 0,40 até 0,60 como concordância moderada; acima de 0,60 até 0,80 como concordância substancial; e acima de 0,80 até 1 como concordância perfeita. FLEISS (1981) caracterizou o coeficiente kappa de forma similar: os valores acima de 0,75 como concordância excelente; de 0,40 até 0,75 como concordância razoável ou boa; e abaixo de 0,40 como concordância fraca.

Através da Tabela 22, é possível verificar que todos os valores do coeficiente kappa estão acima de 0,40 até 0,60 e que, de acordo com LANDIS & KOCH (1977), é um indicador moderado e, de acordo com FLEISS (1981), é um indicador razoável ou bom. Através destes testes, se conclui que utilizar a rede neural proposta parece ser um método promissor para identificar os usuários

confiáveis de comunidades.

Outro fato interessante é que usando apenas 20% *Core* como conjunto de dados de treinamento também pode trazer bons resultados. Além disso, mais do que trazer bons resultados, é importante ressaltar que o uso de 20% do *Core* traz resultados muito parecidos quando comparados com o uso do Core inteiro como dados de treinamento. Isto pode significar mais agilidade e facilidade para configurar uma rede neural.

5.2.4. Capacidade de Generalização da Rede Neural Artificial

Um fato notável é que comunidades online são dinâmicas. Em outras palavras, em comunidades online, o número de usuários podem aumentar ou diminuir ao longo do tempo, novas interações podem ocorrer a qualquer momento etc. Desta forma, é necessário averiguar se a rede neural utilizada traz resultados similares a aqueles apresentados na seção anterior, quando os dados envolvidos em sua configuração e aplicação mudam. Como já visto, o uso do *Core* ou 20% dele como dados de treinamento da rede neural podem trazer bons resultados. Contudo, com o decorrer do tempo, o próprio *Core* da comunidade irá se modificar, à medida que a comunidade evolui.

Dado este fato, buscou-se nesta dissertação uma maneira para avaliar a capacidade de generalização da rede neural utilizada. Para avaliar isto, foi utilizado um método de validação cruzada denominado *k-fold* (GEISSER, 1993). Os métodos de validação cruzada são amplamente empregados em problemas onde o objetivo da modelagem é a predição. Desta forma, busca-se estimar o quão preciso é este modelo na prática, ou seja, o seu desempenho para um novo conjunto de dados (KOHAVI, 1995). No caso deste trabalho, o modelo utilizado é uma rede neural artificial. Neste contexto, como o objetivo da rede neural usada é predizer qual é a classificação dos usuários de acordo com seu nível de confiabilidade (conforme a Tabela 16) e as comunidades online são dinâmicas, é fundamental verificar capacidade de generalização do modelo. O método de validação cruzada *k-fold* consiste em dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho. No caso deste trabalho, o número de subconjuntos utilizados foram dez (k = 10). A partir disto, dez testes foram feitos na rede neural, onde em cada teste, nove subconjuntos dos dados (k - 1) eram usados na fase de treinamento e um subconjunto era usado na fase de aplicação. Em outras palavras, este processo foi realizado dez vezes, alternando de forma circular o subconjunto de teste. A Tabela 23 mostra os resultados da validação cruzada aplicados aos usuários de cada comunidade, após aplicar o processo descrito dez vezes. As siglas utilizadas na Tabela 23 têm os mesmos significados daqueles apresentados na Tabela 22. Para a realização da validação cruzada, foi utilizada a ferramenta Weka.

Tabela 23. Validação Cruzada - k-fold

	BQA	ELU	PQA	MQA	TA
Usuários classificados corretamente	88,6 %	91,6%	90,1%	92,1 %	91,6%
Usuários classificados não corretamente	11,4%	8,4%	9,9%	7,9%	8,4%
Kappa	0,4506	0,5099	0,481	0,467	0,5475

Através dos resultados é possível verificar que, após a validação cruzada, foi obtido resultados muito próximos daqueles apresentados da Tabela 22, onde é mostrado o uso do Core e 20% dele para treinamento da rede neural. A partir disto se pode concluir que a rede neural utilizada tem capacidade de generalização à medida que seu conjunto de dados é modificado.

5.2.5. *Core* de Outras Comunidades como Dados de Treinamento

Neste trabalho foi avaliada uma maneira para averiguar a possibilidade de se utilizar redes neurais treinadas com dados de uma comunidade e, logo após, aplicá-la em outra comunidade.

Tabela 24. Comparação: neural rede x reputação (usando Core de diferentes comunidades)

Comunidade	Usuários classificados corretamente	Usuários classificados não corretamente	Kappa	Dados de treinamento
BQA	86,49%	13,51%	0,3795	Core ELU
BQA	87,69%	12,31%	0,3974	Core PQA
BQA	78,67%	21,33%	0,3225	Core MQA
BQA	87,87%	12,13%	0,4091	Core TA
ELU	90,53%	9,47%	0,4518	Core BQA
ELU	91,39 %	8,61%	0,4837	Core PQA
ELU	91,26 %	8,74 %	0,4627	Core MQA
ELU	90,90%	9,10%	0,5103	Core TA
PQA	89,53 %	10,47 %	0,427	Core BQA
PQA	89,83%	10,17%	0,4974	Core ELU
PQA	90,20%	9,80 %	0,4555	Core MQA
PQA	90,00%	10,00%	0,476	Core TA
MQA	90,34%	9,66%	0,3999	Core BQA
MQA	90,50%	9,50%	0,4751	Core ELU
MQA	91,24%	8,76%	0,4304	Core PQA
MQA	90,65%	9,35 %	0,443	Core TA
TA	90,66 %	9,34 %	0,4687	Core BQA
TA	90,94%	9,06 %	0,5424	Core ELU
TA	90,51%	9,49 %	0,4461	Core PQA
TA	87,84%	12,16 %	0,4248	Core MQA

Tabela 25. Comparação: neural rede x reputação (usando 20% do componente Core de diferentes comunidades)

Comunidade	Usuários classificados corretamente	Usuários classificados não corretamente	Kappa	Dados de treinamento
BQA	87,57%	12,43%	0,3763	20% Core ELU
BQA	87,61 %	12,39 %	0,401	20% Core PQA
BQA	88,34 %	11,66%	0,434	20% Core MQA
BQA	87,69%	12,31 %	0,3761	20% Core TA
ELU	89,58 %	10,42%	0,4338	20% Core BQA
ELU	91,25%	8,75%	0,5258	20% Core PQA
ELU	91,23%	8,77%	0,4448	20% Core MQA
ELU	91,35%	8,65%	0,5122	20% Core TA
PQA	89,32%	10,68%	0,4369	20% Core BQA
PQA	90,16%	9,84%	0,4523	20% Core ELU
PQA	89,89 %	10,11%	0,4047	20% Core MQA
PQA	90,19 %	9,81%	0,4652	20% Core TA
MQA	89,85%	10,15%	0,3888	20% Core BQA
MQA	91,18%	8,82%	0,4372	20% Core ELU
MQA	90,93%	9,07 %	0,4559	20% Core PQA
MQA	91,02%	8,98%	0,4401	20% Core TA
TA	89,91%	10,09%	0,4528	20% Core BQA
TA	91,01%	8,99%	0,4837	20% Core ELU
TA	91,22%	8,78%	0,5289	20% Core PQA
TA	90,66%	9,34%	0,44	20% Core MQA

A fim de verificar a possibilidade de se usar uma rede neural que foi previamente treinada em uma determinada comunidade e utilizar em outras comunidades, foram realizados alguns testes, treinando a rede neural (*Perceptron* Multicamadas) com os atributos dos usuários do *Core* de uma comunidade e, em seguida, tentando classificar todos os usuários de outra comunidade. As Tabelas 24 e 25 mostram os resultados da classificação usando o *Core* de outras comunidades. As siglas utilizadas nas Tabelas 24 e 25 têm os mesmos significados daqueles apresentados na Tabela 22.

Através das Tabelas 24 e 25 se percebe que as redes neurais conseguem classificar corretamente boa parte dos usuários. A maioria dos cenários apresentados nas Tabelas 24 e 25 mostram uma porcentagem maior que 80% de usuários classificados corretamente, tanto quando se utiliza o *Core* quanto 20% dele na fase de treinamento.

Além disso, através das Tabelas 24 e 25, é possível verificar que grande parte dos valores do coeficiente kappa está acima de 0,40 até 0,60, o que, de acordo com LANDIS & KOCH (1977), é um indicador moderado e, de acordo com FLEISS (1981), é um indicador razoável ou bom. Todos os casos em que o coeficiente kappa ficou menor que 0,40 estava envolvida a comunidade *Biology*

Q&A (BQA), ora como comunidade onde a rede neural foi testada e ora como conjunto de dados de treinamento. Esta comunidade é a menor entre as analisadas e, talvez por isto, os resultados providos pelo coeficiente kappa não foram satisfatórios. Desta forma, quando se desejar utilizar essa abordagem para a classificação de usuários, se deve atentar ao tamanho da comunidade em questão. Todavia, em alguns casos, mesmo utilizando comunidades menores como a *Biology Q&A* e *Travel Answers* (TA), o coeficiente de kappa obteve bons resultados. Porém, aparentemente escolher comunidades menores para essa abordagem não parece uma opção segura. Dado esses resultados e ainda considerando o escopo deste trabalho, se pode concluir que para a utilização de uma rede neural previamente treinada e, em seguida, aplicada a outra comunidade é mais seguro utilizar dados provenientes de comunidades um pouco maiores (no mínimo do tamanho similar a *Physics Q&A*) e estas devem ter tamanhos similares. Para este cenário, não foi realizada a validação cruzada, visto que, isto envolveria considerar as cinco comunidades como uma que, por sua vez, é um cenário não realista na prática. A ideia aqui é somente verificar simplificadamente se uma rede neural com boa capacidade de generalização e com boa taxa de assertividade de uma comunidade, pode ser útil para classificar usuários de outra.

5.3. Algoritmo de Agrupamento

Com a finalidade de verificar se outra abordagem pode trazer resultados melhores ou similares ao uso da rede neural, foi testado um algoritmo de agrupamento denominado *k-means* para encontrar os usuários confiáveis das comunidades. Os algoritmos de agrupamento também são conhecidos como algoritmos de clusterização.

A ideia do uso do algoritmo de agrupamento neste trabalho não é conseguir dar uma classificação exata sobre a confiabilidade de um usuário (como feito na rede neural) mas identificar relações em grupos de usuários formados que permitam encontrar aqueles com uma confiabilidade mínima desejada.

5.3.1. K-means

O algoritmo *k-means* foi apresentado por MACQUEEN (1967) e trata-se de um dos algoritmos mais conhecidos de agrupamentos. Este algoritmo tenta fornecer uma classificação de acordo com os próprios dados. A classificação é feita por similaridade onde um objeto é atribuído a um grupo (*cluster*) ao qual é mais semelhante.

A ideia principal desse algoritmo é escolher k objetos (o número k representa também o número de grupos a serem formados e podem ser escolhidos de forma aleatória ou através de

alguma heurística) que serão a base de cada grupo (denominados centróides). Os demais objetos serão associados ao centróide mais próximo (ou similar, através da função de similaridade). A cada passo, os centróides são recalculados dentre os objetos de seu grupo e os objetos são realocados mais uma vez para os centróides mais próximos. Este procedimento é repetido até o algoritmo encontrar um nível de convergência satisfatório. No contexto do trabalho, os objetos a serem agrupados são os usuários das comunidades.

A Figura 14 mostra um exemplo do *k-means* em funcionamento.

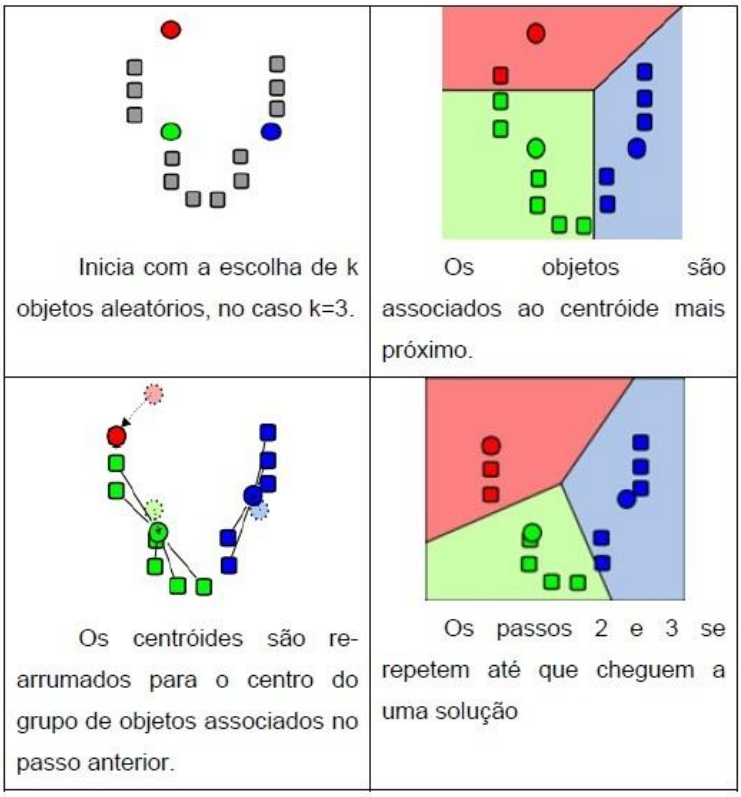

Figura 14. Exemplo do k-means (WIKIPEDIA, 2014)

5.3.2. *K-means* para Encontrar Grupos de Usuários Confiáveis

Com o objetivo de encontrar grupos de usuários confiáveis das comunidades analisadas nesta dissertação, o *k-means* foi implementado utilizando a linguagem de programação Python e configurado para formar cinco grupos de usuários, recebendo como entrada seus atributos. A ideia é que o *k-means* forme grupos com usuários similares baseado em seus atributos. Os atributos usados no *k-means* foram os mesmos utilizados na rede neural, conforme a Figura 13. Além disso, decidiu

por formar cinco grupos de usuários para não ser possível, em nenhuma hipótese, que em cada grupo contenha somente usuários com classificações iguais, conforme aquelas mostradas na Tabela 16. O que se deseja com o uso do *k-means* é descrever o que existe dentro de cada grupo e, a classificação da Tabela 16, será usada somente como um parâmetro de comparação que auxiliará na descrição dos dados de cada grupo.

Tabela 26. Grupos formados pelo k-means

Grupo	Comunidade	Número de Usuários	Número Médio de Respostas	Número de Não Confiáveis	Número de Pouco Confiáveis	Número de Razoavelmente Confiável	Número de Confiável	Número de Muito Confiável	Número de Extremamente Confiável	Probabilidade de Ser Confiável ou Acima	Probabilidade de Ser Razoavelmente Confiável ou Abaixo
0	BQA	20	62.45	0	0	0	7	8	5	100%	0%
1	BQA	2 000	0.4455	1 904	81	14	1	0	0	0.05%	99.95%
2	BQA	7	172.1429	0	0	0	1	3	3	100%	0%
3	BQA	55	23.45455	0	3	13	29	10	0	70.90909%	29.09091%
4	BQA	235	4.438298	86	73	63	13	0	0	5.531915%	94.46809%
0	PQA	175	77.44	0	2	14	120	37	2	90.85714%	9.142857%
1	PQA	27	288	0	0	1	1	12	13	96.2963%	3.703704%
2	PQA	7	1202	1	0	0	0	0	6	85.71429%	14.28571%
3	PQA	14 967	0.702546	13 752	915	296	4	0	0	0.026725%	99.97327%
4	PQA	577	17.75043	40	88	284	159	6	0	28.59619%	71.40381%
0	ELU	11	1 007.273	0	0	0	0	2	9	100%	0%
1	ELU	18 898	0.716372	17 922	841	135	0	0	0	0%	100%
2	ELU	51	378.9216	1	0	0	6	29	15	98.03922%	1.960784%
3	ELU	205	92.53659	1	5	33	140	24	2	80.97561%	19.02439%
4	ELU	1 243	11.66613	224	433	500	86	0	0	6.918745%	93.08126%
0	MQA	47 080	0.424533	45 083	1 599	381	17	0	0	0.036109%	99.96389%
1	MQA	18	1 806.389	0	0	0	0	3	15	100%	0%
2	MQA	3 223	10.1542	1 167	1 120	751	180	5	0	5.739994%	94.26001%
3	MQA	767	85.11864	8	53	303	348	51	4	52.54237%	47.45763%
4	MQA	157	395.7134	2	0	2	63	60	30	97.45223%	2.547771%
0	TA	2 950	0.445763	2 891	56	3	0	0	0	0%	100%
1	TA	31	125.4516	0	0	0	16	11	4	100%	0%
2	TA	81	25.44444	1	11	34	35	0	0	43.20988%	56.79012%
3	TA	514	3.402724	274	175	63	2	0	0	0.389105%	99.61089%
4	TA	3	473	0	0	0	0	1	2	100%	0%

A Tabela 26 mostra os grupos criados pelo *k-means* após aplicá-lo nas cinco comunidades. Através da Tabela, cada grupo criado pelo *k-means* tem o número de usuários que pertence a ele, o número médio de respostas fornecidas à comunidade pelos usuários de cada grupo e o número de cada tipo de usuários que cada grupo contém de acordo com a classificação mostrada na Tabela 16.

As siglas utilizadas na Tabela 26 têm os mesmos significados daqueles apresentados na Tabela 22.

Através da Tabela 26 se pode notar que, em geral, quanto maior for a média de respostas fornecidas pelos usuários de um grupo, maior é a probabilidade de encontrar um usuário confiável nele. Este fato pode ser melhor notado através das Figuras 15 e 16, onde as linhas azuis (sólidas) representam a probabilidade de um usuário confiável ou mais do que confiável e a probabilidade de um usuário ser razoavelmente confiável ou menos do que razoavelmente confiável respectivamente. As linhas vermelhas (tracejadas) nas Figuras 15 e 16 mostram linhas de tendência logarítmica (regressões) dos dados apresentados pelas linhas azuis. Por meio das linhas de tendências apresentadas, se conclui que quanto maior for o número médio de respostas de um grupo, maiores são as chances de um usuário ser no mínimo confiável.

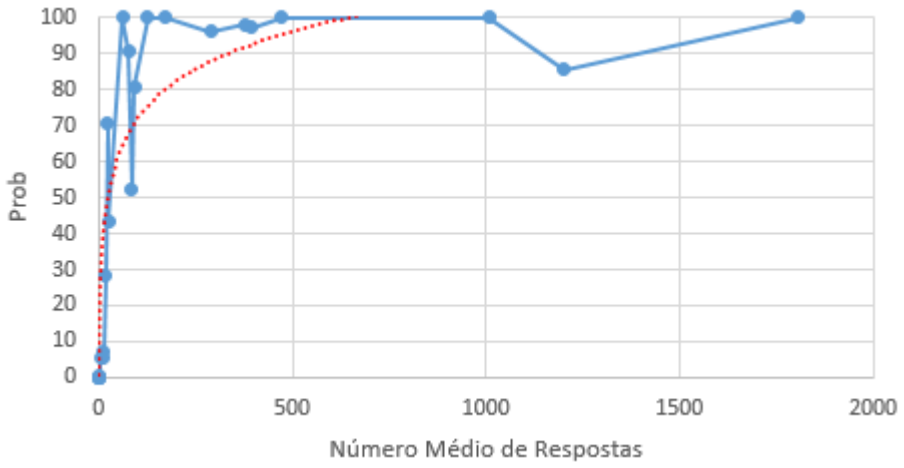

Figura 15. Probabilidade do Usuário Ser Confiável ou Acima

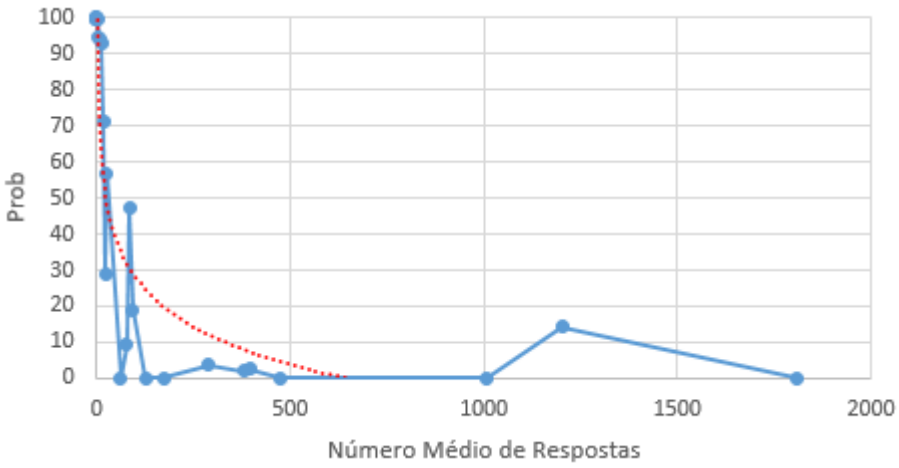

Figura 16. Probabilidade do Usuário Ser Razoavelmente Confiável ou Abaixo

É importante ressaltar que as probabilidades foram calculadas com base na divisão do número de casos favoráveis pelo número de casos possíveis. Por exemplo, no grupo zero da Tabela 26, para calcular a probabilidade de um usuário ser confiável ou acima, primeiramente, foram somados o número de usuários confiáveis, o número de usuários muito confiáveis e o número de usuários extremamente confiáveis. Em seguida, essa soma foi dividida pelo número de usuários para se obter a probabilidade (representada em porcentagem).

5.4. Comentários Finais

Neste capítulo foram apresentados conceitos relacionados a duas técnicas de aprendizado de máquina: uma rede neural artificial e um algoritmo de agrupamento.

A ideia deste capítulo foi mostrar métodos viáveis para encontrar os usuários confiáveis em uma comunidade online. Um dos métodos escolhido foi uma rede neural denominada *Perceptron* Multicamadas e foi utilizado o componente *Core* de cada comunidade como conjunto de dados de treinamento. Os resultados mostraram que o uso de uma rede neural pode ser uma boa solução para encontrar os usuários confiáveis em comunidades. Porém, apesar dos resultados serem animadores, essa abordagem é ainda dependente de um esquema de classificação baseado em um indicador de competência já fornecido (como a reputação) para assim, ser possível configurar a rede neural para uso. Em outras palavras, de acordo com os resultados obtidos, é necessário saber um indicador de competência previamente estabelecido para 20% do componente *Core*. Desta forma, é possível inferir o grau de confiabilidade para os demais usuários da comunidade com uma boa assertividade. Além disto, foi aplicada a validação cruzada na rede neural e, através dos resultados apresentados, pode-se concluir que a rede neural tem capacidade de generalização.

Foram feitos também testes com uma rede neural treinada com dados do *Core* (ou 20% dele) de uma comunidade e, em seguida, essa mesma rede neural foi aplicada a outra comunidade com o objetivo de classificar seus usuários. A abordagem se mostrou promissora, contudo, os melhores resultados obtidos foram de dados provenientes de comunidades um pouco maiores e de tamanhos similares.

O outro método de aprendizado de máquina escolhido para encontrar os usuários confiáveis foi o algoritmo de agrupamento *k-means*. Foram realizados testes utilizando o *k-means* com a finalidade de encontrar grupos similares de usuários. Depois de alguns testes, foi descoberto que quanto maior for o número médio de respostas de um grupo, maiores são as chances de um usuário ser no mínimo confiável. Dado este fato, uma abordagem para saber o grau de confiabilidade de cada usuário dentro de um grupo seria verificar o número de respostas de cada

usuário. Como já demonstrado, o número de respostas de um usuário tem uma correlação positiva com a sua reputação.

6. Conclusão

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

6.1. Comentários Finais e Conclusão

Esta dissertação teve como foco principal realizar um estudo empírico em cinco comunidades online de perguntas e respostas. Esse estudo objetivou encontrar maneiras para encontrar os usuários confiáveis destas comunidades. Contudo, antes do estudo, foram apresentados conceitos fundamentais que nortearam todo o trabalho conduzido.

Primeiramente, foi argumentado que a educação atual não se limita aos moldes tradicionais que já se conhece há anos. Isto significa que, hoje é possível educar e aprender em ambientes mais informais, como as comunidades online, graças à evolução tecnológica. Todavia, mesmo que novas formas para apoiar a educação tenham aparecido, a importância de um professor (tutor, especialista ou uma pessoa mais experiente) em qualquer ambiente onde seja possível aprender, ainda continua sendo desejada. Neste contexto, foram apresentados trabalhos da comunidade científica cujo objetivo era simplesmente encontrar esses possíveis especialistas em comunidades online. Nesta dissertação, esses possíveis especialistas presentes nas comunidades online foram denominados de usuários confiáveis.

Foram também apresentadas discussões sobre como os usuários se tornam confiáveis em uma comunidade online. Foi argumentado que, para saber se alguém é bom em alguma coisa, é necessário submeter pessoas a avaliações. Foi visto que as avaliações são amplamente usadas em escolas e universidades com a finalidade de verificar se um aluno está apto a ser aprovado ou não. De forma similar, algumas comunidades online implementaram mecanismos de avaliações, onde cada usuário pode ser avaliado por outro. Desta forma, se torna possível saber quais usuários das comunidades online fazem contribuições consistentes em algum assunto. A partir desse esquema de avaliações, um usuário pode construir a sua reputação. Essa reputação pode ser positiva caso os usuários tenham boas avaliações e negativa caso tenha avaliações ruins. Essa reputação ou indicador de competência serviu como parâmetro de comparação com análises realizadas nesta dissertação.

Em seguida, iniciou-se o estudo empírico nas cinco comunidades selecionadas para as análises que este trabalho se propôs a fazer com a finalidade de verificar a hipótese apresentada o capítulo 1. Inicialmente foram apresentados como os dados dessas cinco comunidades foram coletados e também foram mostradas algumas características de cada uma das comunidades. Além disso, foi mostrado que neste trabalho as comunidades foram representadas de forma abstratas, através de grafos e de um modelo de classes. A partir dessas representações abstratas, foi possível iniciar uma série de análises de métricas relativas aos usuários dessas comunidades. A ideia dessa análise de métricas era verificar quais delas poderiam ser um indicador que um usuário tem alta reputação.

Neste cenário, foram analisadas algumas métricas dos usuários, como número de respostas, grau de entrada e entropia (foco em assuntos específicos), através de sua correlação com a reputação do usuário. Foi visto que todas as métricas analisadas obtiveram boas correlações (moderada ou forte) com a reputação do usuário. Foi também proposta uma métrica chamada Índice de Confiança com a finalidade de verificar se essa medida pode ser útil para encontrar os usuários confiáveis de uma rede, o que se mostrou uma boa alternativa.

Foi investigada também uma maneira para dividir as comunidades estudadas neste trabalho em partes. As cinco comunidades analisadas neste trabalho foram divididas de acordo com os componentes (partes) da estrutura Bow Tie. Logo após, foram analisados e correlacionados alguns atributos dos usuários das cinco comunidades com as suas respectivas reputações considerando cada componente da estrutura Bow Tie. As melhores correlações foram obtidas no componente *Core* de cada comunidade. Isto pode significar que, para encontrar os usuários confiáveis das comunidades, talvez seja interessante considerar somente os atributos dos usuários no componente *Core* em vez de amostras aleatórias da comunidade.

Foi estudada uma rede neural artificial, denominada *Perceptron* Multicamadas, com o objetivo de verificar se seu uso pode classificar corretamente os usuários das comunidades em confiáveis ou não. Para isto, foram feitos estudos onde foram selecionados os membros do *Core* de cada comunidade analisada para serem os dados de treinamento da rede neural. Os resultados foram animadores e a rede neural se mostrou capaz de classificar qualquer usuário uma vez treinada.

Foram também feitos testes com a rede neural sendo treinada com dados de uma comunidade e, em seguida, aplicada a outra comunidade. A abordagem se mostrou promissora, contudo, os melhores resultados obtidos foram de dados provenientes de comunidades um pouco maiores e de tamanhos similares.

Por fim, foi estudado um algoritmo de agrupamento, o *k-means*, para encontrar os usuários

confiáveis das comunidades. Foi descoberto com os testes que quanto maior for o número médio de respostas de um grupo, maiores são as chances de um usuário ser no mínimo confiável.

Considerando o escopo deste trabalho, pode-se considerar que a hipótese apresentada foi comprovada. Foi visto que a extração de métricas que possam ser um indicador de competência é fundamental para poder classificar um usuário conforme sua confiabilidade. Isso pode ser demonstrado ao longo dos capítulos 3,4 e 5 onde sempre as métricas relacionadas com o indicador de competência estavam presentes em todas as análises. Além disso, os resultados estatísticos apresentados corroboram que através das métricas é possível identificar os usuários confiáveis das comunidades.

Alguns resultados obtidos nesta dissertação foram apresentados em artigos:

- Finding Experts on Facebook Communities: Who Knows More? WORLD SUMMIT ON THE KNOWLEDGE SOCIETY (WSKS), 2014.
- Finding Reliable People in Online Communities of Questions and Answers: Analysis of Metrics and Scope Reduction. International Conference on Enterprise Information Systems (ICEIS), 2014.
- Encontrando Usuários Confiáveis em Comunidades Online de Perguntas e Respostas Através de seu Índice de Confiança. Simpósio Brasileiro de Sistemas de Informação (SBSI), 2014.
- Finding Reliable Users on Online Communities Using Artificial Neural Networks. International Conference WWW/Internet (ICWI), 2014.

Além desses artigos, outros trabalhos extraídos desta dissertação foram aceitos em outras duas conferências, porém, por restrições de tempo estes não puderam ser apresentados. Por conseguinte, estes não serão publicados. Esses trabalhos foram:

- Finding Reliable Users on Online Communities Using Artificial Neural Networks. European Conference On Technology Enhanced Learning (EC-TEL), 2014. Trata-se de uma extensão do trabalho do ICWI, 2014.
- Finding Reliable People in Online Communities of Questions and Answers: Analysis of Metrics and Scope Reduction. Springer Lecture Notes in Business Information Processing Series Book, 2014. Trata-se de uma extensão do trabalho do ICEIS, 2014.

6.2. Contribuições

A principal contribuição dessa pesquisa foi o conjunto de procedimentos apresentados para encontrar os usuários confiáveis de uma comunidade é a contribuição principal deste trabalho. Foi através desses conjuntos de procedimentos é que foi possível comprovar a hipótese apresentada nesta dissertação. Através desses procedimentos são demonstrados como verificar métricas e utilizar algoritmos que podem ajudar a identificar um usuário confiável.

Além disto, as contribuições técnicas ou secundárias da pesquisa são:

- Construção de *scripts* escritos na linguagem de programação Python para a extração de dados das comunidades online estudadas neste trabalho.
- Construções de algoritmos escritos na linguagem de programação Python para tornar possível a realização das análises apresentadas.
- Um levantamento bibliográfico detalhado visando mostrar características de diversos trabalhos já realizados na comunidade científica sobre o tema e também a comparação entre eles.
- Levantamento detalhado de métricas que possam ser um indicador que um usuário é confiável em uma comunidade online.
- Estudo detalhado de diversas características de comunidades online, ressaltando suas semelhanças e diferenças.
- Elaboração de uma análise para encontrar os usuários confiáveis baseado em partes de uma comunidade online.
- Configuração de uma rede neural artificial para classificar os usuários de uma comunidade online.
- Configuração de um algoritmo de agrupamento e definições de estratégias para encontrar grupos com maiores probabilidades de encontrar usuários confiáveis.
- Definição de estratégias para escolher os dados de treinamento durante a configuração da rede neural artificial utilizada neste trabalho.

6.3. Limitações

Essa pesquisa se limitou a análise de somente cinco comunidades. São através dos resultados dessas cinco comunidades é que se derivaram as conclusões aqui apresentadas.

Contudo, sabe-se que análises em mais comunidades são necessárias para que seja possível alcançar resultados mais seguros, tornando assim possível, a elaboração de um método que seja capaz de encontrar os usuários confiáveis em qualquer comunidade online (ou talvez algumas com determinadas características).

Além disso, essa pesquisa se limitou somente em classificar os usuários e, através dessa classificação, identificar os usuários confiáveis. Sabe-se, de fato, que somente identificar os usuários confiáveis não é suficiente. As comunidades analisadas discorrem sobre diversos assuntos de grandes áreas do conhecimento. Contudo, dentro de cada uma dessas grandes áreas, existem subáreas ou especializações sobre o assunto principal abordado na comunidade. Uma abordagem mais ampla seria identificar os usuários confiáveis em assuntos específicos.

Outro ponto de atenção é a consideração da reputação como indicador de competência. Apesar da argumentação apresentada e a apresentação de trabalhos relacionados indicando que um usuário com alta reputação é aquele que fornece boas contribuições para a comunidade, é provável que existam casos onde usuários com contribuições não tão boas tenham alta reputação. Contudo, para a verificação de soluções para o problema que este trabalho endereça, geralmente as pesquisas utilizam alguma forma avaliação dos conteúdos produzidos em comunidades durante os momentos de participação dos usuários, envolvendo muitas vezes uma análise humana externa (como aquela apresentada na seção 3.3 sobre o grupo do Facebook). Sobre essa perspectiva, as reputações construídas nas cinco comunidades estudadas são também oriundas de avaliações, porém, tais avaliações são feitas pelas próprias comunidades.

6.4. Trabalhos Futuros

Um trabalho futuro possível pode ser a realização de análises dentro de cada categoria das comunidades objetivando verificar se é possível encontrar usuários confiáveis em algum assunto. Além disso, o estudo apresentado se limitou a identificar atributos ou estratégias que podem indicar que um usuário é confiável. Um trabalho futuro possível é elaborar um modelo que permita encontrar as pessoas mais adequadas para responder a uma determinada pergunta.

Outro trabalho futuro seria testar outras formas (classificadores) para encontrar os usuários confiáveis, além da rede neural artificial proposta. Uma ideia pode ser utilizar redes bayesianas para encontrar os usuários confiáveis e, em seguida, comparar seus resultados com a rede neural. Infelizmente, por questão de tempo, as redes bayesianas não puderam ser testadas neste trabalho.

Outra ideia interessante sobre uma possível extensão do trabalho seria representar todas as análises realizadas em uma ontologia. Desta forma, para encontrar os usuários confiáveis deveria ser feita inferências lógicas nesta ontologia.

Pensando a nível de ambientes para a aprendizagem, uma opção para trabalhos futuros seria implementar a proposta dessa dissertação em uma plataforma de ensino a distância, para verificar se a abordagem pode ser interessante sobre a perspectiva qualitativa. Através dessa pesquisa se sabe que é possível encontrar os usuários confiáveis de ambientes online. Contudo, se testada a abordagem em um ambiente real de aprendizagem, talvez seja possível responder perguntas como:

- Há ganhos na aprendizagem quando um aluno é apoiado por um usuário confiável em um ambiente online? Em quais cenários?
- Os alunos apoiados por usuários confiáveis tendem a se tornarem confiáveis?
- Em quais são os tópicos os usuários confiáveis são mais solicitados?

Outra opção para extensão do trabalho é buscar uma maneira de não encontrar somente os usuários confiáveis. Muitas vezes, pessoas necessitam de ajudas mais simples como pedir um livro emprestado ou informações sobre horários de aulas. Neste cenário, essas pessoas não necessitam de um especialista para ajudá-la. Uma outra abordagem poderia ser identificar quais são os tipos de ajudas solicitadas em ambientes online e quais são os tipos de usuários capazes para fornecê-la.

Um outro trabalho futuro interessante seria também avaliar o desempenho da rede neural artificial e do algoritmo de agrupamento em uma comunidade online real. Desta forma, será possível descobrir se é viável a implementação de tais técnicas em ambientes reais ou mesmo descobrir quais são os pontos que devem ser otimizados para a viabilização do uso.

Outra questão futura consiste na identificação e o estudo de mais métricas dos usuários. Contudo, a descoberta de novas métricas muitas vezes é dependente da forma de como é representada a comunidade. Desta maneira, buscar por novos modelos (além do grafo e do modelo de classes) que descrevam a comunidade, os usuários e as participações podem ser considerados um trabalho futuro.

Além disso, pode-se considerar a realização de melhorias no Índice de Confiança. Atualmente, o Índice de Confiança está atrelado a data da primeira participação do usuário. Uma ideia de evolução dessa métrica é não considerar a data da primeira participação e sim, datas pré-definidas ao longo do tempo. Essas datas poderiam servir de marcos para recalcular o Índice de Confiança, tornando possível uma análise temporal da métrica ao longo do tempo. Desta forma, será possível verificar se o Índice de Confiança de um usuário varia muito ou pouco com o decorrer do tempo e responder a seguinte pergunta: é mais confiável aquele usuário cujo Índice de Confiança varia menos?

7. Referências

ABASSI, Alireza., HOSSAIN, Liaquat., LEYDESDORFF, Loet. (2012). Betweenness centrality as a driver of preferential attachment in the evolution of research collaboration networks. Journal of Informetrics 6: 403–412. doi:10.1016/j.joi.2012.01.002.

ADAMIC, L., ZHANG J., BAKSHY E. and ACKERMAN, M. S. (2008). Knowledge sharing and yahoo answers: everyone knows something, Proceedings of the 17th international conference on World Wide Web, April 21-25, Beijing, China.

ACKERMAN, M.S., MCDONALD, D.W. (1996). Answer Garden 2: merging organizational memory with collaborative help. In: Proceedings of CSCW '96, Boston, MA, ACM Press, 97-105.

ACKERMAN, M.S., WULF, V., PIPEK, V. (2002). (eds.). Sharing Expertise: Beyond Knowledge Management. MIT Press.

ALAN, Wang G., JIAN, Jiao, ABRAHAMS, ALAN S., FAN, Weiguo, ZHANG, Zhongju. (2013). ExpertRank: A topic-aware expert finding algorithm for online knowledge communities, Decision Support Systems, Volume 54, Issue 3, February, Pages 1442-1451, ISSN 0167-9236, http://dx.doi.org/10.1016/j.dss.2012.12.020.

ANTTILA, J. (2006). Advanced web 2.0 based interactive technology to support informal learning for enhancing quality of business management – a modern approach of information society to knowledge work environment for management. Disponível em: http://www.qualityintegration.biz/Mumbai2006.html.

BADER, D., MADDURI, K. (2006). Parallel algorithms for evaluating centrality indices in real-world networks, in Proc. 35th Int'l Conf. on Parallel Processing (ICPP). Columbus, OH: IEEE Computer Society.

BALOG, K., AZZOPARDI, L., RIJKE, M. D. (2009). A language modeling framework for expert finding. In: Information Processing and Management, 45(1). (2009) 1–19.

BANERJEE, A., BASU, S. (2008). A social query model for decentralized search. Proc. 2nd Workshop on Social Network Mining and Analysis, ACM Press.

BERNERS-LEE, T. et al. (2006). A framework for web science, Foundations and Trens in Web Science, v.1, n.1, p.1-130. Disponível em: http://www.nowpublishers.com/media/Journal-Article-PDFs/1800000001.pdf.

BONACICH P., LLOYD P. (2001). Eigenvector-like measures of centrality for asymmetric relations. Social Networks 23(3):191-201.

BOSU, A., CORLEY, C. S., HEATON, D., CHATTERJI, D., CARVER, J. C, KRAFT, N. A. (2013). Building reputation in stackoverfow: an empirical investigation. In Proceedings of the Tenth International Workshop on Mining Software Repositories, pages 89–92. IEEE Press, 2013.

BRANDES, U., (2001). A faster algorithm for betweenness centrality, J. Mathematical Sociology, vol. 25, no. 2, pp. 163-177.

BRASIL (2008). LEI Nº 11.738, de 16 de Julho de 2008. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11738.htm.

BRODER, A., KUMAR, R., MAGHOUL, F., RAGHAVAN, P., RAJAGOPALAN, S., STATA, R., TOMKINS, A., WIENER, J. (2000). Graph structure in the Web. Computer Networks, 33 (1-6). 309-320.

CAMPBELL, C.S., MAGLIO, P.P., COZZI, A., DOM, B. (2003). Expertise identification using email communications. In: the twelfth international conference on Information and knowledge management, New Orleans, LA. (2003) 528-231.

CARVALHO. R. A., NEVADO, R. A, MENEZES, C. S. (2005). Arquiteturas Pedagógicas para Educação a Distância: Concepções e Suporte Telemático. Anais XVI Simpósio Brasileiro de Informática na Educação. Juiz de Fora – MG.

CASTELLS, M. (1999). A sociedade em rede. São Paulo: Paz e Terra.

CASTRO, A. M., MENEZES, C. (2011). Sistemas Colaborativos para uma nova sociedade e um novo ser humano. In: 1. ed. Rio de Janeiro: Sistemas Colaborativos. Elsevier, v. 1, cap 9.

COMSCORE (2010). Orkut Continues to Lead Brazil's Social Networking Market, Facebook Audience Grows Fivefold. Disponível em: http://www.comscore.com/Insights/Press_Releases/2010/10/Orkut_Continues_to_Lead_Brazil_s_Social_Networking_Market_Facebook_Audience_Grows_Fivefold.

COMSCORE (2011). The Netherlands leads Global Markets in Twitter.com reach. Disponível em: http://www.comscoredatamine.com/2011/02/the-netherlands-leads-global-markets-in-twitter-reach/ .

CUZZOCREA, Alfredo., PAPADIMITRIOU, Alexis., KATSAROS., Dimitrios., MANOPOULUS., Yanis. (2012). Edge betweenness centrality: A novel algorithm for QoS-based

topology control over wireless sensor networks. Journal of Network and Computer Applications 35: 1210–1217. doi:10.1016/j.jnca.2011.06.001.

DANGALCHEV Ch. (2006), Residual Closeness in Networks, Phisica A 365, 556.

DAVENPORT, T., PRUSAK L. (1998). Working Knowledge: How Organizations Manage What They Know. Harvard Business School Press, 1998.

DAVITZ, J., YU, J., BASU, S., GUTELIUS D., HARRIS, A. (2007). iLink: search and routing in social networks. In: Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press. (2007) pp. 931-940.

DENG, Liping, TAVARES, Nicole Judith. (2013). From Moodle to Facebook: Exploring students' motivation and experiences in online communities, Computers & Education, Volume 68, October 2013, Pages 167-176, ISSN 0360-1315, http://dx.doi.org/10.1016/j.compedu.2013.04.028.

DOM, B., EIRON, I., COZZI, A., ZHANG, Y. (2003). Graph-based ranking algorithms for e-mail expertise analysis. In: DMKD, New York, NY. (2003) ACM Press, 42-48.

ENGLISH, R. M., 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.

FAGUNDES, L., C., SATO, L., S., MAÇADA, D., L. (1999). Aprendizes do Futuro – as inovações já começaram. Brasília, MEC.

FISCHER, G. (2001). Communities of interest (CoIs): Learning through the interaction of multiple knowledge systems. In IRIS 24th annual information systems research seminar in Scandinavia, Ulvik, Hardanger Fjord, Norway, August 11–14.

FLEISS, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). In: New York: John Wiley. ISBN 0-471-26370-2.

FRITZEN, EDUARDO, PRATES, JOÃO CARLOS, SIQUEIRA, Sean W. M., Braz, Maria Helena L.B., DE ANDRADE, LEILA C.V. (2013). Contextual web searches in Facebook using learning materials and discussion messages. Computers in Human Behavior, v. 29, p. 386-394.

FÜLLER, J. G., JAWECKI, G., MÜHLBACHER, H. (2007). Innovation creation by online basketball communities. Journal of Business Research, 60, 60–71.

GEISSER, Seymour. (1993). Predictive Inference. New York, NY: Chapman and Hall.

ISBN 0-412-03471-9.

GILBERT, J., MORTON, S., & ROWLEY, J. (2007). E-Learning: The student experience. British Journal of Educational Technology, 38(4), 560–573. doi:10.1111/j.1467-8535.2007.00723.x.

GOMES, A. S. (2012). Educar com o Redu, 1ª Edição, Recife.

GRAIL RESEARCH (2010). Consumers Tomorrow Insight and Observation about Generation Z. Disponível em: http://grailresearch.com/pdf/ContenPodsPdf/Consumers_of_Tomorrow_Insights_and_Observations_About_Generation_Z.pdf.

HERTEL, G., GEISTER, S., KONRADT, U. (2005). Managing virtual teams: A review of current empirical research. Human Resource Management Review, 15, 69–95.

HERTEL, G., NIEDNER, S., HERRMANN, S. (2003). Motivation of software developers in Open Source projects: An Internet-based survey of contributors to the Linux kernel. Research Policy, 32, 1159–1177.

HOLLOWAY, T., BOZICEVIC, M, BORNER, K. (2007). Analyzing and visualizing the semantic coverage of wikipedia and its authors: Research articles. Complexity, 12(3):30–40, 2007.

HOROWITZ, D., KAMVAR, S. (2010). The anatomy of a large-scale social search engine. Proc. of the 19th International Conference on World Wide Web (WWW), ACM Press, 2010, pp. 431-440.

HUBERMAN, B., ROMERO D., WU, F. (2009). Social networks that matter: Twitter under the microscope. First Monday, vol. 14, 2009, pp. 1-8.

JADIN, T., T. GNAMBS, B. BATINIC. (2012). Personality traits and knowledge sharing in online communities. Computers in Human Behavior.

KOHAVI, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. v. 14, p. 1137–1145.

KOLLOCK, P (1999). The economies of online cooperation: gifts and public goods in cyberspace. In Smith, M.A. and Kollock, P. eds. Communities in Cyberspace, Routledge, London, 1999.

KRULWICH, B., BURKEY, C. (1996). ContactFinder agent: answering bulletin board

questions with referrals. In the 13th National Conference on Artificial Intelligence, Portland, OR, 1996, 10-15.

LAKHANI, K., VON HIPPEL, E (2000). How open source software works: "free" user-to-user assistance. Research Policy, 32 (6), 923-943.

LANDIS, J.R., KOCH, G.G. (1977). The measurement of observer agreement for categorical data. In: Biometrics 33 (1): 159–174. doi:10.2307/2529310. JSTOR 2529310. PMID 843571.

LÉVY, Pierre. (1999). Cibercultura. São Paulo. Editora 34.

LIEBOWITZ, J. (2007). Social Networking. The Essence of Innovation. Ed. Rowman & LittleField.

LITTLEPAGE, G.E., MUELLER, A.L. (1997). Recognition and utilization of expertise in problem-solving groups: Expert characteristics and behavior. In: Group Dynamics: Theory, Research, and Practice, 1. 324-328.

LIU, X., WANG, G.A., JOHRI A., ZHOU, M., FAN, W. (2012). Harnessing global expertise: a comparative study of expertise profiling methods for online communities, Information Systems Frontiers 1–13.

LUCENA, C. MACULAN N. (2008). Brazilian Institute for Web Science. Disponível em: http://webscience.org.br/files/INCT_Intro_08_46.pdf.

MACQUEEN, J. B., (1967). Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. University of California Press. pp. 281–297. MR 0214227. Zbl 0214.46201. Retrieved 2009-04-07.

MUI, Y., WHORISKEY, P. (2010) Facebook passes Google as most popular site on the Internet, two measures show. The Washington Post, 2010.

MARCON, K., MACHADO, J. B., CARVALHO, M. J. S., (2012) Arquiteturas Pedagógicas e Redes Sociais: Uma experiência no Facebook, Simpósio Brasileiro de Informática na Educação, Rio de Janeiro

MCCULLOCH, W., PITTS, W. (1943). A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 7:115 - 133.

MEC (2000). Ministério da Educação: Plano Nacional de Educação. Disponível em: http://portal.mec.gov.br/arquivos/pdf/pne.pdf.

MEIRA, S. R. L., COSTA, R. A., JUCÁ, Paulyne Matthews, SILVA, E. M. (2011). Sistemas Colaborativos. 1. ed. Rio de Janeiro: Elsevier, v. 1, cap 4.

MENEZES, C., NEVADO, R., CASTRO JR, A., SANTOS, L. (2008). MORFEU – Multi-Organizador Flexível de Espações Virtuais para Apoiar a Inovação Pedagógica e, EAD. Anais do XIX Simpósio Brasileiro de Informática na Educação. Fortaleza – CE, 2008.

MORRIS, M., TEEVAN, J., PANOVICH, K. (2010). What do people ask their social networks, and why? A survey study of status message Q&A behavior. Proc. 28th International Conference on Human Factors in Computing Systems (CHI), ACM Press, 2010, pp. 1739-1748.

NICOLACI-DA-COSTA, A. M. (2002). Revoluções tecnológicas e transformações subjetivas. Psicologia Teoria e Pesquisa, v.18, n.2, p.193-202.

NICOLACI-DA-COSTA, A. M., PIMENTEL, M. (2011). Sistemas Colaborativos para uma nova sociedade e um novo ser humano. In: 1. ed. Rio de Janeiro: Sistemas Colaborativos. Elsevier, v. 1, cap 1.

NIELSEN. (2009). Global Faces and Networked Places. Disponível em http://www.nielsen.com/content/dam/corporate/us/en/newswire/uploads/2009/03/nielsen_globalfaces_mar09.pdf.

O'HARA, K., HALL, W (2008). Web Science, Association of Learning Technologies Newsletter: Issue 12, May 2008.

PAGE, L., BRIN, S., MOTWANI, R., WINOGRAD, T. (1998). The Pagerank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998.

PAUL, S., HONG L., CHI, E. (2013). Is twitter a good place for asking questions? a characterization study. Proc. Fifth AAAI International Copyright (c) IARIA, 2013. ISBN: 978-1-61208-280-6 152 ICIW 2013: The Eighth International Conference on Internet and Web Applications and Services Conference on Weblogs and Social Media (ICWSM), 2011, pp. 578-581.

PINHATI, Fernando (2013). Plataforma Mignone: uma arquitetura para ambientes virtuais e um modelo para construção de objetos de aprendizagem especializados para educação. Dissertação de Mestrado, Universidade Federal do Estado do Rio de Janeiro, Rio de Janeiro, Brasil.

PRENSKY, M. (2001). Digital natives, digital immigrants, Part II: Do they really think differently? On the Horizon 9(6) (2001) 1–9.

PROCACI, T. B., SIQUEIRA, S. W. M., ANDRADE, L. C. V. (2014a). Finding Experts

on Facebook Communities: Who Knows More? In: WORLD SUMMIT ON THE KNOWLEDGE SOCIETY (WSKS), Venice. Proceedings of the 7th WORLD SUMMIT ON THE KNOWLEDGE SOCIETY. Athens: ORS, 2014. v. 7. p. 1-10.

PROCACI, T. B., SIQUEIRA, S. W. M., ANDRADE, L. C. V. (2014b). Finding Reliable People in Online Communities of Questions and Answers - Analysis of Metrics and Scope Reduction. In: 16th International Conference on Enterprise Information Systems, Lisbon. Proceedings of the 16th International Conference on Enterprise Information Systems. v. 16. p. 526-535.

PROCACI, T. B., SIQUEIRA, S. W. M., ANDRADE, L. C. V. (2014c). Encontrando Usuários Confiáveis em Comunidades Online de Perguntas e Respostas Através de seu Índice de Confiança. In: Simpósio Brasileiro de Sistemas de Informação (SBSI), Londrina. Anais. Londrina: SBC, 2014. v. 10. p. 675-686.

PROCACI, T. B., SIQUEIRA, S. W. M., BRAZ, M.H.L.B., ANDRADE, L. C. V. (2014d). Finding Reliable Users on Online Communities Using Artificial Neural Networks. In: International Conference WWW/Internet (ICWI), Porto. Proceedings, 2014. p. 1-8.

RHEINGOLD, H. (2000). The Virtual Community. Homesteading on the Electronic Frontier. MIT Press.

ROSENBLATT, Frank (1957). The Perceptron. A perceiving and recognizing automaton. Report 85-460-1, Cornell Aeronautical Laboratory.

RÓZEWSKI, P., KUSZTINA, E., TADEUSIEWICZ, R., ZAIKIN, O. (2011). Intelligent Open Learning Systems: Concepts, Models and Algorithms. Springer.

SANTOS, Edméa. O., SILVA, Marco. (2009). Conteúdos de aprendizagem na educação on-line: inspirar-se no hipertexto. Educação & Linguagem, v. 12, p. 124-142.

SCHROER, J., HERTEL, G. (2009). Voluntary engagement in an open web-based encyclopedia: Wikipedians and why they do it. Media Psychology, 12, 96–120.

SINGH, S. (1999). O último teorema de Fermat. São Paulo: Editora Record.

SMITH, S. D., CARUSO, J. B. (2010). The ECAR Study of Undergraduate Students and Information Technology, EDUCAUSE Center for Applied Research, Boulder, CO Research Study, Vol. 6.

SOUZA, C. C. MAGALHÃES, J. J., COSTA, E. B. FECHINE, J. M. (2013). Social

Query: A Query Routing System for Twitter. In: The Eighth International Conference on Internet and Web Applications and Services (ICIW). Roma. Proceedings of the International Conference on Internet and Web Applications and Services.

SPYER, J. (2007). Conectado. Rio de Janeiro: Jorge Zahar.

STREETER, L., LOCHBAUM, K. (1988). Who Knows: A System Based on Automatic Representation of Semantic Structure. In Proceedings of RIAO, 1988, 380-388.

TEEVAN, J, MORRIS, M., PANOVICH, K. (2010). Comparison of information seeking using search engines and social networks. Proc. 4th International AAAI International Conference on Weblogs and Social Media (ICWSM), AAAI Press, 2010, pp. 291-294.

UGULINO, W., MARQUES, A.M., PIMENTEL, M. SIQUEIRA, S. W. M. (2009). Avaliação Colaborativa: um Estudo com a Ferramenta Moodle Workshop. In: II Workshop sobre Avaliação e Acompanhamento da Aprendizagem em Ambientes Virtuais, XX Simpósio Brasileiro de Informática na Educação (SBIE), Florianópolis - SC. Porto Alegre, RS: SBC, 2009. ISSN 2176-4301. 10p.

VASILESCU, B., CAPILUPPI, A., SEREBRENIK, A. (2012). Gender, representation and online participation: A quantitative study of StackOverfow. In Proc. ASE SocialInformatics, IEEE (2012), 332–338.

WASKO, M.S. FARAJ TEIGLAND, R. (2004). Collective action and knowledge contribution in electronic networks of practice, Journal of the Association for Information Systems 5 (11–12) (2004) 494–513.

WASSERMAN, S., FAUST, K. (1994). Social Network Analysis: Methods and Applications. Cambridge University Press, Cambridge.

WEI, C.-T., YOUNG, S. S. C. (2011). Investigating the Role and Potentials of Using Web2.0 in Music Education from Student Perspective, 2011 IEEE 11th International Conference on Advanced Learning Technologies, pp. 344–346.

WIKIPEDIA (2014). K-means clustering. Disponível em: http://en.wikipedia.org/wiki/K-means_clustering.

YIMAM-SEID, D., KOBSA, A. (2003). Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach, Sharing Expertise: Beyond Knowledge Management, MIT Press, Cambridge, MA.

YU, T. K., LU, L. C., LIU, T. F. (2010). Exploring factors that influence knowledge sharing behavior via weblogs. Computers in Human Behavior, 26, 32–41.

ZHANG, J., ACKERMAN, M.S, ADAMIC, L. (2007). Expertise networks in online communities: structure and algorithms, In: Proceedings of the 16th international conference on World Wide Web, May 08-12, Banff, Alberta, Canada.

ZHUNGE, H, ZHANG, J., (2010). Topological Centrality and Its e-Science Applications, J. Am. Soc. for Information Science and Technology, vol. 61, no. 9, pp. 1824-1841.

ZIMMER, MARCO VINICIO (2001). A criação de conhecimento em equipes virtuais: um estudo de caso em empresa do setor de alta tecnologia. Dissertação de Mestrado, Universidade Federal do Rio Grande do Sul, Rio Grande do Sul, Brasil.