Encontrando Usuários Confiáveis em Comunidades Online de Perguntas e Respostas Através de seu Índice de Confiança

Thiago B. Procaci[1], Sean W. M. Siqueira[1], Leila C. Vasconcelos de Andrade[1]

[1]Programa de Pós-Graduação em Informática – Universidade Federal do Estado do Rio de Janeiro (UNIRIO) – Rio de Janeiro – RJ – Brasil

{thiago.procaci, sean, leila}@uniriotec.br

Abstract. *Online communities of questions and answers became important places for users to get information and share knowledge. We investigated a new metric that allows the identification of people that are willing to help and provide good answers in a community, which we call the reliable users. We analyzed three online communities of questions and answers aiming to explore its features and then propose a new metric. Our new metric is called Reliable Index.*

Resumo. *As comunidades online de perguntas e respostas se tornaram lugares importantes para usuários buscarem e compartilharem conhecimentos. Esse trabalho objetiva propor uma nova métrica que permita identificar pessoas que têm maiores chances para dar uma boa resposta para uma pergunta em uma comunidade. No contexto do trabalho, essas pessoas são chamadas de usuários confiáveis. Para isso, foram feitas análises em três comunidades online de perguntas e respostas da Web visando explorar suas características para tornar possível a elaboração da nova métrica. A métrica proposta é chamada de Índice de Confiança.*

1. Introdução

Procurar por soluções ou respostas para problemas pode não ser uma tarefa trivial. Algumas pessoas realizam buscas na Web, contatam amigos ou especialistas conhecidos, porém, em algumas situações, a melhor opção é utilizar as comunidades online de perguntas e respostas para obter o esclarecimento desejado. Contudo, nessas comunidades, pessoas podem não receber respostas ou mesmo receber respostas contraditórias. Neste cenário, esse trabalho tem como finalidade propor uma nova métrica, denominada Índice de Confiança, que permita encontrar os usuários confiáveis de uma comunidade, ou seja, aqueles que têm mais chances de fornecer uma boa resposta para uma determinada pergunta em uma comunidade online.

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 *et al.*, 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. Alan *et al.* (2013) afirmam que as comunidades online destinadas ao compartilhamento de conhecimento são lugares eficientes para se procurar ajuda, pois, em geral, são compostas por indivíduos que compartilham interesses comuns e voluntariamente trabalham para expandir a sua compreensão sobre um domínio do conhecimento. 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].

As comunidades online destinadas ao compartilhamento de conhecimento 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. 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 as pessoas (ou 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].

A ideia desse trabalho é propor uma nova métrica com a finalidade de encontrar usuários confiáveis em comunidades online. Para tornar clara a ideia do trabalho, imagine que um aluno de ciência da computação 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 usando o motor de busca. 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 a suas perguntas. Desta forma, o aluno posta a sua pergunta e aguarda por respostas. A ideia o trabalho é encontrar uma maneira que possa ajudar o aluno do exemplo a encontrar pessoas mais experientes (mais confiáveis).

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 e Basu, 2008]. Esse processo pode ser visto como uma alternativa aos motores de busca. Segundo Horowitz *et al.* (2010), alguns problemas são melhores resolvidos por pessoas, por exemplo, 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. Segundo Fritzen *et al.*, (2013), os resultados dos motores de busca não necessariamente refletem o que se busca em um determinado momento. Huberman *et al.* (2013) e Mui *et al.* (2010) afirmam que ambientes que permitem a formação comunidade 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 a presença de muitos usuários que, por sua vez, aumentam as chances de se receber algum tipo de informação ou resposta.

Todavia, o uso de *social query* 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; nunca receber uma resposta, uma vez que, algumas comunidades tendem a priorizar a visualização das postagens mais recentes [Morris *et al.*, 2013], [Paul *et al.*, 2010].

Diante desse cenário e visando buscar caminhos para minimizar algumas das limitações da *social query* (respostas erradas e ausência de respostas), esse trabalho realiza um estudo em três comunidades online, objetivando propor uma nova métrica, chamada de Índice de Confiança, que permita inferir qual usuário é confiável em uma rede. Essa nova métrica proposta considera o grau de participação do usuário, o seu foco em determinados assuntos e o seu tempo de vida na rede.

Este trabalho está organizado da seguinte forma: a seção 2 será destinada a apresentação de trabalhos relacionados; na seção 3 será feito um estudo empírico em três comunidades online distintas visando caracterizar e explorar as redes. Além disso, será apresentada a nova métrica proposta e também a sua avaliação. Por fim, a seção 4 será destinada a conclusão, trabalhos futuros e comentários finais.

2. Trabalhos Relacionados

Uma alternativa aos motores de busca para a resolução de problemas ou dúvidas são as comunidades online de perguntas e respostas como o *Stackoverflow*, *Quora*, *Yahoo! Answers*, onde os usuários perguntam e respondem de forma voluntária. Todavia, existem pessoas que preferem postar perguntas para pessoas que pertencem somente ao seu círculo de amizades a postar para pessoas desconhecidas em comunidades de perguntas e respostas [Morris *et al.*, 2013].

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. Neste trabalho, foi concluído que 93,5% dos usuários tiveram suas perguntas respondidas e, em 90,1% dos casos, os usuários obtiveram respostas em menos de um dia. Paul *et al.* (2010) fez estudos similares no Twitter, porém, com resultados diferentes. Neste trabalho foi concluído 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). Uma das explicações da baixa porcentagem de respostas recebidas se deve ao do fato Twitter priorizar a visualização de postagens mais recentes. Logo, é provável que alguns seguidores nem fiquem sabendo da existência de uma determinada pergunta.

Estudos para encontrar os usuários confiáveis (*experts*) em comunidades foram explorados em outros trabalhos científicos. 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 [Krulwich e Burkey, 1996] [Ackerman e 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 parser de seus textos produzidos na comunidade e, em seguida, processando-os [Zhang *et al.*, 2007]. Segundo

Littlepage e Mueller (1997) essa abordagem tem se mostrado limitada. Balog *et al.* (2009) propuseram um forma para identificar os *experts* baseado em consultas feitas em um ambiente e uma coleção de textos associados aos candidatos a *experts*. Este trabalho baseado em técnicas de recuperação de informações e métodos probabilísticos visa determinar a relevância entre uma consulta e os candidatos a *experts*. Outro trabalho similar foi proposto por Liu *et al.* (2012), em que 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 é baseada em algoritmos de ranqueamento em grafos para encontrar os usuários *experts* 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) utilizaram o algoritmo de ranqueamento *HITS* em grafos para encontrar os *experts* 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 (adaptação do algoritmo *Page Rank*) e analisaram métricas (número de respostas, por exemplo) 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 necessárias adaptações nas medidas ou nas técnicas utilizadas. Alan *et al.* (2013) propuseram uma nova forma de identificar os experts, construindo um modelo híbrido da abordagem baseada em recuperação de informações com a baseada em algoritmos de ranqueamento em grafos.

Banerjee e 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 (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) propuseram um algoritmo para encontrar os usuários *experts* 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 *experts* no Twitter.

A ideia deste trabalho é revisitar algumas métricas e estratégias já desenvolvidas em outros trabalhos e, em seguida, compará-las com a métrica proposta. Todavia, a métrica proposta neste trabalho é diferenciada dos demais trabalhos porque considera três fatores extraídos das comunidades: o grau de participação do usuário, o seu foco em determinados assuntos e o seu tempo de vida na comunidade.

3. Estudo Empírico nas Comunidades

O objetivo desta seção é mostrar como foi conduzido o estudo empírico nas

comunidades. Nessa seção será mostrada algumas características gerais das comunidades analisadas e também métricas já estudadas em outros trabalhos que podem indicar que um determinado usuário é confiável na rede. Por fim, será apresentada a métrica proposta neste trabalho.

3.1. *Dataset* e Características Gerais das Comunidades

Com a finalidade de testar a proposta desse trabalho, foi necessário extrair um conjunto de dados de comunidades online reais. Para isso, foram escolhidas três comunidades distintas de perguntas e resposta: *Stackoverflow*, que é uma rede destinada assuntos relacionados a programação de computadores; *English Language and Usage*, que é uma comunidade voltada para o aprendizado da língua inglesa; e *Travel Answers*, que é uma comunidade destinada ao esclarecimento de dúvidas sobre viagens.

Em geral, nessas comunidades, 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 verbos etc.) e cada usuário é avaliado por outros usuários baseado em suas perguntas ou respostas postadas. Esse esquema de avaliação permite que os usuários construam a sua reputação na rede, podendo ser positiva ou negativa. Em síntese, as três comunidades são parecidas, porém, com públicos diferentes.

Tabela 1. 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
Stackoveflow	1.000.925	149.269	248.047	603.609	3	270	180.740
English Language and Usage	326.915	30.044	79.978	216.893	6	236	20.408
Travel Answers	42.322	5.529	10.526	26.267	4	275	3.579

A coleta dos dados das comunidades foi através de um *crawler* que consumia dados de cada comunidade. Através desse *crawler*, foi possível consumir todos os dados das comunidades *English Travel and Usage* e *Travel Answers*. Por outro lado, da comunidade *Stackoverflow*, por ser maior que as demais, foi extraída somente uma amostra de seus dados. A Tabela 1 mostra os dados coletados e algumas características gerais das comunidades. Através da Tabela 1, se percebe que a comunidade *Stackoverflow* é a maior das três, seguida da *English Language and Usage* e, por último, a *Travel Answers*. Esse fato pode ser percebido através do número de mensagens, respostas, comentários, *threads* e usuários. As quantidades médias de caracteres escritos nas postagens são bem parecidas nas três comunidades. Já quanto ao tamanho médio de uma *thread*, a comunidade *English Language and Usage* apresenta o maior tamanho. Isso pode significar que nessa comunidade existem discussões mais longas quando comparada com as demais.

3.2. Representação das Comunidades Através de um Grafo

Para realizar algumas análises necessárias neste trabalho, foi preciso representar as comunidades através de um grafo. Zhang *et al.* (2007) propõe o uso de grafo direcionado para representar esse esquema de perguntas e respostas. Nessa representação, os nós do grafo representam os usuários e as arestas representam as interações entre usuários. 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 a pergunta. Do lado direito da figura é mostrado o grafo correspondente a esse esquema de perguntas e respostas.

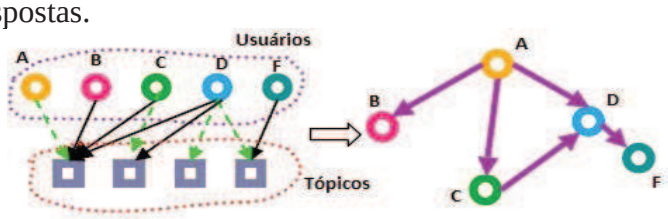

Figura 1. Exemplo de uma comunidade com seu respectivo grafo

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 comunidades. 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 no usuário Z. Desta forma, seguindo este modelo, as comunidades analisadas neste trabalho foram representadas através de grafos, conforme descrito na Tabela 2.

Tabela 2. Dados do Grafo das Comunidades

Comunidade	Número de Nós	Número de Arestas
Stackoverflow	180.740	508.410
English Language and Usage	20.408	149.993
Travel Answers	3.579	16.792

3.3. Atributos dos Usuários

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

- Entropia do usuário: a entropia é uma medida 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 qual a relação do foco do usuário em assuntos específicos e sua reputação. A seção 3.4 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 3.5 explica com mais detalhes esse atributo.
- Grau de entrada: como a comunidade será representada através de um grafo, conforme relatado na seção 3.2, o grau de entrada significa o número de pessoas que um usuário respondeu. Acredita-se que a reputação de um usuário tem relação com a quantidade de pessoas que ele responde 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), porém, se deseja comparar essa abordagem com a proposta. A representação através do grafo é necessária para o seu uso.
- 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 o quão recentes são as participações.

Por fim, todas essas medidas foram extraídas ou calculadas e, em seguida, foram correlacionadas estatisticamente com a reputação de cada usuário fornecida pelas próprias comunidades. Essa reputação é oriunda do esquema de avaliações de perguntas e respostas dos usuários presentes nas comunidades. A ideia é verificar quais métricas representam essa reputação de forma automática. Para as correlações foram selecionados aleatoriamente: 8.000 usuários da comunidade *Stackoverflow*; 8.000 usuários da comunidade *English Language and Usage*; e 2.000 usuários da comunidade *Travel Answers*.

3.4. Entropia do Usuário

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:

$$entropia = -\sum_{i} Pi \, x \log_2(Pi)$$

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. Porém, três das dez respostas foram relacionadas à categoria Java, três relacionadas à categoria arquitetura de computadores e quatro relacionadas à categoria compiladores. 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 Java é 0,3 pois, 3 das 10 respostas postadas foram para a categoria Java (é a divisão 3/10). Desta forma, para calcular a entropia do usuário basta realizar o seguinte cálculo:

$$-((Pjava\, x \log_2(Pjava)) + (Parq\, x \log_2(Parq)) + (Pcomp\, x \log_2(Pcomp)))$$

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

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 cada usuário. Em outras palavras, a reputação de cada usuário foi correlacionada com a sua entropia.

Tabela 3. Coeficiente de Correlação de Pearson (entropia vs reputação)

	Stackoverflow	*English Language and Usage*	*Travel Answers*
Entropia	0,43	0,36	0,44

A Tabela 3 mostra as correlações da entropia com a reputação dos usuários das comunidades. Através dela, pode-se concluir que a entropia se correlaciona moderadamente com a reputação do usuário, quando se analisa as redes. Diante disso, após as análises nas três 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.

3.5. Correlacionando a 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 (confiável), foram extraídas algumas medidas dos usuários (bem como do nó do grafo que o representa) para serem correlacionadas com a reputação dos usuários. As medidas escolhidas para serem extraídas foram: 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.

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 e com o número de respostas de um usuário, uma vez que, responder a muitas questões pode ser um indício que um usuário é um 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:

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

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* nesse trabalho foi usado para identificar os usuários mais relevantes na rede.

Uma vez extraídas todas as medidas das redes, estas foram correlacionadas com a reputação do usuário. A Tabela 4 mostra as correlações dos atributos dos usuários de cada comunidade com a reputação adquirida na rede. Na tabela abaixo, a legenda "Num Resp" significa número de resposta, 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.

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

Atributo	*Stackoverflow*	*English Language and Usage*	*Travel Answers*
Num Resp	0,66	0,92	0,94
Num Com	0,54	0,76	0,83
R + C	0,60	0,82	0,91
z-score	0,58	0,81	0,76
Grau Entrada	0,61	0,88	0,93
Page Rank	0,52	0,86	0,91

Analisando os resultados da Tabela 4, se pode concluir as correlações obtidas nas comunidades menores (*English Language and Usage* e *Travel Answers*) foram fortes (acima de 0,7). Isto significa que os atributos escolhidos, quando apresentarem um alto valor em relação à rede, podem ser fortes indícios de que um usuário é confiável. Já a comunidade *Stackoverflow* (a maior comunidade) apresentou correlações moderadas dos atributos com a reputação (valores entre 0,3 e 0,7). Todavia, se pode considerar que estes atributos podem ser um indício mais fraco de que um usuário é confiável.

3.6. 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 representar melhor 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 do número de comentários, de respostas ou do 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 (conforme mostrado na seção 3.5). 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 assunto somente e um usuário que participam 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) a aquele 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:

$$Índice\ de\ Confiança = \frac{participação \times entropia}{tempo} = \frac{num\ resp \times entropia}{data\ atual - data\ primeira\ participação}$$

Uma vez calculado o Índice de Confiança para os usuários da rede, estes foram correlacionados estatisticamente com a reputação provida pela rede (Tabela 5). Através da Tabela 5, se percebe que o Índice de Confiança pode ser mais útil em comunidades maiores, como no *Stackoverflow*, onde o aumento da correlação foi maior. Na comunidade *Travel Answers* também houve um aumento na correlação, porém, na comunidade *English Language and Usage*, a correlação se manteve constante. Desta 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 5. Coeficiente de Correlação de Pearson (ind. confiança vs reputação)

	Stackoverflow	*English Language and Usage*	*Travel Answers*
Índice de Confiança	0,69	0,92	0,96

4. Conclusão e Trabalhos Futuros

Neste trabalho foi apresentado um estudo em três comunidades distintas visando explorar métricas que possam indicar que um usuário é confiável. 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 de 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 a melhor alternativa para este fim no escopo do trabalho.

Como trabalho futuro se pretende realizar análises dentro de cada categoria da comunidade objetivando verificar se é possível encontrar os especialistas em algum assunto. Além disso, o estudo apresentado se limitou a somente em identificar atributos que podem indicar que um usuário é um confiável. Contudo, somente identificar um usuário confiável não é suficiente para dizer se uma pessoa é adequada para responder a uma determinada pergunta. Por exemplo, considere um usuário especialista em engenharia de software, isto não significa que ele seja a pessoa mais indicada para responder uma pergunta sobre teoria dos compiladores. Logo, somente identificar os confiáveis, não é suficiente para dizer se um usuário é apto a responder com qualidade uma determinada pergunta. Dado esse problema, um trabalho futuro possível é elaborar um modelo que permita encontrar as pessoas mais adequadas para responder a uma determinada pergunta.

Agradecimentos

Este trabalho foi parcialmente financiado pela FAPERJ (projeto E-26-102.256/2013 – Associa: Explorando um Ambiente Semântico e Social de Ensino-Aprendizagem).

Referências

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

Ackerman, M.S., Wulf, V. and Pipek, V. (2002). (eds.). Sharing Expertise: Beyond Knowledge Management. MIT Press, 2002.

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, 2008, Beijing, China.

Alan, Wang G., Jian, Jiao, Abrahams, Alan S., Fan, Weiguo. and Zhang, Zhongju. (2013) ExpertRank: A topic-aware expert finding algorithm for online knowledge communities, Decision Support Systems, Volume 54, Issue 3, February 2013, Pages 1442-1451, ISSN 0167-9236,

Balog, K., Azzopardi, L. and Rijke, M. D. (2009). A language modeling framework for expert finding. Information Processing and Management, 45(1), 1–19.

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

Campbell, C.S., Maglio, P.P., Cozzi, A. and Dom, B. (2003). Expertise identification using email communications. In the twelfth international conference on Information and knowledge management, New Orleans, LA, 2003, 528-231

Davitz, J., Yu, J., Basu, S. Gutelius D. and Harris, A. (2007). iLink: search and routing in social networks. Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2007, pp. 931-940.

Fritzen, Eduardo., Prates, João Carlos., Siqueira, Sean W. M., Braz, Maria Helena L.B. and 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, 2013.

Horowitz, D. and 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. and Wu, F. (2009), Social networks that matter: Twitter under the microscope. First Monday, vol. 14, 2009, pp. 1-8.

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. and Burkey, C. (1996), ContactFinder agent: answering bulletin board questions with referrals. In the 13th National Conference on Artificial Intelligence, Portland, OR, 1996, 10-15

Littlepage, G.E. and Mueller, A.L (1997). Recognition and utilization of expertise in

problem-solving groups: Expert characteristics and behavior. Group Dynamics: Theory, Research, and Practice, 1. 324-328.

Liu, X., Wang, G.A., Johri A., Zhou, M. and Fan, W. (2012). Harnessing global expertise: a comparative study of expertise profiling methods for online communities, Information Systems Frontiers (2012) 1–13.

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

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

Page, L., Brin, S., Motwani, R. and Winograd. (1998), T. The Pagerank Citation Ranking: Bringing Order to the Web, Stanford Digital Library Technologies Project, 1998.

Paul, S., Hong L. and 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.

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), 2013. Roma. Proceedings of the International Conference on Internet and Web Applications and Services.

Teevan, J., Morris, M. and 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.

Wasko, M.S. and 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.

Yimam-Seid, D., and 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, 2003.

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