UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

HENRIQUE DIAS PEREIRA DOS SANTOS

Identificação de Autoridades em Tópicos na Blogosfera Brasileira usando Comentários como Relacionamento

Dissertação apresentada como requisito parcial
para a obtenção do grau de
Mestre em Ciência da Computação

Prof. Dr. Leandro Krug Wives
Orientador

Porto Alegre, Janeiro de 2013

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Henrique Dias Pereira dos Santos,

Identificação de Autoridades em Tópicos na Blogosfera Brasileira usando Comentários como Relacionamento /

Henrique Dias Pereira dos Santos. – Porto Alegre: PPGC da UFRGS, 2013.

56 f.: il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2013. Orientador: Leandro Krug Wives.

1. Autoridade. 2. Blogosfera Brasileira. 3. Análise de Redes Sociais. 4. Ranqueamento. I. Leandro Krug Wives, . II. Título.

”A informação pode nos dizer tudo.
Ela tem todas as respostas.
Mas ela tem respostas para perguntas que não fizemos,
e sem dúvida ainda nem surgiram.”
— BAUDRILLARD, JEAN

AGRADECIMENTOS

Três pesquisadores foram responsáveis pelas principais contribuições ao meu trabalho e a eles faço os principais agradecimentos. Meu orientador na UFRGS, Leandro Krug Wives, que soube conduzir esse trabalho na melhor direção. Pedro Henrique Calais Guerra, doutorando da DCC-UFMG, que foi fundamental nas suas observações ao trabalho em diversas visitas que fiz a ele. Michele Coscia, Pós-Doutorando em Harvard, que me ajudou nas análises de redes complexas sobre os grafos que gerei durante esse trabalho. Discuti meu trabalho também com outros colegas pesquisadores da UFRGS, PUCRS e UFMG, que sempre foram solícitos nas conversas. Meu chefe no CPD-UFRGS, Ricardo Vieira, me deu a oportunidade de continuar meus estudos e sempre me apoiou no mestrado.

Aos meus pais, expresso meu profundo agradecimento, por me darem todas as oportunidades possíveis para cursar a faculdade e consequentemente o mestrado. Sem o ambiente favorável aos estudos que eles sempre criaram, nunca seria possível terminar essa pós-graduação. Meu irmão, Augusto Santos, mestrando na computação da UFRGS, também sempre me ajudou a conduzir as ideias desse trabalho com importantes discussões. Minhas irmãs, Juliana Santos e Ana Helena Ulbrich, mestre em Medicina e Doutora em Química, respectivamente, serviram como audiência às minhas apresentações e ajudaram na melhor exposição dos resultados do trabalho em tabelas e gráficos.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

InWeb	Instituto Nacional de Ciência e Tecnologia para a Web
VSM	Vector Space Model (Modelo Espaço Vetorial)
TF-IDF	Frequência de Termos - Frequência Inversa de Documentos
LDA	Latent Dirichlet Allocation (Alocação Latente de Dirichlet)
ODP	Open Directory Project
KL	Kullback-Leibler
NTOT	Ngram Topic over Time
TOT	Topic over Time
API	Application Program Interface
XML	Extensible Markup Language
JSON	JavaScript Object Notation
BSP	Blog Service Provider
URL	Uniform Resource Locator

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Com o aumento dos usuários acessando a internet no Brasil, cresce a quantidade de conteúdo produzido por brasileiros. Assim se torna importante classificar os melhores autores para que se tenha mais confiança nos textos lidos. Nesse sentido, esta dissertação faz um estudo sobre a descoberta de autoridades em tópicos na blogosfera brasileira. O escopo de estudo e análise é a plataforma de publicação de blogs, Blogspot, sobre os blogueiros que se identificam como brasileiros. Para tanto, foram coletados nove milhões de postagens do ano de 2012 e considerados os comentários como fonte de relacionamento entre os blogueiros para gerar uma rede social. Essa rede foi usada para experimentos do algoritmo de identificação de autoridades em tópicos. O algoritmo utilizado como base é o Topic PageRank, separando os diversos tópicos da blogosfera pelas tags que os usuários definem em suas postagens e posteriormente construindo a lista das autoridades em tais tópicos. Experimentos realizados demonstram que o método proposto resulta em melhor ranqueamento que o algoritmo original do PageRank. Cabe salientar que foi feita uma caracterização dos dados coletados por um questionário aplicado a quatro mil autores.

Palavras-chave: Autoridade, Blogosfera Brasileira, Análise de Redes Sociais, Ranqueamento.

Topical Authority Identification in the Brazilian Blogosphere using Comments as Relationships

ABSTRACT

With the intesification of users accessing the Internet in Brazil, the amount of content produced by Brazilians increases. Thus, it becomes important to classify the best authors to have more confidence in the texts read. In this sense, this work presents a study on subject of topic authorities discovery in the Brazilian blogosphere. The scope of the study is the Blogspot platform, focusing on bloggers who identify themselves as Brazilians. To this end, we collected nine millions posts in the year of 2012 and considered the comments as a source of relationship between bloggers to generate a social network. This network was used for performing experiments considering the proposed approach to identify topic authorities. The algorithm used is based on the Topic PageRank, which can separate the different blogosphere's topics by tags that users use on their posts, and then building the list of authorities on such topics. The experiments conducted show that the proposed approach results in better ranking than the original PageRank algorithm. We also characterize the collected database with a survey of over four thousand authors.

Keywords: Authority, Brazilian Blogosphere, Social Network Analysis, Ranking.

1 INTRODUÇÃO

Diversas redes sociais surgiram rapidamente nos últimos anos, apresentando benefícios diferenciados para diferentes tipos de usuários. Para indivíduos, tais comunidades na Web ajudam a encontrar amigos e a solucionar problemas, permitindo-os compartilhar interesses em comum. Para empresas de publicidade, essas comunidades podem ser exploradas para encontrar o que os usuários estão interessados na intenção de focar seus objetivos (SHEN et al., 2006). Isso torna tais plataformas extremamente interessantes e ricas para a pesquisa.

Há algum tempo os blogs têm se tornado uma mídia social na internet que permite aos usuários, conhecidos como "blogueiros", facilmente publicar conteúdo para sua comunidade. A comunidade de blogs surge de acordo com o comportamento dos blogueiros, visto que cada blogueiro lê e se comunica com os demais. O blogueiro se torna tanto produtor quanto consumidor dos conteúdos gerados na rede de blogs (LIN et al., 2007). Ao conjunto de blogs disponíveis na internet dá se o nome de "blogosfera" (AGARWAL; LIU, 2008). Em comparação com sites tradicionais, a blogosfera possui um estilo diferenciado, o da conversação. Uma conversação geralmente se inicia com alguém que coloca alguma informação, ideia ou opinião nova no seu blog e a espalha para seus amigos, família ou usuários (SONG et al., 2007).

Segundo Ali-hasan e Adamic, a maioria dos blogueiros escreve sobre sua vida cotidiana com pouca audiência de leitores assíduos. Alguns desses leitores interagem com os autores deixando comentários em resposta a uma postagem (i.e., post) específica. Alguns leitores são também autores em seus próprios blogs. Além disso, eles podem listar seus blogs favoritos em seu blog (listagem denominada de *blogroll*) ao lado de seus posts. Também podem criar links para outros blogs e sites dentro de seus posts, referenciando-os através de citações (receber links pelo *blogroll* de outra pessoa, possuir citações e comentários são sinais da popularidade de um blog). Além disso, blogueiros tendem a visitar e criar relações com blogs que compartilham os mesmos interesses (ALI-HASAN; ADAMIC, 2007).

Mesmo existindo um grande número de autores escrevendo conteúdo, apenas uma pequena fração deles concentra a maior parte do tráfego de leitura. A identificação desses indivíduos, aqui chamados de autoridades, é essencial para descobrir qual conteúdo é mais importante e quais indivíduos têm mais influência social na rede. Também é mais eficiente para empresas, clientes e publicitários saberem quem são essas pessoas para utilizar essa influência para entender as vontades dos consumidores, administrarem suas marcas e promover seus produtos (SONG et al., 2007).

Poucos trabalhos como o de Recuero (2008) focam na análise de blogs brasileiros mostrando que existem oportunidades para descobrir informações latentes sobre os dados de blogs e o desenvolvimento de novos algoritmos, especialmente em blogs do Brasil.

Portanto este trabalho visa dar atenção a rede de blogueiros brasileiros, identificando as autoridades em tópicos para o contexto onde o Brasil está inserido.

1.1 Motivação

Na literatura, alguns dos trabalhos que utilizam blogs analisam os conteúdos gerados pelos autores para criar comunidades eminentes (ADAMS; PHUNG; VENKATESH, 2010; SHEN et al., 2006), para realizar a classificação de conteúdo (ADAMIC; GLANCE, 2005; JIANG; ARGAMON, 2008) ou para a definição de autoridades (SONG et al., 2007). Outros utilizaram links entre os blogs para criação de redes sociais (ALI-HASAN; ADAMIC, 2007; LIN et al., 2007) ou para a descoberta de autoridades (PAL; COUNTS, 2011). No entanto, algumas limitações podem ser apontadas nesses trabalhos, dando abertura para novas pesquisas. A maioria dos trabalhos encontrados fez sua análise sobre uma pequena parte da rede social de blog por não ter uma forma automatizada de coletar os dados, construindo algoritmos pouco escaláveis e não aplicáveis num cenário real. Outros problemas estão relacionados com a identificação dos usuários que escrevem e usuários que comentam, dificultando a construção da rede entre os blogs.

Motivado pelo trabalho desenvolvido sobre um grande volume de dados provenientes de microblogs na UFMG (CALAIS GUERRA et al., 2011), a intenção do trabalho aqui apresentado foi fazer uma coleta sobre os blogs brasileiros, obtendo uma grande quantidade de dados para a análise dos relacionamentos entre os blogueiros e identificar autoridades. Para tanto, foi desenvolvido um coletor de blogs para utilizar esses dados na criação de uma rede social entre os blogueiros. Obtendo esses dados, a descoberta das autoridades dentro da rede de blogs brasileiros pode ser mais expressiva.

1.2 Problema

Devido à grande quantidade de conteúdo disponível na internet, não se sabe qual informação é mais confiável ou quais usuários são mais importantes em determinado assunto. Devem ser criados algoritmos para determinar os conteúdos e os autores mais relevantes na internet, tendo como base a sua popularidade ou qualidade de texto.

1.3 Hipótese

Algoritmos de ranqueamento de grafos, como o PageRank (PAGE et al., 1998), podem ser usados para determinar autoridades em redes sociais. É possível determinar autoridades em tópicos utilizando a ideia de Haveliwala (2002) para segmentar o cálculo do PageRank em diversos grafos separados por assuntos diferentes. Essa separação dos tópicos usando os rótulos, *tags*, criados pelos usuários em suas postagens é uma alternativa para segmentação do grafo.

1.4 Objetivos

Esse trabalho visa definir as autoridades em tópicos nas redes sociais com base em blogs com o algoritmo PageRank (PAGE et al., 1998) e fazer análises comparativas com outras formas de ranqueamento de autores.

1.5 Contribuições

Esta dissertação apresenta uma solução para a descoberta de autoridades em tópicos específicos utilizando algoritmo de análise de grafos sobre a base de dados de blogueiros brasileiros. A técnica utilizada como base para o ranqueamento dos autores é o algoritmo de PageRank (PAGE et al., 1998) aplicado sobre os grafos de cada tópico, como sugere Haveliwala (2002).

As principais contribuições deste trabalho são:

- Uma grande base de dados envolvendo nove milhões de postagens sobre a Blogosfera brasileira que pode ser usada em outras pesquisas;
- Descoberta de autoridades utilizando o Topic PageRank segmentando os tópicos pelas tags das postagens e os comentários dos blogs como relacionamento entre os usuários;
- Uma pesquisa realizada sobre quatro mil autores de blogs para ajudar a caracterizar a base de dados coletada e entender melhor o perfil dos usuários.

A técnica proposta foi aplicada à base coletada, a qual contém postagens, conteúdo e comentários coletados da plataforma Blogspot[1].

1.6 Estrutura do Texto

No próximo capítulo serão apresentados alguns conceitos importantes para a compreensão das estruturas básicas da dissertação como blogs, Redes Sociais e Autoridade. No capítulo **??** serão abordados trabalhos que de alguma forma utilizaram os dados de blogs como redes sociais para extrair informações mais ricas que somente as postagens, além de outros trabalhos relacionados à descoberta de autoridades. Após, no capítulo 3, é apresentado o método utilizado para descobrir as autoridades e como os dados foram coletados. Os resultados, experimentos e sua respectiva validação são discutidos no capítulo 4, demonstrando quais tópicos foram utilizados e as melhorias obtidas. Por fim, a dissertação termina com as conclusões e discussões de trabalhos futuros. Os apêndices mostram mais informações sobre a infraestrutura utilizada, experimentos complementares, o exemplo de um grafo e os questionários aplicados aos usuários.

[1] http://blogspot.com/

2 REFERENCIAL TEÓRICO E TRABALHOS RELACIONADOS

Este capítulo apresenta conceitos importantes que são empregados na descoberta de autoridades em redes sociais referenciados ao longo da dissertação. Na seção 2.1 são descritos os conceitos gerais sobre blogs e depois, na seção 2.2, são descritas as redes sociais. Em seguida, na seção 2.3, são abordados conceitos relacionados à definição de autoridade. Na seção 2.4 são apresentados os principais trabalhos que consideraram a Blogosfera como uma rede social e depois, na seção sec:autoridades, os trabalhos que procuram identificar autoridades em blogs.

2.1 Blogs

Blog, abreviatura de web log, é um diário de textos, imagens e outras mídias, compartilhando novidades ou conteúdos encontrados em páginas da Web. Um blog consiste em um título, uma informação de assinatura e em múltiplas postagens que são organizadas de forma descendente com relação à sua data de publicação. Em geral, a postagem de um blog é combinada com sua data de publicação, um texto sobre algum assunto, *hyperlinks*, imagens e comentários. O usuário que escreve em um blog, o seu autor, é chamado de blogueiro. Em um blog, outros blogueiros podem comentar em uma postagem. Além disso, blogueiros podem acompanhar outros blogs, significando que eles têm interesse no tópico desses blogs. Blogueiros também podem adicionar blogs na sua lista de favoritos, conhecida como *blogroll*, que é listada na capa do blog, geralmente em uma lista ao lado, indicando os links que o blogueiro mais gosta.

A Figura 2.1 mostra os tipos de interações que podem haver entre os blogs. Os comentários feitos pelos usuários mostram que eles têm interesse pelo tópico discutido no texto. Também existem os *trackbacks*, que são links de retorno identificando que um blog citou uma postagem de outro blog. Além disso, os blogueiros podem listar seus blogs favoritos numa barra lateral às postagens que indicam blogs que este autor segue ou lê. Essas são as interações mais comuns entre os blogs (LIN; TANG; KAO, 2009). Se um usuário pesquisador desejar fazer mineração em blogs, esses são os tipos de dados que podem ser explorados (i.e., o conteúdo e os relacionamentos entre blogueiros).

De acordo com Adams, Phung e Venkatesh (2010), os blogs são descentralizados e dinâmicos, frequentemente correspondem a eventos do mundo real noticiados pela mídia convencional. Geralmente são construídos em uma rede confiável de relacionamento, que cria uma atmosfera de responsabilidade, tornando a Blogosfera em um bom termômetro e propagadoras de opiniões entre os usuários. Esse fato cria um grande interesse comercial de grandes empresas de mídia e empresas de pesquisa (como Google, Yahoo!, etc.)

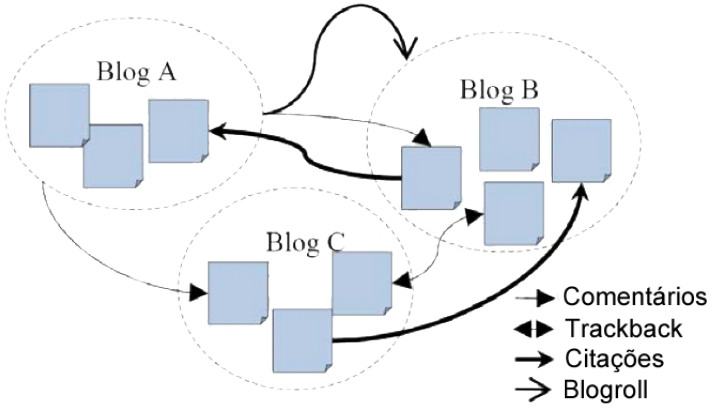


Figura 2.1: Interações existentes entre os blogs.
Fonte: imagem traduzida e adaptada de (LI; LAI; CHEN, 2009, p. 411)

aos quais suas aplicações e produtos podem receber retorno de satisfação e servir para realização de marketing viral sobre os blogs (ADAMS; PHUNG; VENKATESH, 2010).

Adams, Phung e Venkatesh (2010) também descrevem que os blogs são mais do que simplesmente um conjunto de textos. Eles são uma ferramenta tecnológica para a manifestação de interação social. A maioria dos blogs está em uma rede de relacionamento entre os blogueiros, pessoas que comentam suas postagens. As ligações que compõem essa rede são dinamicamente envolvidas e têm nuances semânticas (ADAMS; PHUNG; VENKATESH, 2010).

Com isso, fica fácil entender como a rede de blogs pode ser considerada uma rede social, como será visto na próxima seção.

2.2 Redes Sociais

As redes sociais são espaços para criar, construir e manter relações pessoais ou profissionais entre pessoas, encontrar oportunidades e aprender novas ideias. A palavra social entra quando essas interações se dão em relações que já existem no mundo real. Obtendo uma rede de contatos e colocando-a no mundo virtual é criada uma comunidade on-line. Existem redes sociais especiais (p.ex., os blogs) que contêm simplesmente sentimentos e atividades que seus autores executam no dia-a-dia.

A Figura 2.2 mostra um exemplo de rede social criada a partir de informações obtidas em blogs. Os nodos azuis representam uma das redes do Kuwait analisadas e os nodos vermelhos outra rede. O tamanho dos nodos é proporcional ao número de arestas de entrada. As informações são provenientes de duas comunidades de blogueiros, azul e vermelho, ligados pelas interações geradas por seus comentários, citações ou *blogroll*, as arestas do grafo. Nessa figura já é possível perceber que existem blogueiros que têm mais influência na rede do que os demais, os nodos maiores, que receberam mais links.

2.3 Autoridades

Autoridades são usuários que têm grande influência dentro da rede social que pertencem. Essa definição também pode ser usada para os autores populares. Essa influência pode ser identificada pelo alto grau de relações entre um usuário e os demais ou analisando o quanto do texto gerado por um usuário é posteriormente utilizado pelos demais

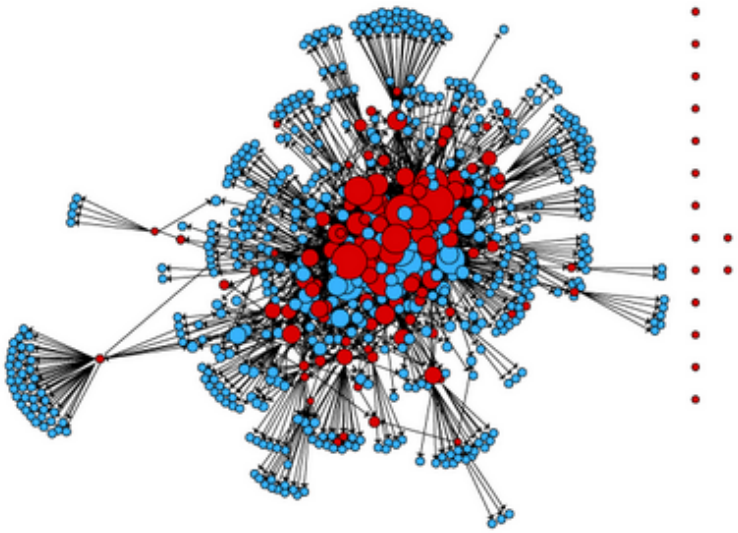


Figura 2.2: Exemplo de uma rede social de blogs.
Fonte: (ALI-HASAN; ADAMIC, 2007, p. 6)

(PAL; COUNTS, 2011).

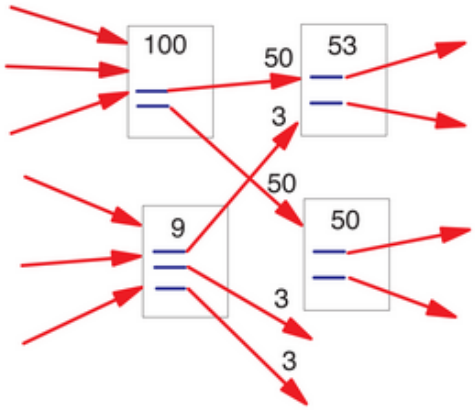


Figura 2.3: Cálculo simplificado do PageRank.Fonte: (PAGE et al., 1998, p. 4)

A Figura 2.3 mostra uma forma simplificada de fazer um cálculo de autoridade utilizando o algoritmo PageRank. O algoritmo dá valor de autoridade aos usuários de acordo com o valor dos usuários que os citam e quantas citações eles têm (PAGE et al., 1998). No exemplo, o usuário com valor 100 cita somente dois usuários, dando o valor de 50 para cada um deles.

A ideia básica do PageRank é que se uma página u tem um link para a página v, então o autor de u dá certa importância para a página v. Intuitivamente, Yahoo! é uma página importante, pois várias páginas apontam para ele. Assim, páginas apontadas pela página do Yahoo! provavelmente serão importantes. Já a importância que uma página u dá aos seus links é medida da seguinte forma: seja Nu o grau de saída de uma página u e considerando que $Rank(p)$ represente a importância (i.e., PageRank) de uma página p. Então o link (u, v) será $Rank(u)/Nu$ unidades para v. Essa ideia rege para as iterações seguintes, que geram o vetor de ranqueamento **Rank*** sobre todas as páginas da Web. Se n é o número de paginas, cada página terá seu valor inicial como $1/n$. Seja Bu o conjunto

de páginas apontando para v. Em cada iteração, propaga-se o ranqueamento como abaixo:

$$PR(u) = c \sum_{v \in B_u} \frac{PR(v)}{N_v} \tag{2.1}$$

Na equação, u é uma página da Web. F_u é o conjunto de páginas que u aponta e B_u o conjunto de páginas que apontam para u. $N_u = |F_u|$ é o número de links saindo de u e c o fator usado para a normalização (assim o total dos ranks de todas as páginas é uma constante). Então PR é definido como o *ranking* simples das páginas na Web e a Equação 2.1 como uma versão simplificada do PageRank original (PAGE et al., 1998).

As iterações continuam até que o **Rank** estabilize até certo limiar. O vetor final **Rank*** contém o vetor PageRank de toda a Web. Esse vetor é computado uma única vez depois de coletar a Web, e os valores podem então serem usados para influenciar o *ranking* de uma máquina de busca. O algoritmo original do PageRank foi desenvolvido para o ranqueamento da popularidade das páginas da Web. A estrutura de links da blogosfera é similar às páginas da Web, mas com características adicionais como comentários e listas de blogs favoritos, podendo enriquecer o algoritmo.

2.3.1 PageRank sensível ao Tópico

O trabalho realizado por Haveliwala (2002) aprimora o uso do PageRank para tópicos e sugere a computação de diversos vetores de PageRank para as páginas, cada vetor representando seu valor de PageRank para cada tópico definido. Esses vetores de tópicos serão usados para influenciar o ranqueamento dos resultados em uma máquina de busca.

Cada página é relacionada com um tópico utilizando o Open Directory Project[1] para rotulá-las. Foram definidos 16 tópicos caracterizando 16 vetores de PageRank para cada página. Os vetores são pré-computados por 16 matrizes de adjacência entre as páginas em um algoritmo de programação dinâmica que reutiliza a computação das outras matrizes para facilitar o processo. O cálculo do PageRank original levou 5 horas em um *dataset* de 80 milhões de URLs e 20 horas para os cálculos de 16 vetores no mesmo *dataset*.

Os resultados foram comparados com a versão original do algoritmo e avaliados por um grupo de usuários. A lista de páginas retornadas nos diversos experimentos mostrou que o uso de vetores de tópicos para a máquina de busca traz resultados mais significativos que utilizando somente um vetor genérico para o PageRank. Essa abordagem demonstra que o uso do algoritmo PageRank para tópicos é válido e melhora os resultados. Entretanto, nesse estudo, o autor necessita que as páginas sejam previamente rotuladas.

No caso desta dissertação, este será o algoritmo utilizado para o cálculo de PageRank por tópicos, mas o rótulo dado para cada documento (i.e. postagens) é feito utilizando as tags dadas pelos usuários no próprio documento, não necessitando de fontes externas.

2.4 Blogs como Redes Sociais

Nesta seção são descritos alguns trabalhos encontrados na literatura que estruturam os blogs como uma rede social, utilizando as interações dos blogueiros como ligações entre os usuários.

Um trabalho interessante foi realizado por Calais Guerra (2011), na Universidade Federal de Minas Gerais, estruturando a plataforma de microblogs, Twitter[2] como um grafo

[1] http://www.dmoz.org/

[2] http://twitter.com

social. Nesse trabalho os *retweets*, mensagens replicadas de outros usuários, são usados como sinal de endosso entre usuários, formando um grande grafo direcionado e conectado. Esse grafo é utilizado para dar polaridade aos usuários, tendo como base perfis de polaridade conhecida e os *retweets* entre os usuários. Essa polaridade é empregada na transferência de aprendizado de palavras novas que surgem no contexto e para identificar qual é a polaridade delas. Essa abordagem também pode ser aplicada à blogosfera para descoberta de endosso pelos comentários dos usuários, aumentando os laços entre os blogueiros, entretanto essa não é o foco desta dissertação.

Na literatura, também é comum encontrar trabalhos que consideram a rede de co-autorias sobre artigos científicos e constroem uma rede social. No trabalho de Chan, Pon e Cardenas (2006), por exemplo, os autores de artigos são modelados como os vértices de um grafo e as arestas são as co-autorias ou citações dos artigos. Então, usando um algoritmo de agrupamento de grafo, é possível encontrar subcomunidades dentro de uma área de pesquisa. Nos blogs, entretanto, existem outros tipos de laços para serem explorados entre os usuários para descobrir subcomunidades. Algoritmos de agrupamento de grafo podem ser usados para encontrar subcomunidades em blogs e para a descoberta de autoridades.

2.4.1 Mineração de amigos latentes em dados de blogs

O crescimento dos blogs como ferramenta de produção textual tem gerado um recurso rico para mineração de comunidades sociais na Web. No trabalho realizado por Shen (2006), os dados de interações nos blogs são usados para resolver o problema de mineração de amigos latentes. Um amigo latentes é definido como alguém que compartilha uma distribuição de tópicos semelhante ao usuário em questão. Essas pessoas podem não se conhecer, mas compartilham interesses dentro da blogosfera.

Três abordagens são desenvolvidas para a detecção de amigos latentes. A primeira consiste em calcular a similaridade do cosseno sobre o conteúdo dos autores. Na segunda é usado o modelo de tópicos LDA (*Latent Dirichlet Allocation*) (BLEI; NG; JORDAN, 2003) sobre o conteúdo para encontrar a similaridade em nível de tópico. E a terceira abordagem é um método híbrido de similaridade baseado nas duas etapas combinadas. Dado um blog, depois de calcular a similaridade entre ele e todos os outros blogueiros, pode-se ordenar os autores de acordo com sua similaridade e obter a lista daqueles que compartilham os mesmos interesses. Essa abordagem é a base de cálculo para os outros métodos propostos.

No trabalho de Shen (2006), foi utilizado um banco de dados com 153 mil páginas de tópicos com distribuição em 74 categorias sobre 10 mil autores e suas postagens.

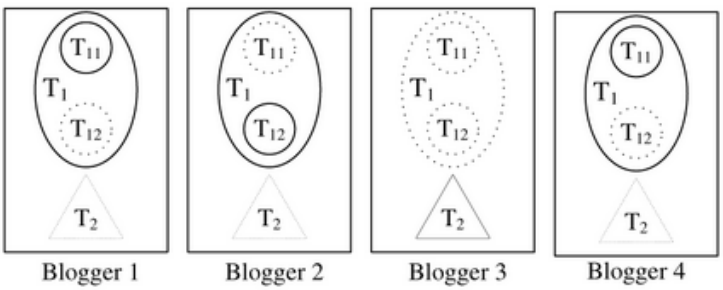


Figura 2.4: Ilustração das vantagens de cada abordagem.Fonte: (SHEN et al., 2006, p. 8)

O método de duas etapas supera todas as desvantagens dos demais métodos, primeiro

calculando a similaridade dos tópicos e depois aplicando uma similaridade fina entre os autores. Por exemplo, na Figura 2.4 para o blogueiro 1 (*Blogger 1*), o método de duas etapas pode retirar o blogueiro 3 (*Blogger 3*) usando a similaridade de tópicos e então colocar o blogueiro 2 (*Blogger 2*) depois o blogueiro 4 (*Blogger 4*), no ordenamento de amigos latentes, de acordo com a similaridade fina.

As vantagens de um método de duas etapas são: A proposta tem como objetivo diminuir a complexidade e tempo de computação; A estratégia é fácil de ser estendida para encontrar amigos com os mesmos interesses.

Shen (2006) propôs explorar o conteúdo dos blogs para descobrir amigos latentes entre os autores de blogs. Os experimentos que ele realizou mostraram que o método de duas etapas é eficiente na solução deste problema. Esse trabalho mostra que é possível, inserindo a complexidade de analise de texto em tópicos, criar comunidades de usuários relacionados com um tema específica. Depois de criada essa comunidade, pode-se então fazer a identificação de autoridade.

2.4.2 Analisando Redes Sociais em Blogs através de Links e Comentários

Os blogs são muito usados como diários, permitindo às pessoas a falar sobre suas vidas cotidianas. Os leitores interagem com os autores contribuindo com comentários em resposta a uma postagem específica. O trabalho de Ali-Hasan e Adamic (2007) apresenta um estudo comparando as amizades existentes entre os usuários na vida real e na blogosfera. Três comunidades foram analisadas: Kuwait Blogs, Dallas/Fort Worth Blogs e United Arab Emirates Blogs.

As ligações analisadas nessas comunidades são as seguintes: *Blogroll* é a lista de blogs que o autor na barra lateral e aponta para outros blogs no qual ele lê ou segue; Citações são links no texto de um usuário onde leva para outro blog, geralmente com conteúdo similar ao escrito; Os comentários são interações entre o leitor e o autor da postagem onde o usuário pode expressar sua opinião sobre o texto escrito, essa é a interação mais interessante pois garante que o usuário leu o conteúdo, se interessou por ele e deixou sua opinião.

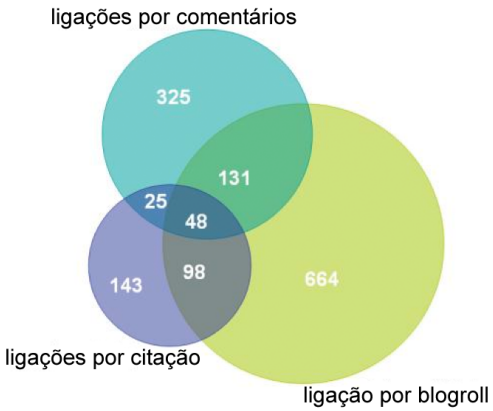


Figura 2.5: Sobreposição das ligações entre blogueiros usando diagrama de Venn.
Fonte: (ALI-HASAN; ADAMIC, 2007, p. 3)

Embora fosse esperado que os usuários deixassem comentários nos blogs que eles seguem, a Figura 2.5 mostra que nem sempre isso acontece. Muitos comentários e citações são feitos em blogs que o usuário não listou na lista de blogs favoritos.

Tabela 2.1: Reciprocidade entre as ligações existentes

	Kuwait	**UAE**	**DFW**
citações	19%	16%	26%
blogroll links	32%	43%	27%
comentários	43%		

Fonte: (ALI-HASAN; ADAMIC, 2007, p. 5)

Uma importante métrica em redes sociais é a reciprocidade entre os usuários, mostrando que a rede tem relações simétricas. Na Tabela 2.1 é mostrada a reciprocidade nas redes analisadas pelos autores e a maior relação reflexiva está nos comentários da rede Kuwait e na *blogroll* da rede UAE.

Na análise dos autores sobre as relações on-line e off-line foram pesquisados 87 blogs da rede Kuwait, 38 da UAE e 67 da DWF, mas somente alguns autores responderam à pesquisa: 63%, 68% e 23% das comunidades citadas, respectivamente. Quando perguntados sobre os comentários recebidos em seus blogs, a maioria dos autores respondeu que os leitores não são indivíduos que eles conhecem pessoalmente. Além disso, os próprios autores, em sua maioria, deixam comentários em blogs de autores que eles desconhecem pessoalmente.

Além das relações criadas on-line, também foi analisado se a blogosfera ajuda a manter as relações off-line. Nesse caso, 99% dos autores da comunidade de blogs do Kuwait, 94% do UAE e 79% do DFW responderam que praticamente nenhum comentário é feito por seus conhecidos que não tem blogs. Isso mostra que os blogs não contribuem para criar relações no mundo reais. Os autores também responderam que os blogs listados nos favoritos não são de conhecidos da vida real. Um comportamento interessante analisado na comunidade do Kuwait é a conversação entre os autores e leitores. Existem dois tipos de usuário: os incitadores, que iniciam a discussão de um assunto, e os apoiadores, que continuam essa discussão nos comentários da postagem.

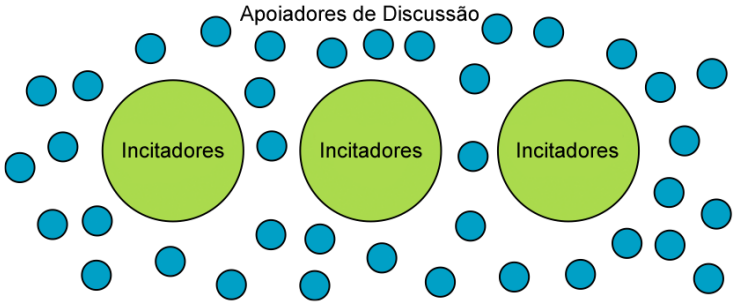


Figura 2.6: Regras sociais na blogosfera do Kuwait.
Fonte: traduzido e adaptado de (ALI-HASAN; ADAMIC, 2007, p. 8)

Como mostra a Figura 2.6, os incitadores têm muitas interações de comentário por seus apoiadores de discussão. O nodo mais central da rede recebeu 346 comentários de 100 usuários diferentes e 24 destes de dentro da comunidade. Esse autor também deixou 89 comentários em 12 blogs diferentes. A maior contribuição do artigo é que, embora a maioria dos trabalhos sobre interações de blogs deem ênfase no *blogroll* e citações, as análises mostram que essas ligações nem sempre se refletem nos comentários. A pesquisa feita sobre os autores de blogs mostra que as relações on-line não têm muita paridade

com as off-line, na vida real. Essas observações nas redes de blogueiros servem para entender melhor as interações entre eles e refletir melhor em como podem ser utilizadas na identificação de autoridades.

2.4.3 Descobrindo subcomunidades latentes de leitores de Blogs

O crescente número de publicações em blogs e de interações entre leitores cria a necessidade de agrupar esses usuários definindo suas preferências e interesses. Nessa linha, Adams e Phung (2010) propõem um método para identificar subcomunidades em tópicos entre os leitores, caracterizando os indivíduos e modelando sua leitura por meio da técnica de LDA (Latent Dirichlet Allocation) e de NTOT (*Ngram Topic over Time*)[3].

A ideia do algoritmo de Adams e Phung consiste em agrupar os usuários que comentam tendo como base o conteúdo das postagens que os induzem a comentar. O repositório de blogs é considerado como documentos e pode-se representar um histograma de palavras por usuário que comenta sobre os termos utilizados nos textos comentados. Os termos são reduzidos pela NTOT, que tem os benefícios de gerar um espaço vetorial compacto e descritivo.

Considerando os comentários como um tipo de relacionamento entre os autores e usuários e definindo que a similaridade entre os usuários que comentam é baseada nos tópicos definidos pelas postagens que eles comentam, é possível descobrir subcomunidades por afinidade de tópicos. Então uma comunidade é definida pelos interesses dos usuários. Existem várias razões para um usuário deixar um comentário: ele pode estar interessado no tópico da postagem, ele pode querer argumentar ou criticar o seu conteúdo ou pode simplesmente querer agradecer seu autor.

Os comentários servem como moeda de troca para os autores, pois mostram que seus textos estão sendo lidos por alguém. Adams e Phung (2010) consideram que se duas pessoas comentam em um mesmo tópico, eles são similares de alguma forma.

Assumindo que existem M usuários e K tópicos, é construída uma matriz TMxK onde cada entrada representa o grau de interesse do usuário x no tópico k. O par de similaridades entre usuários, S(x,y), é dado pelos interesses dos usuários nos tópicos T(x) e T(y) utilizando a distância de vetores. Essa similaridade provê uma detecção básica de comunidades latentes entre os leitores de blogs. A descoberta de comunidades é tratada como um problema de agrupamento definido por duas restrições: o número de grupos é desconhecido e o espaço de grupo não tem métricas.

No que diz respeito ao reconhecimento de grupos similares de usuários, não existe um padrão simples de unificação a ser usado. O primeiro conjunto de dados utilizado pelo artigo tem 12 blogs (3250 postagens e 6500 comentários), onde os usuários autores deram retorno sobre o resultado das comunidades descobertas. Esse conjunto ofereceu respostas úteis para os autores e leitores, considerando os grupos recomendados sobre os tópicos. Segundo os autores, o algoritmo foi testado em um conjunto de dados com um número bem maior de comentários e n-gramas, e também ofereceu bom resultados, provando ser um algoritmo com boa escalabilidade.

A contribuição se dá em duas frentes: a descoberta de subcomunidades dá ao autor a possibilidade de entender como seu texto é consumido pelos usuários e provê uma base para uma nova modalidade de visualização e navegação sobre a blogosfera, dando uma perspectiva mais social, recomendando aos leitores conteúdo de seu interesse. Novamente, se antes de identificar a autoridade há a necessidade de classificar as comunidades

[3]Tanto o LDA quanto NTOT são técnicas para identificar tópicos em documentos.

nos tópicos definidos pelos textos, o NTOT mostra ser uma alternativa interessante sobre o LDA.

2.5 Identificação de Autoridades em Blogs

Nesta seção serão abordadas pesquisas que propuseram algoritmos de descoberta de autoridades em blogs. Os trabalhos mais importantes serão detalhados nas subseções seguintes.

Um trabalho interessante analisando autoridades em tópicos na blogosfera foi o realizado por Liu (2011). O autor discute o desafio de identificar de forma eficiente autoridades em tópicos e propõe um novo modelo para quantificar as autoridades. Além disso, o trabalho apresenta uma nova abordagem para identificar comunidades de blogs relacionadas a determinado tópico. Experimentos demonstram que essa abordagem pode extrair autoridades em tópicos e comunidades de forma eficiente.

2.5.1 Utilizando Relações Sociais para Mineração da Popularidade de Blogs

O trabalho de Lin, Tang e Kao (2009), desenvolvido na Universidade Nacional de Cheng Kung, Taiwan, é o principal trabalho na descoberta de blogs populares utilizando todas as ligações que uma ferramenta de blog permite criar entre os diversos blogs. Esse trabalho cria um algoritmo similar ao PageRank, chamado BRank, que pondera cada ligação entre os blogs (comentários, citações, lista de blogs e *trackbacks*) para medir a popularidade de um blog. O método se baseia não só nas características dos blogs, que inclui a estrutura de links, mas também outros em aspectos de similaridade entre blogs.

Dois tipos de relacionamentos sociais entre blogs são definidos no modelo de rede de blogs proposto: relações de suporte e similaridade. Interações ou ligações do blog A para o blog B indicam que o autor de A é leitor de B, e essas relações são chamadas de relações de suporte. Existem quatro tipos de comportamentos de interações: comentários, *trackbacks*, *blogroll* e citação, e todas são relacionadas com relações de suporte. Links em comum entre o blog A e o blog B definem as chamadas relações de similaridade entre A e B. Se o blog A e o blog B são similares, existirá uma probabilidade dos leitores de A lerem B.

O BRank é uma modificação do PageRank levando em conta as características sociais que a rede de blogs tem. A grande modificação é dada pela probabilidade de um leitor do blog A ir para o blog B (PAB). Essa probabilidade leva em consideração o produto de três fatores de cada relacionamento entre o blog A e B: o peso do relacionamento, visto que cada um tem um significado diferente para o blogueiro, o grau para expressar a força do relacionamento, e, por último, a qualidade do blog, que é expressa por uma combinação de atributos. Esses atributos se referem à atividade do blog, tamanho médio dos posts, quantidade de postagens, comentários e *trackbacks*, pois se considera que blogs com alta qualidade terão mais chance de receberem leitores. A equação generalizada do BRank é:

$$BRank(A) = \frac{1-d}{n} + d * \sum_{X \in S(A)} BRank(X) * P_{X \to A'} \tag{2.2}$$

onde d é o fator de amortecimento definido pelo PageRank e $S(A)$ é o conjunto de blogs que se relacionam com A.

Os experimentos focaram em cinco BSPs (Blog Service Provider) em Taiwan: Wretch Blogs, Yahoo Blogs, Yam Blogs, Xuite Blogs e Pixnet Blogs. O processo de coleta come-

çou sobre os blogs mais populares entre setembro de 2007 e maio de 2008. As estatísticas do conjunto de dados estão na Tabela 2.2:

Tabela 2.2: Informações dos blogs e suas interações

	Blog	**Posts**	**Comentários**	**TrackBack**	**Citações**	**Blogroll**
Wretch	592.123	6.880.087	16.527.101	316.263	154.190	236.168
Yahoo	294.352	727.335	1.589.940	137.232	253.928	110.837
Yam	84.536	1.895.319	2.318.052	104.594	65.125	15.583
Xuite	27.320	1.270.830	822.398	21.053	325.854	12.791
Pixnet	41.507	2.511.188	4.356.075	14.336	57.504	27.602

Fonte: (LIN; TANG; KAO, 2009, p. 415)

Para quantificar a comparação de resultados entre os dois *rankings*, PageRank e BRank, foram utilizados o coeficiente de correlação e o coeficiente Kappa. No coeficiente de correlação, um (1) indica um par perfeito de relação, já -1 indica um par com relação oposta e 0 um par sem nenhuma relação. A Figura 2.7 mostra o resultado na comparação do coeficiente de relação entre o PageRank e o BRank.

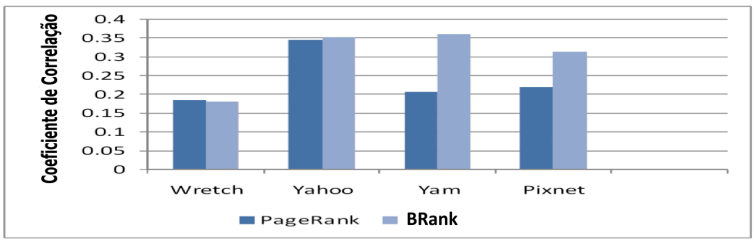


Figura 2.7: Coeficiente de correlação de comparação do PageRank e BRank.
Fonte: (LIN; TANG; KAO, 2009, p. 417)

Nota-se que o resultado dos dois algoritmos é muito semelhante nas bases de dados do Wretch e Yahoo e somente nos blogs Yam e Pixnet verifica-se uma pequena melhoria nos resultados utilizando mais informações sociais. Uma justificativa dos autores é que, no caso dos blogs Wretch, os comentários são muito frequentes pois se trata de blogs com postagem de assuntos pessoais e comentários curtos e pouco informativos. Esse tipo de ruído pode aumentar os pontos de blogs pequenos, mas com alta conectividade na comunidade.

Foi utilizada uma avaliação humana sobre o ranqueamento dos blogs. A avaliação consistitu em 400 blogs aleatoriamente selecionados do site BlogLook sobre os BSPs selecionados. As características manualmente definidas são: detalhamento das discussões, abundância de informação, clareza nas categorias, frequência de atualização, qualidade das postagens e pontuação de popularidade. O *ranking* definido manualmente foi comparado com o *ranking* definido pelo BRank e obteve um pequeno índice de relação entre os métodos.

Como a análise de conteúdo consome muito tempo, a coocorrência de *hyperlinks* foi usada para determinar similaridade entre os blogs. O BRank aplicado a um BSP pode ser executado com eficiência e provendo um resultado eficaz. A comparação dos resultados mostra uma pequena concordância entre o julgamento humano e o método proposto de ranqueamento de blogs.

2.5.2 Identificando Autoridades por Tópicos em Microblogs

Conteúdos de microblogs como o Twitter são produzidos por milhões de usuários. Essa diversidade é desafiadora para encontrar autores interessantes e autoridades em tópicos. Para resolver esse problema, Pal e Counts (2011) desenvolveram um algoritmo de agrupamento para identificar uma lista de autoridades por tópicos.

Foi utilizada uma lista de métricas extraídas e calculadas para cada autoridade em potencial. Dado a natureza dos *tweets* (texto curto que às vezes contém URLs) e a maneira que eles são frequentemente usados (para conversas rápidas via respostas e difusão de informação via retweet), os autores focaram em métricas que refletem o impacto dos usuários no sistema, especialmente com relação ao tópico de interesse.

Diversas métricas sobre os *tweets* e *retweets* foram definidas para comparar os usuários. Todas as métricas são bastante simples de calcular para um autor e sua rede. Além disso, essas métricas podem ser calculadas em paralelo para todos os usuários, sendo possível integrar com um *framework* de computação paralela como o MapReduce[4] (DEAN; GHEMAWAT, 2004).

Utilizando tais métricas, os autores usaram o Modelo de Mistura Gaussiana, que é um modelo probabilístico para representar subpopulações, para dividir os usuários em dois grupos. O motivo de usar o agrupamento é diminuir o espaço de busca. Isso também torna a lista de autoridades mais robusta, pois evita o aparecimento de usuários fora-de-série, tais como celebridades.

Os experimentos foram executados sobre a base completa do Twitter entre cinco dias de intervalo (6 a 10 de Junho de 2010). Numa parceria dos autores com o Twitter, eles obtiveram toda a base de 90 milhões de *tweets* nesse período de tempo. Foi extraídos *tweets* de três tópicos: *oil spill, world cup e iphone*.

Foram comparados diversos métodos:

- método dos autores: método sobre as métricas
- *baseline1*: propriedades de grafo sobre as métricas RI (*Retweet Impact*), MI (*Mention Impact*), ID (*Information Diffusion*). Depois é executado o PageRank sobre o grafo direcionado com peso sobre as menções.
- *baseline2*: análise textual sobre TS (*Topical Signal*), SS (*Signal Strenght*), CS (*Non-Chat Signal*)
- *baseline3*: são selecionados aleatoriamente usuários fora do grupo alvo.

Foi extraída a lista dos primeiros dez autores descobertos pelos algoritmos selecionados. A Tabela 2.3 mostra o número médio de seguidores da lista das dez autoridades encontradas.

A Tabela 2.3 mostra que o número médio de seguidores para o método dos autores é menor que o *baseline1* e maior que o *baseline2*, indicando que atinge um equilíbrio entre as propriedades da rede e textuais de um usuário influente em certo tópico.

Os resultados desse método confirmam que foi possível encontrar autores que produzem textos interessantes e que são considerados autoridades. Foi mostrado que o modelo probabilístico de agrupamento é efetivo na descoberta de grande número de elementos fora-de-série no uso de métricas e seleciona um alto nível de usuários como autoridades onde uma lista de ranqueamento pode ser aplicada com mais robustez.

[4]Modelo de programação paralela e distribuída desenvolvido pelo Google.

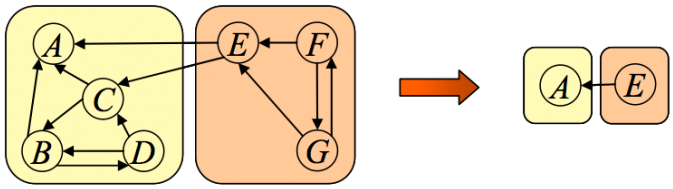
Tabela 2.3: Número médio de seguidores para a lista de dez autores para os diversos algoritmos.

	método dos autores	**baseline1**	**baseline2**	**baseline3**
iphone	282.665	1.364.015	117.250	1.252
oil spill	462.507	871.159	411.210	840
world cup	29.373	32.121	18.017	277

Fonte: (PAL; COUNTS, 2011, p. 49)

2.5.3 Identificando Líderes de Opinião na Blogosfera

Líderes de opinião são aqueles que trazem novas informações, ideias e opiniões, e então disseminam e influenciam a opinião e as decisões de outros usuários pelo mecanismo boca-a-boca. Essa é a definição de autoridade para Song et al. (2007). Os líderes de opinião são responsáveis pela maior parte das opiniões dadas nas redes sociais, então é importante entender esse comportamento na blogosfera. Nesse trabalho, Song et al., definiram o InfluenceRank, que identifica esses personagens da rede não só por sua importância como também pela quantidade de informações novas que eles geram.



(a) A rede de um blog (b) Líderes de Opinião

Figura 2.8: Exemplo de motivação para o InfluenceRank.
Fonte: adaptado e traduzido de (SONG et al., 2007, p. 971)

A Figura 2.8 ilustra sete blogs e suas referências ou ligações quando os autores publicam suas opiniões. Os blogs A, B, C e D discutem sobre *How to use Riya image search engine*. Na figura anterior, o blog E inicia uma discussão sobre uma possível compra da Riya pela Google e cita os blogs A e C. Seguindo o blog E, os blogs F e G começam uma discussão sobre esse boato de compra. Nesse exemplo, os blogs A e E são líderes de opinião, pois iniciam a discussão de um assunto que é difundindo entre os outros blogs. Esse grafo gerado pelas citações é utilizado para verificar as semelhanças entre blogs.

Para encontrar esses líderes de opinião o algoritmo InfluenceRank deve identificar novas informações na rede e onde elas são usadas novamente. Para medir a novidade de uma informação, primeiro é considerada cada postagem como um documento. Para reduzir a quantidade de dados é executado o LDA para gerar um espaço de tópicos sobre os documentos. Depois é executado o KL (Kullback-Leibler) para descobrir a similaridade dos documentos ou a diferença entre eles. É também criado um nodo escondido no grafo para considerar uma fonte externa de novidade como um jornal, televisão ou revista que o autor tenha utilizado como fonte para escrever seu texto.

Depois de criar o espaço de tópicos dos documentos, é calculado o nível de novidade oferecido pelos nodos escondidos e existentes sobre as postagens dos blogueiros como define a fórmula:

$$Nov(A|Out(A)) = \frac{\sum_{Ae \in A} Nov(Ae|Out(Ae))}{card(Set(Ae))} \tag{2.3}$$

onde Ae é uma entrada (postagem) no grafo feito pelo blog A, $card()$ é o total de entradas geradas pelo blog A. $Nov()$ é o cálculo de novidade gerado pelo blog A sobre as suas citações recebidas em $Out()$.

Dado o nível de novidade sobre as entradas para cada blog, pode-se calcular o InfluenceRank. Os n blogs fazem parte de um grafo direcionado de citações **G**. Gij=1 se o blog i tem ligação para j, se não é zero. O cálculo é executado sobre os blogs ligados, tendo como base a equação a seguir:

$$IR^t = (1 - \beta) * IR^t * W + \beta * Nov^t \tag{2.4}$$

onde β é o parâmetro que defini quão significante é a informação para o líder que se espera detectar e **W** é a matriz de adjacência normalizada de **G**. Quando β é grande, tende-se a achar usuários que contribuem com informações mais novas. O InfluenceRank pode ser rodado da mesma forma que o RandonWalk e utilizando a mesma estratégia do PageRank para pular nodos circulares ou sem grau de saída.

Para os experimentos foi utilizada uma base de dados coletada entre Julho de 2005 e Outubro de 2006, contendo 67 mil entradas, 11 mil termos e 12 mil links entre os blogs. Para avaliar o desempenho dos algoritmos foram utilizadas três métricas: cobertura, diversidade e distorção. Cobertura avalia a influência do nodo sobre a rede, diversidade verifica as abrangência de termos entre os pares de postagens existentes e distorção é utilizada o KL para verificar as diferença entre os pares.

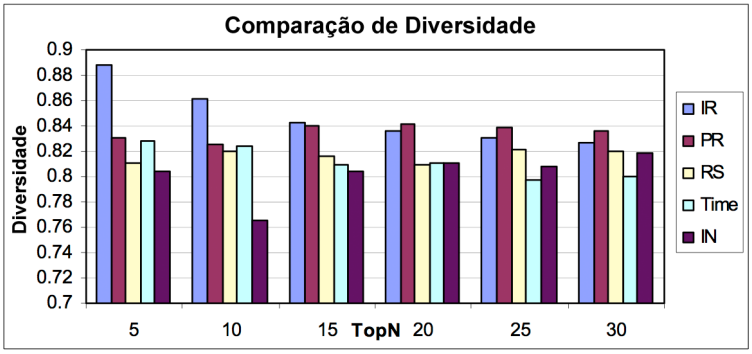


Figura 2.9: Comparação de diversidade entre os algoritmos.
Fonte: (SONG et al., 2007, p. 974)

Foram analisados o desempenho do InfluenceRank (IR) com o PageRank (PG), Amostragem Aleatória (RS), Ranqueamento baseado no tempo (Time) e Ranqueamento por novidade de informação (IN).

A Figura 2.9 mostra onde o InfluenceRank melhor se comportou na identificação de líderes de opinião. Foi comparada a diversidade dos líderes de opinião identificados pelos cinco algoritmos. Os líderes identificados pelo InfluenceRank proverão maior diversidade comparado com os líderes identificados pelos demais algoritmos de *baseline*.

O algoritmo proposto por Song et al. (2007) teve um desempenho pouco melhor que o PageRank, o que não justifica todo o trabalho de análise de texto sobre os blogs para obter uma pequena melhora, embora a técnica talvez possa ser estendida para soluções de outros

problemas. Nesse caso ainda podemos considerar o PageRank como uma abordagem mais direta para descoberta de líderes ou autoridades, como é feito nesta dissertação.

2.6 Resumo do Capítulo

As três estruturas vistas estão muito relacionadas: os blogs devem ser vistos como uma rede social e as autoridades dessa rede são os autores mais importantes entre os blogueiros. Considerando os relacionamentos entre usuários e as interações dentro dos blogs, (links e comentários) essas informações podem ser usadas para fazer uma análise de comportamento entre os usuários. Essa plataforma de publicação de informação, que são os blogs, gera um conteúdo aberto que pode ser usado para definir uma consciência coletiva que age como um armazenamento de conhecimento desses usuários.

Esse conhecimento abre a possibilidade de fazer descobertas e explorar essas informações para tirar proveito dessa consciência coletiva. Utilizar essas informações para a busca de autoridades sobre tópicos é um trabalho desafiador. As pessoas constantemente influenciam umas as outras e a descoberta dos usuários mais influentes e críticos pode ajudar as empresas a entender como seus produtos e serviços são consumidos e podem ser melhorados.

3 METODOLOGIA PROPOSTA PARA A IDENTIFICAÇÃO DE AUTORIDADES

Os trabalhos relacionados descritos na Seção **??** mostram que é relevante a pesquisa sobre a Blogosfera, e a variedade de estruturas encontradas nessa rede social traz diversos desafios para a descoberta de conhecimento. Um desses desafios é a identificação de autoridades para um dado tópico. Resolvendo esse problema, é possível saber qual o blogueiro mais importante para um determinado assunto e recomendar conteúdos melhores para os usuários.

Baseado na afirmação feita por Ali-hasan e Adamic (2007), que diz que os comentários são o relacionamento mais verdadeiro entre os usuários de uma rede social baseada em blogs, esta dissertação foca em um algoritmo de classificação de autores usando esse tipo de relação. Para realizar essa tarefa, foi criado um conjunto de dados específico contendo autores e os comentários dados pelos usuários (detalhes sobre a base de dados são abordados na seção 4.1). Utilizando esses dados coletados, são construídos diferentes grafos sobre os diversos tópicos escolhidos. Nesses grafos, os nodos representam os autores e as arestas representam os comentários escritos pelos usuários. Mais detalhes sobre a construção do grafo são dados na seção 3.2.

Nas próximas seções, serão descritos alguns desafios ao se aplicar mineração sobre blogs, será descrito como o grafo é construído a partir das postagens, e será feita a descrição do algoritmo proposto para o ranqueamento de usuários e explicado como ele é utilizado para a identificação das autoridades em cada tópico.

3.1 Mineração em Dados de Blogs

A blogosfera oferece várias perspectivas interessantes de mineração de dados, já que é possível extrair conhecimento pelas postagens, comentários e outros tipos de interações entre os usuários. Por enquanto, das postagens, é possível inferir tópicos e suas tendências, identificando comunidades, rede de usuários em tópicos, e classificando o blog por seu tópico (ADAMS; PHUNG; VENKATESH, 2010). Dos comentários e *trackbacks*, é possível definir os disseminadores de opinião, encontrar comunidades e definir autoridades (LIN; TANG; KAO, 2009).

Como existem diversas plataformas de publicação de blogs disponíveis para coleta, deve-se primeiro escolher quais delas serão utilizadas, aquelas que são mais importantes para a tarefa de mineração. Alguns usuários usam os serviços gratuitos para publicar seu conteúdo, enquanto outros preferem personalizar seu blog e instalá-lo em seus próprios servidores.

Extrair as informações dos blogs em serviços gratuitos se torna mais fácil quando eles

proveem uma API, interface direta para acessar as informações, ou uma lista de todos os blogs que essa plataforma mantém. Outra forma de extrair dados de blogs consiste em coletar os blogs já catalogados por máquinas de busca especializadas como BlogPulse[1] ou Technorati[2], mas geralmente essas ferramentas só catalogam os blogs mais populares.

Depois de escolher a parte da blogosfera que será extraída, o próximo passo consiste em criar um coletor para obter os dados. O coletor precisa ser personalizado para a estrutura de cada tipo de blog, pois cada plataforma contém um modelo diferente de HTML, ou então trabalhar com coletores que aprendem esses modelos (ZHANG et al., 2009). O número de modelos disponíveis na internet para fazer essa coleta é relativamente baixo, entretanto esses modelos mudam e se atualizam, sendo necessário atualizar o coletor também. Na literatura, é possível encontrar abordagens que automaticamente ou semiautomaticamente constroem ferramentas de coleta, tais como a de Lam, Gong e Mayeba (2008). Outra maneira de coletar os dados é utilizando a API que provê as informações de forma organizada como XML ou JSON. Então, nesta dissertação, como a extração de informações não é o foco deste trabalho, foi escolhida uma plataforma de publicação que oferece uma API de comunicação para a extração dos dados.

3.2 Construção do Grafo

De acordo com Wasserman e Faust (1994), análise de redes sociais é o estudo quantitativo dos relacionamentos entre os indivíduos ou organizações. A análise de redes sociais representa os relacionamentos em grafos onde indivíduos são representados como nodos (também referenciados como atores ou vértices) e as conexões entre eles são as arestas (também conhecidos como ligações ou vínculos). Para quantificar as estruturas sociais, a análise de redes sociais pode determinar os nodos mais importantes de uma rede (WASSERMAN; FAUST, 1994).

Para representar a blogosfera como uma rede social, os usuários (i.e., blogueiros) são os nodos e os comentários são os relacionamentos entre os usuários. A maioria dos trabalhos usa os blogs como a entidade representada nos nodos. Entretanto, neste trabalho, como é procurada uma pessoa que represente a autoridade de um tópico específico, serão utilizados os usuários como vértices desse grafo e os comentários serão suas arestas. A escolha pelos comentários como aresta é feita com base na afirmação de Ali-Hasan e Adamic (2007), considerando que os comentários proveem a melhor ligação entre os usuários. Em trabalhos futuros, outros vínculos podem ser usados para enriquecer o grafo da rede social de blogs.

O objetivo principal deste trabalho é construir uma lista dos principais autores (i.e., autoridades) sobre um determinado tópico e, para alcançar isso, somente os comentários pertencentes às postagens sobre esse tópico serão usados para construir o grafo. Usar somente os comentários relacionados com um determinado tópico reduz o tamanho do grafo, diminuindo a complexidade da tarefa (i.e., o espaço de busca) e foca a iteração do algoritmo nos usuários que têm interesse no tópico específico.

Por simplicidade e porque não é o objetivo principal deste, as postagens, utilizadas para montar o grafo dos comentários, são coletadas usando as tags das postagens como consulta. Nesse caso, a tag é armazenada no banco depois de passar pelo RSLP (ORENGO; HUYCK, 2001) e usada para consultar as postagens sobre determinado tópico. Entretanto, outras abordagens mais complexas podem ser usadas para fazer uma

[1] http://www.blogpulse.com (desativada)
[2] http://technorati.com/

seleção mais relevante de postagens. Pode ser utilizado, por exemplo, o nível de entropia entre as palavras para selecionar aquelas que pertencem ao mesmo contexto (SHANNON, 2001). Também se pode usar o método LDA (BLEI; NG; JORDAN, 2003) (Latent Dirichlet Allocation), que é capaz de classificar documentos em tópicos (ANDRZEJEWSKI; BUTTLER, 2011; TSAI, 2011), ou NMF (LEE; SEUNG, 1999) (Non-negative Matriz Factorization), que pode ser usado para fazer agrupamento de documentos (XU; LIU; GONG, 2003).

Neste trabalho, a abordagem LDA poderia ter sido usada na extração de postagens, considerando mais palavras, e para encontrar postagens com distribuição de tópico similar, o que poderia servir como um conjunto mais interessante para o dado tópico. Novamente, no entanto, esse não é o objetivo deste trabalho e é uma tarefa extremamente custosa em termos de recursos computacionais, o que poderia levar mais tempo para ser finalizada do que havia sido previsto.

Então, nesse caso, os comentários utilizados para construir o grafo são aqueles feitos em postagens que contêm a tag relacionada ao tópico de interesse (p.ex., saúde, política). É importante salientar que os usuários identificados são de dois tipos: o autor, a pessoa na qual escreve a postagem, e o leitor, aquele que faz o comentário na postagem.

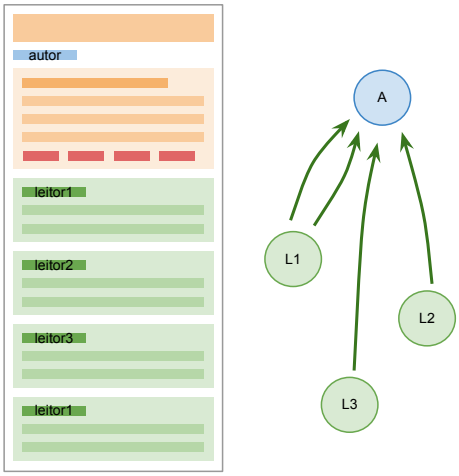


Figura 3.1: Criando o grafo utilizando os comentários e as postagens.
Fonte: figura elaborada pelo autor

A Figura 3.1 mostra como são construídas as relações entre os blogueiros. Cada comentários realizado por um usuário gera uma aresta direcionada entre o leito e o autor, podendo haver diversas arestas iguais.

Como está ilustrado pela Figura 3.1, o grafo dos tópicos G(V,E) é defino como:

- V = todos os usuários que escrevem ou comentam em postagens sobre o tópico;
- E = o conjunto de comentários entre dois usuários, o leitor e o autor, em uma postagem sobre o tópico.

Cada comentário realizado por um leitor gera uma aresta direcionada para o nodo do autor no grafo. Todos os nodos podem ter arestas entre si, exceto arestas reflexivas. As arestas podem se repetir entre os nodos tantas vezes quantos forem os comentários existentes, em ambas as direções. É importante ressaltar que o grafo é direcionado, requisito

necessário para a execução do algoritmo PageRank. Isso significa que se um usuário a é comentado na postagem de um usuário b, a aresta do grafo será direcionada de a para b. Se b também comentar em uma postagem de a, haverá outra aresta (no sentido contrário) entre a e b. Com a geração do grafo do tópico, é possível identificar a autoridade usando o algoritmo explicado a seguir.

3.3 PageRank de Tópicos sobre os Blogs

O PageRank (PAGE et al., 1998) é um algoritmo popular para análise de ligações em grafos e ranqueamento de páginas populares na Web. Para um conjunto de páginas ligadas, o PageRank associa um valor para cada página, representando a sua importância relativa sobre as outras páginas. A versão original do PageRank utiliza o imenso grafo de toda a web coletada para computar o valor de cada nodo. O valor global é de cada página é dado considerando o valor de cada aresta de chegada.

Nesse ponto, escolheu-se outra abordagem para a construção do grafo, fazendo a seleção dos nodos pelo seu contexto semântico, considerando somente aqueles que estejam envolvidos em certo tópico. Com isso é possível criar uma lista melhor e mais específica de autoridades sobre um determinado assunto. Além disso, a computação em um grafo de tópicos é mais rápida que a versão global, quando uma abordagem de tópico é aplicada ao problema.

No trabalho de Haveliwala (2002) a iteração também ocorre sobre todos os nodos de todos os tópicos, dando diversos valores de PageRank para cada nodo, um valor para cada tópico. Nesse trabalho, o PageRank inicia com um (1) caso o nodo faça parte do tópico e zero caso contrário, nos diversos vetores de tópicos existente. Assim, o grafo não é divido, somente os nodos iniciam com zero, mas a computação é executada sobre todos os nodos e, por programação dinâmica, os diversos vetores são preenchidos simultaneamente. Nesta dissertação, a computação é feita separadamente para cada tópico de interesse.

Depois de o grafo ser criado, as iterações do PageRank podem ser executadas até convergir. A convergência é definida pelo erro quadrático, a diferença entre os valores de todos os nodos entre duas iterações. Se o erro for menor que o desejado, o algoritmo para. Desse grafo resultante, é possível obter uma lista dos autores mais populares para o tópico, em uma ordem decrescente do valor do PageRank associado aos nodos.

É importante destacar que essa abordagem é a mesma utilizada por Haveliwala (2002), entretanto aqui os experimentos levam em conta blogs ou invés de páginas da Web. Além disso, foram utilizados os comentários realizados pelos usuários para ligar os nodos do grafo, diferentemente do que é feito no algoritmo original do PageRank, onde os nodos são ligados por links em páginas. A estrutura de blogs tem links para serem explorados e esse é um trabalho futuro a ser desenvolvido: comparar a abordagem de comentários com a utilização de links para descoberta de autoridades em tópicos.

Nas próximas seções será explicado como foi feita a coleta dos dados. Na Seção 4 são demonstrados alguns experimentos sobre os tópicos selecionados, utilizando a abordagem descrita até o momento.

3.4 Coleta de Blogs

A plataforma Blogspot foi escolhida para a coleta dos blogs, pois é gratuita, além de ser bastante conhecida, primeiro porque ela oferece uma interessante API de desen-

volvimento (Blogger API[3]) que facilita o processo de extração, e segundo porque é a plataforma mais utilizada no Brasil (GOOGLE, 2012). O Blogger API provê serviços para coletar a lista de postagens de um dado blog, extrair o conteúdo de uma postagem e recuperar os comentários de uma dada postagem. Para cada blog, suas postagens e comentários, é associado um número de identificação do autor e leitor que comenta, permitindo a identificação dos usuários (de fato, por motivos de privacidade, não foi coletada nenhuma informação pessoal, somente o número de identificação, que é usado para identificar os usuários em suas postagens e comentários).

Outra plataforma de publicação de blogs bastante utilizada no Brasil é o Wordpress[4], que oferece informações mais ricas, porém mais difíceis de serem coletadas. Essa plataforma tem recursos de *trackback* e informação de preferência dos usuários sobre as postagens, que pode ser usando como dado de endosso. Mas essas informações não podem ser obtidas facilmente e é necessário o desenvolvimento de um interpretador específico para essa plataforma. Como os dados do Blogspot já são suficientes para os experimentos, o Wordpress foi descartado.

3.4.1 Coletor

Foi escolhido como estudo de caso a blogosfera brasileira. Para isso, para coletar alguns blogs brasileiros, foi necessário procurar por usuários que se autodeclaram brasileiros em seus perfis no Blogspot. A plataforma provê a busca de usuários por localidade, e essa é a semente utilizada pelo coletor para começar a coleta. A Figura 3.2 exemplifica como funciona o processo de coleta de dados dos blogueiros brasileiros.

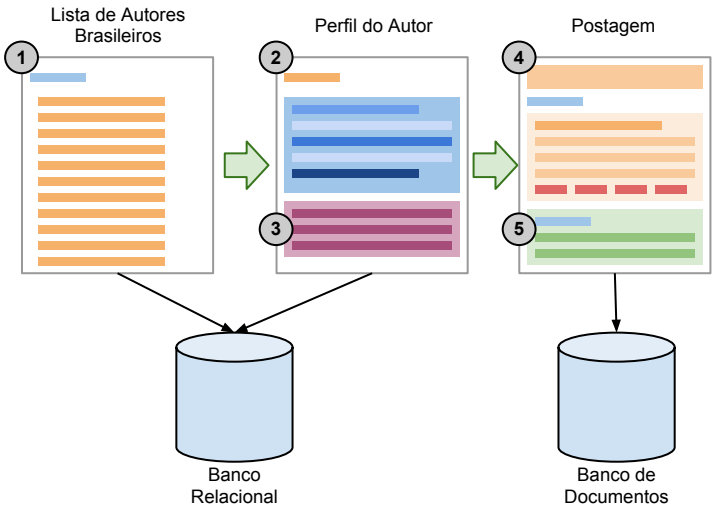


Figura 3.2: Fluxo de coleta dos dados sobre o Blogspot.
Fonte: figura elaborada pelo autor

O primeiro processo, identificado na Figura 3.2 com o número 1, é responsável pela busca de perfis de usuários brasileiros na lista fornecida pelo Blogspot. A lista contém diversos usuários que são mostrados aleatoriamente e paginados de 20 em 20 até o número 4.000. O processo percorre todas as páginas dessa lista e começa novamente com a requisição de uma nova lista aleatória. Não foi possível precisar o número total de usuá-

[3]https://developers.google.com/blogger/
[4]http://wordpress.org/

rios brasileiros no Blogspot, pois a cada consulta o valor total de usuários muda sem uma progressão definida. O maior valor mostrado foi o de 16 milhões de usuários brasileiros.

Depois de obter alguns usuários, o segundo processo, identificado pelo número 2 na Figura 3.2, executando paralelamente com o processo 1, busca a lista de blogs que o usuário contribui e seus blogs favoritos (*blogroll*). O processo 3 trabalha sobre a lista de blogs encontrados nos perfil brasileiros. Como os usuários podem contribuir ou seguir blogs não brasileiros, o processo 3 obtém o idioma escrito nos blogs, informação fornecida pela API da plataforma. Todas as informações dos blogueiros e suas listas de blogs são armazenadas em um banco de dados relacional, que facilita a execução de consultas de agrupamento, para fazer estatísticas dos perfis e blogs mais encontrados.

Tendo a lista dos blogs brasileiros, o processo de número 4, em paralelo com os demais, faz a coleta das postagens desses blogs com os seus respectivos comentários. Esses dados são armazenados em um banco de dados orientado a documentos (MongoDB), facilitando o armazenamento dos dados de forma semiestruturada, sem a necessidade do uso de relacionamentos e número fixo de colunas para guardas os dados. Essa flexibilidade facilita a inserção dos dados, pois como não há a ideia de relacionamento nesse tipo de banco, não é necessário fazer busca de chave-estrangeira para adicionar os comentários em seus posts. O processo 5 é responsável por catalogar no banco de dados relacional todos os usuários encontrados na lista de comentários dos blogs. Como alguns deles não foram obtidos pela lista de blogueiros brasileiros, é necessário definir a nacionalidade desses blogueiros e obter a lista de seus blogs caso ele seja brasileiro.

O banco de dados escolhido para armazenar os documentos foi o MongoDB[5]. Além de armazenar os dados de forma semiestruturada, ele permite a alocação dos dados de forma distribuída em clusters de computadores. Essa técnica distribuiu a carga dos dados em diversos computadores e permite a computação distribuída e paralela dos dados.

Os dados das postagens são armazenados no banco de dados orientado a documento MongoDB, que permite distribuir os dados em diversas máquinas para aumentar o poder de computação, utilizando o modelo de programação paralelo e distribuído MapReduce (DEAN; GHEMAWAT, 2004). Assim, o PageRank pode ser executado paralelamente em cada uma das máquinas utilizando o paradigma MapReduce (BAHMANI; CHAKRABARTI; XIN, 2011), podendo ser escalado para outras máquinas facilmente. Os processos de busca pelas tags e construção do grafo para executar o PageRank também são executados paralelamente e distribuídos entre as *shards*, pedaços do banco de dados.

Mais informações sobre a infraestrutura utilizada para executar o coletor e armazenar os dados podem ser encontradas no Apêndice A desta dissertação.

3.5 Resumo do Capítulo

Organizado o modo como o trabalho será desenvolvido e o escopo de estudo, pode-se dar início à coleta dos dados sobre os blogs. Assim os dados coletados já serão armazenados de forma estruturada, visando a resolução do problema de descoberta de autoridades nos tópicos.

[5]http://www.mongodb.org/

4 RESULTADOS, EXPERIMENTOS E VALIDAÇÃO

No início do capítulo serão mostrados os dados coletados e estatísticas dos usuários. Na seção 4.4 serão apresentados os resultados da abordagem de tópicos e global do PageRank aplicados aos grafos da rede social de blogs. Serão descritas as métricas utilizadas para a comparação dos métodos. Então será demonstrada a eficácia desta abordagem para identificar autoridades em tópicos. Nos experimentos realizados, focou-se nas *tags* mais utilizadas nas postagens para gerar grafos mais representativos.

4.1 Base de Dados

O processo de coleta focou nos posts gerados pelos blogueiros nos meses de maio a julho de 2012. A coleta foi realizada sobre 2,7 milhões de blogs da plataforma Blogspot, entre os dias 4 e 7 de setembro de 2012. Como o algoritmo trabalha sobre as interações de comentários e tags, os textos da postagem e comentários não são armazenados no banco de dados. Nas próximas seções serão descritas as estatísticas dos dados coletados.

A coleta foi focada em blogueiros brasileiros. Na Tabela 4.1 é apresentado o montante de usuários, blogs, postagens e comentários coletados. Foram contabilizados somente os usuários que tiveram interações de comentários com outros usuários. Alguns blogs não tiveram nenhuma postagem e não tiveram nenhum comentário, como será explicado na Seção 4.1.1.

Tabela 4.1: Visão geral sobre os dados coletados

Conteúdo	**Quantidade**
Blogs	2.758.286
Postagens	9.635.902
Comentários	4.074.244
Usuários	511.162

Fonte: tabela elaborada pelo autor com base nos dados coletados.

No Blogspot, a menos que seja bloqueado pelo blogueiro, os comentários podem ser feitos por qualquer usuário da internet, mesmo aqueles não registrados na plataforma. Entretanto, para melhor identificar os usuários e construir uma base de dados mais coerente, somente foram coletados usuários que tinham um número de identificação do Blogspot.

4.1.1 Atividade dos Blogs

A maioria dos trabalhos (ALI-HASAN; ADAMIC, 2007; LIN; TANG; KAO, 2009; ADAMIC; GLANCE, 2005) focou suas coletas nos blogs mais populares. Como o in-

teresse deste trabalho está na análise de todos os usuários, foram coletadas postagens de todos os blogs, independente de usa popularidade. Os blogs foram divididos em três categorias, de acordo com a quantidade de comentários recebidos, como demostra a Figura 4.1.

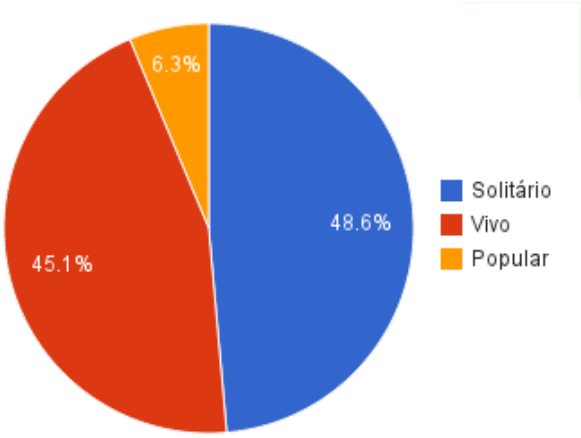


Figura 4.1: Popularidade dos blogs de acordo com os comentários recebidos.
Fonte: elaborada pelo autor

Alguns blogs publicam diversas postagens, mas não recebem nenhum comentário, esses blogs foram classificados como Solitários no gráfico da Figura 4.1. Vivos são os blogs que publicaram conteúdo e receberam pelo menos um comentário em 2012. Os blogs classificados como Populares são aqueles que receberam pelo menos um comentário em cada postagem feita em 2012.

4.1.2 Perfil dos Blogueiros

Conhecer o perfil dos blogueiros brasileiros ajuda a entender os seus interesses e comportamentos. Na Tabela 4.2 é possível perceber que a grande maioria dos comentários feitos nos blogs brasileiros é das mulheres. Isso é interessante e curioso, pois mais da metade das postagens são feitas pelos blogueiros homens e apenas um quinto dos comentários é feito por eles. As mulheres se mostram mais comunicativas, considerando os comentários.

Tabela 4.2: Sexo dos blogueiros brasileiros na base de dados

Sexo	**% Blogueiros**	**% Postagens**	**% Comentários**
Feminino	57.2	46.8	77.6
Masculino	42.8	53.2	22.4

Fonte: tabela elaborada pelo autor com base nos dados coletados.

Também é interessante saber a ocupação desses usuários e como eles interagem na rede. A Tabela 4.3 resume as profissões mais representativas dos blogueiros brasileiros.

Alguns setores têm maior representatividade em número de blogueiros, mas Artes, mesmo tendo um número menor de usuários, tem a mesma quantidade de comentários que Estudantes e profissionais da Educação. Essa análise é importante para verificar as subcomunidades que mais interagem.

É importante ressaltar um comportamento comum na blogosfera verificado por Ali-Hasan e Adamic (2007): existem os usuários que iniciam as discussões em seus blogs e

Tabela 4.3: Ocupação dos usuários na base de dados

Ocupação	**% Blogueiros**	**% Postagens**	**% Comentários**
Estudante	18.3	13.9	16.7
Educação	14.2	12.0	15.0
Artes	12.3	9.3	16.2
Comunicação	5.4	11.7	6.0
Moda	4.7	3.5	7.5
Internet	4.4	6.9	3.8
Tecnologia	3.8	3.8	2.9
Negócios	3.1	2.5	1.8
Esportes e Recreação	2.5	2.7	1.2
Religião	2.5	2.4	0.7
Outros	28.9	31.3	26.0

Fonte: tabela elaborada pelo autor com base nos dados coletados.

usuários que somente comentam em outras postagens. É um comportamento comum na Web, analisado por Nonnecke e Preece (2000), existindo a proporção 90-9-1: onde 90% dos usuários são observadores e não contribuem nem comentam nos conteúdos; 9% são pessoas que contribuem em conteúdos criados por outros; e 1% dos usuários é de fato criador de conteúdo e geram mais informações novas.

Outras características dos blogueiros brasileiros podem ser encontradas no Apêndice D desta dissertação, onde foi realizada uma pesquisa com quatro mil blogueiros pertencentes ao banco de dados coletado.

4.1.3 Reciprocidade

A reciprocidade é importante para medir o comportamento social dos usuários. Se as interações e os relacionamentos são verdadeiros formadores de uma comunidade, esses laços serão recíprocos (BERSCHEID; REIS, 1998). Na base de dados coletados, a reciprocidade nos comentários é de 12,56%. Isso é bastante baixo comparado com a rede social encontrada nos blogs do Kuwait (ALI-HASAN; ADAMIC, 2007), mas acredita-se que isso aconteça pelo fato do trabalho ter sido realizado em um conjunto pequeno e focado de blogs do Kuwait, com comunidades restritas. Mesmo com um número baixo de comentários recíprocos, essa informação pode ser utilizada para encontrar subcomunidades que tenham mais reciprocidade. Essas comunidades auxiliam a identificação de laços e relacionamentos mais fortes entre os blogueiros.

Outra distribuição do banco de postagens é o tempo que a postagem recebe comentários. No conjunto de dados coletados, 58% dos comentários são feitos no primeiro dia da postagem, provavelmente usuários que acompanham o blog. Para obter 90% dos comentários feitos em uma postagem é necessário esperar o 9º dia de vida de uma postagem. Portanto, se um coletor tem a intenção de pegar a maioria dos comentários em uma postagem, ele deve esperar 10 dias.

Com esses dados coletados e analisados, é possível executar os algoritmos de autoridades sobre os diversos tópicos propostos, como será abordado na próxima seção.

4.2 Tópicos Selecionados

Para servir como conjunto de teste para os experimentos foram selecionadas as seis tags mais frequentes nas postagens coletadas: *Moda, Música, Decoração, Filmes, Livros* e *Fotografia*. Esses tópicos foram selecionados, pois eles têm uma quantidade grande de dados e interações entre os usuários. A Tabela 4.4 apresenta a distribuição considerando os dados extraídos.

Tabela 4.4: Distribuição dos dados coletados sobre cada um dos tópicos selecionados

Tag	**Postagens**	**Comentários**	**Blogueiros**
Moda	11.479	81.308	14.743
Música	6.245	28.519	11.461
Decoração	4.773	32.617	9.606
Filmes	4.751	22.483	9.766
Livros	4.616	34.904	9.795
Fotografia	4.368	18.949	8.361

Fonte: tabela elaborada pelo autor com base nos dados coletados.

A Tabela 4.4 mostra o comportamento entre os diferentes tópicos selecionados. Mesmo *Música* tendo mais postagens e usuários envolvidos com o tema, *Livros* tem mais comentários, mostrando ser uma comunidade mais interativa. A quantidade de postagens utilizadas para a construção do grafo é muito inferior à coletada, nove milhões, pois somente 10% das postagens recebeu comentário.

4.3 Avaliação e Validação dos Experimentos

Três análises serão feitas sobre os grafos gerados pelos relacionamentos dos autores através dos comentários. Primeiro será feita uma análise quantitativa da produção de postagens e dos comentários recebidos dos autores mais importantes de cada grafo para analisar o envolvimento deles com o tópico. Depois serão feitas comparações com a lista de autores ordenada pelo PageRank com outros ordenamentos baseados em número de postagens e comentários recebidos, a fim de avaliar a correlação dos diferentes ordenamento. Além disso, será feita uma análise da topologia dos grafos para entender melhor a estrutura gerada por esses relacionamentos dos blogueiros de forma geral e pelos diferentes tópicos.

Como medida principal para avaliar a eficiência dos algoritmos, utilizou-se o resultado baseado no ranqueamento dos primeiros 10 autores da lista, conforme indicado por Kelly, Fu e Shah (2010). A métrica principal para medir o desempenho dos algoritmos é a soma de postagens criadas pelos 10 primeiros autores e os comentários recebidos por eles. Aqui é considerado que um usuário que escreve mais postagens sobre o tópico e recebe muitos comentários nelas é o usuário mais importante.

Outra métrica considerada é a correlação de Spearman (1904) sobre o *ranking* gerado pelas duas abordagens do PageRank, Global e por Tópicos, e os diversos ordenamentos possíveis: por visualização de perfil, número de posts criados, quantidade de comentários recebidos. Utilizou-se a correlação de Spearman ao invés da de Pearson, pois ela trabalha melhor com pequenas variações de itens fora de série. As visualizações foram retiradas dos perfis dos usuários, mas é uma informação com certo viés, pois não considera o período da coleta dos dados. As visualizações é o número de visitas recebidas pelo perfil

desde a sua criação, portanto, usuários mais antigos têm mais chance de receber visitas, mas é um valor possível para comparação de popularidade.

É importante salientar que a correlação de Spearman retorna um valor entre -1 e 1 quando duas listas são comparadas. Quando mais próximo de -1, as listas têm uma correlação inversa, ou seja, os primeiros colocados em uma lista são os últimos na outra. Se o valor obtido é próximo de zero, as listas não têm nenhuma correlação observável. E se o valor é mais próximo de 1 (um positivo), as listas têm forte relação de ordenamento.

A seguir são mostrados os resultados das comparações entre os ordenamentos do PageRank Global e de Tópicos sob diferentes perspectivas. O PageRank Global foi gerado executando o algoritmo do PageRank sobre o grafo geral de usuários e depois foram selecionados os autores que escreveram sobre o tópico. Portanto, o tópico foi selecionado *a posteriori*. No PageRank de Tópicos, os grafos são gerados selecionando somente os autores que escreveram sobre este assunto, depois foi executado o algoritmo do PageRank para definir as autoridades. Então, o tópico foi selecionado *a priori*.

4.4 Resultados

Para medir o conteúdo publicado pelos 10 autores ranqueados em ambas as abordagens, na Tabela 4.5 encontram-se os resultados sobre os tópicos em questão. O número de postagens criadas e comentários recebidos pelos 10 primeiros autores da lista referente ao tópico são representados por *TP* e *TC*, respectivamente, nos métodos de ordenamento GPR (Global PageRank), TPR (Topic PageRank).

O algoritmo PageRank classifica os nodos de um grafo de acordo com a sua importância, dando valores para os nodos. Portanto, os nodos podem ser ordenados de acordo com o seu valor de PageRank, tanto na abordagem Global quanto de Tópicos. Sendo assim, foram selecionados os primeiros 10 autores com maior valor de PageRank entre os experimentos realizados. Para o PageRank Global, foram selecionados os 10 autores, de cada tópico, com maior valor de PageRank. O mesmo foi feito com o PageRank de Tópicos: foram selecionados os 10 autores com maior valor de PageRank de cada grafo de tópico.

Assim, tem-se duas listas de 10 autores para cada tópico selecionado para os experimentos. Cada um dos autores escreveu postagens sobre o tópico e recebeu comentários. Aqui é considerado o melhor ordenamento aquele que tem autores que escreveram mais sobre o tópico e receberam mais comentários.

Tabela 4.5: Distribuição dos primeiros 10 autores sobre o PageRank Global e de Tópicos

Tag	**GPR #TP**	**TPR #TP**	**GPR #TC**	**TPR #TC**
Moda	90	4.533	**444**	**12.519**
Música	6	273	**91**	**1.816**
Decoração	26	1.402	**151**	**4.208**
Filme	237	1.605	**267**	**2.174**
Livro	9	134	**233**	**5.962**
Fotografia	13	314	**198**	**3.572**

Fonte: tabela elaborada pelo autor com base nos experimentos realizados.

Como era esperado, já que foram usados somente as postagens e comentários que estão envolvidos com a *tag* para gerar o grafo, o PageRank de Tópicos pode retornar uma lista dos 10 autores com mais postagens e comentários sobre o tópico, comparado com o PageRank Global. Somente no tópico *Filme* as duas listas têm uma quantidade

parecida de postagens e comentários, nos demais casos a abordagem por tópicos leva grande vantagem.

Esse resultado mostra que a utilização da abordagem do PageRank por Tópicos, como sugerido por Haveliwala (2002) para sites e links na internet, pode ser utilizada como algoritmo de descoberta de autoridades também na blogosfera.

4.4.1 Correlação das Listas

Outra análise interessante a ser feita sobre as listas geradas pelas duas abordagens do PageRank é a correlação delas com o ordenamento feito por outras métricas. Foram comparadas as listas com a ordenação dos autores pela visualização (número de visitas ao perfil), número de comentários recebidos de outros usuários e número de postagens criadas pelo usuário. A Tabela 4.6 mostra a correlação de Spearman entre essas diversas listas. Quando utilizada a lista gerada pelo PageRank Global na correlação é usada a legenda *GPR*, quando é usado o PageRank de Tópicos a legenda é *TPR*.

Tabela 4.6: Correlação de Spearman

Tag	Visualizações	Comentários	Postagens	GPR
Moda GPR	**0.8815**	0.8028	0.8084	1.0000
Moda TPR	0.1021	**0.9835**	**0.9931**	0.8155
Música GPR	**0.8457**	0.2567	0.2448	1.0000
Música TPR	-0.1726	**0.9966**	**0.9916**	0.2553
Decoração GPR	**0.8058**	0.6475	0.6625	1.0000
Decoração TPR	0.1882	**0.9268**	**0.9648**	0.6931
Filme GPR	**0.8868**	0.1080	0.1043	1.0000
Filme TPR	-0.0199	**0.9945**	**0.9942**	0.1123
Livro GPR	**0.8547**	0.7011	0.7141	1.0000
Livro TPR	0.1582	**0.9255**	**0.9548**	0.7500
Fotografia GPR	**0.8507**	0.2432	0.2388	1.0000
Fotografia TPR	0.2090	**0.9988**	**0.9974**	0.2440

Fonte: tabela elaborada pelo autor com base nos experimentos realizados.

Como pode ser observado na Tabela 4.6, e marcado em negrito, o PageRank Global tem forte relação com o número de visualizações recebidas pelos usuários, mesmo elas tendo sido feitas durante todo o tempo de existência do usuário no Blogspot, e o PageRank Global criado sobre a coleta nos meses de maio a julho de 2012. Oposto a isso, a lista gerada pelo PageRank de Tópicos tem uma relação muito próxima com a ordenação das postagens e comentários.

Essas observações de correlação mostram duas vertentes na pesquisa utilizando o PageRank: se forem consideradas as visualizações como *gold-standard* de autoridade o PageRank Global é suficiente para se utilizar como métrica de autoridade. Em contrapartida, utilizando o conteúdo gerado diretamente (as postagens) ou indiretamente (os comentários recebidos) como *gold-standard*, somente obter qual usuário recebe mais comentários é o suficiente para se ter a melhor lista. Como o PageRank de Tópicos tem uma forte relação com os comentários recebidos pelos usuários, se não for necessário ter uma lista tão precisa de autoridades, somente ordenar os usuários pelo número de comentários recebidos já é suficiente para obter uma boa ordenação de autoridades.

Essa última observação nos mostra que é possível comparar duas plataformas de blogs diferentes, como Blogspot e Wordpress. Somente utilizando a informação de comentários

recebidos pelos autores nas postagens, pode-se geral uma lista confiável de autoridades.

4.4.2 Correlação de Visualizações e Comentários

Outro dado coletado que pode ser comparado com a correlação de Spearman é a quantidade de visitas recebidas pelo perfil do usuário e o número de comentários recebidos. Nesse caso, compara-se todas as visitas e todos os comentários recebidos durante todo o período do usuário no Blogspot. Na Figura 4.2 os autores foram separados por ano de entrada na plataforma Blogspot para fazer a comparação entre o ordenamento dos usuários pelas visitas e o ordenamento pelos comentários.

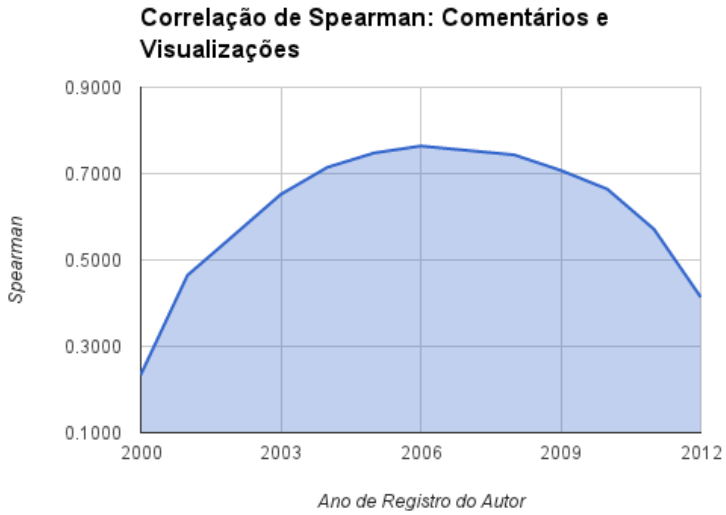


Figura 4.2: Correlação de Spearman sobre popularidade dos autores.
Fonte: figura elaborada pelo autor

Como pode ser observado pelo gráfico na Figura 4.2, a correlação entre o ordenamento de visitas e comentários sempre é positiva, mostrando que existe uma correlação entre as visitas recebidas e os comentários. Mesmo que em alguns casos essa correlação seja menor, como em 2002 e 2003, ainda assim é uma correlação positiva.

De qualquer forma, como as visualizações não são uma informação disponível pelos blogs em geral, pois somente é possível obtê-las com o servidor de hospedagem do usuário, não é interessante considerá-las como forma de ranqueamento dos autores de blogs. Nesse caso, a quantidade de comentários recebidos é observada novamente como uma boa forma de ranqueamento de autoridades.

4.5 Características dos Grafos

Uma importante análise sobre as redes sociais é a topologia que elas formam com relação aos usuários e relacionamentos entre eles. Na Tabela 4.7[1] são mostradas algumas medidas sobre os grafos gerados entre os comentários e blogueiros dos tópicos selecionados.

Considerando todos os comentários e blogueiros coletados, o grafo geral é mais disperso, tendo um diâmetro de 33 nodos, mas mantendo uma média de grau semelhante ao

[1]Não foi possível calcular a modularidade do grafo geral por falta de processamento, mas se estima que o valor seja de aproximadamente 0,2 por ser um grafo mais esparso que os demais.

Tabela 4.7: Medidas dos Grafos de Tópicos

Tag	Grau Médio	Diâmetro	Modularidade
Todos	5,708	33	?
Moda	5,780	12	0,514
Música	3,149	16	0,690
Decoração	4,025	14	0,639
Filme	2,777	14	0,779
Livro	3,698	17	0,633
Fotografia	2,364	12	0,852

Fonte: tabela elaborada pelo autor com base nos dados coletados.

obtido nos grafos de tópicos. A modularidade (BLONDEL et al., 2008) é um valor entre 0 e 1 que mede a estrutura da rede, mostrando a força de divisão da rede em módulos ou agrupamentos. Para modularidade, todos os tópicos têm um valor semelhante perto de 0,6 mostrando que os grafos têm agrupamentos bem definidos por grande número de conexões, mas também com alguns nodos esparsos. Um exemplo de grafo pode ser visto no Apêndice B desta dissertação.

Outras características observadas para análise de redes complexas é a definição de rede pequeno mundo ou livre de escala. As redes pequeno mundo se caracterizam por ter comportamento semelhante ao mundo real: os nodos não podem estar, na média, muito distantes um dos outros, mostrando que as pessoas no mundo não estão tão distantes umas das outras. Esse comportamento foi observado e demonstrado no trabalho de Watts e Strogatz (1998), que indicam que as redes pequeno mundo têm a distância média similar ao logaritmo dos nodos.

Já as redes livres de escala se caracterizam por ter uma distribuição, de arestas por nodos, semelhante à distribuição de Lei de Potências. Essas redes têm alguns nodos, denominados *hubs*, que têm seu grau bem acima da média e concentram as arestas. Elas também têm agrupamentos mais afastados entre si e muitos nodos com poucas arestas. A rede livre de escala foi definida no trabalho de Cohen e Havlin (2003), e seu cálculo de semelhança foi demonstrado através do uso da Lei de Potências por Clauset (2009), chamado Fator de Lei de Potências.

Tabela 4.8: Classificação dos Grafos de Tópicos

Tag	Log(Nodos)	Distância média	Fator de Lei de Potências
Todos	5,7009	30,687	0,9800
Moda	4,1657	4,289	0,0108
Música	4,0592	5,057	0,0084
Decoração	3,9825	4,995	0,0556
Filme	3,9897	5,109	0,0000
Livro	3,9910	4,774	0,1600
Fotografia	3,9223	3,989	0,4644

Fonte: tabela elaborada pelo autor com base nos dados coletados.

Na Tabela 4.8 são caracterizados cada um dos grafos dos tópicos e o grafo geral de comentários. Como é possível notar, os grafos, quando subdivididos nos tópicos, têm uma forte relação com as redes pequeno-mundo (o logaritmo do número de nodos é muito semelhante à distância média entre eles). Isso mostra que os tópicos têm grupos bem

definidos e agrupamentos mais fortes. Considerando o grafo geral dos comentários e autores, nota-se uma forte relação com a Lei de Potência, caracterizando uma rede livre de escala, que tem poucos autores que recebem muitos comentários, acima da média, e a maioria recebe poucos comentários, ficando mais esparsos na rede.

4.6 Resumo do Capítulo

Com a coleta de nove milhões de postagens, foi possível fazer diversos experimentos e validações sobre os dados. Os resultados mostraram que a abordagem de tópicos sobre o PageRank pode ser utilizada para a descoberta de autoridades na blogosfera. Além disso, o uso de tópicos no PageRank mostra-se mais efetivo que a sua versão original, assim como foi demonstrado por Haveliwala (2002).

Analisando a topologia dos grafos, pode-se observar uma tendência na correlação entre o PageRank e o *inDegree* (links de entrada) dos nodos: para os grafos livre de escala, o PageRank tem baixa relação com o *inDegree*, entretanto, os grafos pequeno-mundo tem alta correlação do PageRank com o *inDegree* dos nodos. Essa observação deve ser analisada com mais profundidade, pois, se essa tendência se repetir com outros grafos pequeno-mundo, mostra-se desnecessário o cálculo do PageRank, podendo usar o valor do *inDegree* como indicador de autoridade.

5 CONCLUSÃO

A blogosfera é uma fonte rica para mineração de informações latentes e de comportamento dos usuários. Os trabalhos realizados sobre dados de blogueiros mostram resultados interessantes na descoberta de conhecimento sobre essa rede. Descobrir autoridades sobre a blogosfera é um dos trabalhos mais realizados, como pode ser percebido no Capítulo **??**, pois isso ajuda na identificação de autores importantes e na recomendação de conteúdo mais relevante para os usuários.

Foram desenvolvidas técnicas que auxiliam na descoberta de autoridades importantes para a rede social de blogueiros do Brasil, em particular os da plataforma Blogspot.

5.1 Objetivos e Hipótese

A proposta da dissertação trata da possibilidade de identificar autoridades de tópicos em blogs utilizando o algoritmo PageRank e separando o grafo de tópicos dos usuários pelas *tags* utilizadas por eles. Os experimentos realizados mostram que esta abordagem tem bons resultados para a descoberta de autoridades na blogosfera, podendo inclusive ser substituída pelo valor do *inDegree* dos nodos. O objetivo deste trabalho foi mostrar, com uma nova base de dados coletada de blogueiros brasileiros, que a técnica funciona sobre o grafo de postagens, considerando os comentários como relacionamento entre os usuários.

5.2 Contribuições

Foi proposta uma abordagem sobre tópicos para obter melhores resultados sobre o algoritmo PageRank. Essa abordagem foi baseada na divisão do grafo da blogosfera, sobre os comentários dos autores, em subgrafos dos tópicos mais escritos na plataforma Blogspot. Aplicando o PageRank neste subgrafos, ao invés de aplicar ao grafo todo, foi obtida uma lista mais significativa de autoridades sobre os tópicos de interesse. Acredita-se que essa abordagem pode ser usada em outros problemas de grafos direcionados, pois diminui a complexidade de computação na análise das ligações entre os nodos.

Os dados coletados também são considerados uma das contribuições deste trabalho, pois já são utilizados em outras pesquisas sobre a blogosfera realizadas na universidade. Também foi realizada uma importante entrevista com blogueiros reais para melhor caracterizar os dados, comparando suas respostas com os dados obtidos na coleta.

A base de dados coletada, o código-fonte utilizado e a entrevista realizada com os autores podem ser encontrados na página principal do autor[1] para uso livre.

[1] http://www.inf.ufrgs.br/~hdpsantos/

5.3 Trabalhos Futuros

Muitos trabalhos podem ser realizados sobre os dados coletados e a entrevista realizada, alguns deles estão relacionados nos trabalhos revisados na Seção **??**. A melhoria direta que pode ser feita nos dados é a análise das postagens sobre os tópicos escritos utilizando técnicas de agrupamento de documentos. Métodos como o TF-IDF, LDA, NMF ou outro estado-da-arte podem gerar grafos mais consistentes sobre os tópicos selecionados.

Além disso, os usuários que não proveem suas informações de preferência nos perfis podem ter elas inferidas fazendo a análise baseada em comportamento mútuo de usuários semelhantes. Outro trabalho futuro pode considerar o questionário realizado sobre os blogueiros para identificar os tópicos mais importantes para eles. Isso é importante para medir a autoridade do usuário baseada na qualidade e relevância do seu conteúdo publicado nos blogs.

A análise temporal das autoridades em tópicos também pode ser observada considerando a evolução das autoridades ao longo dos meses do ano. Os usuários que recebem muitos comentários somente num curto período de tempo e depois não continuam com esse comportamento, não devem ser considerados boas autoridades. Uma boa autoridade vai permanecer autoridade durante muito tempo, portanto se o usuário, ao longo do ano, receber uma média alta de comentários, ele pode ser considerado uma autoridade estável no assunto.

REFERÊNCIAS

ADAMIC, L. A.; GLANCE, N. The political blogosphere and the 2004 U.S. election: divided they blog. In: LINK DISCOVERY, 3., New York, NY, USA. **Proceedings...** ACM, 2005. p.36–43. (LinkKDD '05).

ADAMS, B.; PHUNG, D.; VENKATESH, S. Discovery of latent subcommunities in a blog's readership. **ACM Trans. Web**, New York, NY, USA, v.4, p.12:1–12:30, July 2010.

AGARWAL, N.; LIU, H. Blogosphere: research issues, tools, and applications. **SIGKDD Explor. Newsl.**, New York, NY, USA, v.10, n.1, p.18–31, May 2008.

ALI-HASAN, N.; ADAMIC, L. A. Expressing social relationships on the blog through links and comments. In: IN PROCEEDINGS OF THE 1ST ANNUAL MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS. **Anais...** [S.l.: s.n.], 2007.

ANDRZEJEWSKI, D.; BUTTLER, D. Latent topic feedback for information retrieval. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 17., New York, NY, USA. **Proceedings...** ACM, 2011. p.600–608. (KDD '11).

BAHMANI, B.; CHAKRABARTI, K.; XIN, D. Fast personalized PageRank on MapReduce. In: MANAGEMENT OF DATA, 2011., New York, NY, USA. **Proceedings...** ACM, 2011. p.973–984. (SIGMOD '11).

BERSCHEID, E.; REIS, H. T. Attraction and close relationships. In: GILBERT, D. T.; FISKE, S. T.; LINDZEY, G. (Ed.). **The handbook of social psychology**: vol. 2. New York: McGraw-Hill, 1998. v.2, p.193–281.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **J. Mach. Learn. Res.**, [S.l.], v.3, p.993–1022, Mar. 2003.

BLONDEL, V. D. et al. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, [S.l.], v.2008, n.10, p.P10008+, Oct. 2008.

CALAIS GUERRA, P. H. et al. From bias to opinion: a transfer-learning approach to real-time sentiment analysis. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 17., New York, NY, USA. **Proceedings...** ACM, 2011. p.150–158. (KDD '11).

CHAN, S.; PON, R. K.; CARDENAS, A. F. Visualization and Clustering of Author Social Networks. In: DISTRIBUTED MULTIMEDIA SYSTEMS CONFERENCE, Grand Cayon, Arizona, USA. **Anais...** [S.l.: s.n.], 2006.

CLAUSET, A.; SHALIZI, C. R.; NEWMAN, M. E. J. Power-Law Distributions in Empirical Data. **SIAM Review**, [S.l.], v.51, n.4, p.661–703, 2009.

Cohen, R.; Havlin, S. Scale-Free Networks Are Ultrasmall. **Phys. Rev. Lett.**, [S.l.], v.90, p.058701, 2003.

CUNHA RECUERO, R. da. Information flows and social capital in weblogs: a case study in the brazilian blogosphere. In: ACM CONFERENCE ON HYPERTEXT AND HYPERMEDIA, New York, NY, USA. **Proceedings...** ACM, 2008. p.97–106. (HT '08).

DEAN, J.; GHEMAWAT, S. MapReduce: simplified data processing on large clusters. In: SYMPOSIUM ON OPEARTING SYSTEMS DESIGN & IMPLEMENTATION - VOLUME 6, 6., Berkeley, CA, USA. **Proceedings...** USENIX Association, 2004. p.10–10. (OSDI'04).

GOOGLE. **Google Insights**. [Online; accessed 15-April-2012], http://www.google.com/insights/search/q=blogspot,wordpressgeo=BR.

HAVELIWALA, T. H. Topic-sensitive PageRank. In: WORLD WIDE WEB, 11., New York, NY, USA. **Proceedings...** ACM, 2002. p.517–526. (WWW '02).

JIANG, M.; ARGAMON, S. Finding Political Blogs and Their Political Leanings. In: TEXT MINING WORKSHOP AT THE SIAM. **Anais...** [S.l.: s.n.], 2008.

KELLY, D.; FU, X.; SHAH, C. Effects of position and number of relevant documents retrieved on users' evaluations of system performance. **ACM Trans. Inf. Syst.**, New York, NY, USA, v.28, n.2, p.9:1–9:29, June 2010.

LAM, M. I.; GONG, Z.; MUYEBA, M. A method for web information extraction. In: ASIA-PACIFIC WEB CONFERENCE ON PROGRESS IN WWW RESEARCH AND DEVELOPMENT, 10., Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2008. p.383–394. (APWeb'08).

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **Nature**, [S.l.], v.401, n.6755, p.788–791, Oct. 1999.

LI, Y.-M.; LAI, C.-Y.; CHEN, C.-W. Identifying bloggers with marketing influence in the blogosphere. In: INTERNATIONAL CONFERENCE ON ELECTRONIC COMMERCE, 11., New York, NY, USA. **Proceedings...** ACM, 2009. p.335–340. (ICEC '09).

LIN, C.-L.; TANG, H.-L.; KAO, H.-Y. Utilizing Social Relationships for Blog Popularity Mining. In: ASIA INFORMATION RETRIEVAL SYMPOSIUM ON INFORMATION RETRIEVAL TECHNOLOGY, 5., Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2009. p.409–419. (AIRS '09).

LIN, Y.-R. et al. Blog Community Discovery and Evolution Based on Mutual Awareness Expansion. In: IEEE/WIC/ACM INTERNATIONAL CONFERENCE ON WEB INTELLIGENCE, Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2007. p.48–56. (WI '07).

LIU, X. et al. Identifying topic experts and topic communities in the blogspace. In: DATABASE SYSTEMS FOR ADVANCED APPLICATIONS - VOLUME PART I, 16., Berlin, Heidelberg. **Proceedings...** Springer-Verlag, 2011. p.68–77. (DASFAA'11).

NONNECKE, B.; PREECE, J. Lurker demographics: counting the silent. In: SIGCHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS, New York, NY, USA. **Proceedings...** ACM, 2000. p.73–80. (CHI '00).

ORENGO, V.; HUYCK, C. A Stemming Algorithmm for the Portuguese Language. **String Processing and Information Retrieval, International Symposium on**, Los Alamitos, CA, USA, v.0, p.0186, 2001.

PAGE, L. et al. **The PageRank Citation Ranking**: bringing order to the web. 1998.

PAL, A.; COUNTS, S. Identifying topical authorities in microblogs. In: ACM INTERNATIONAL CONFERENCE ON WEB SEARCH AND DATA MINING, New York, NY, USA. **Proceedings...** ACM, 2011. p.45–54. (WSDM '11).

SHANNON, C. E. A mathematical theory of communication. **SIGMOBILE Mob. Comput. Commun. Rev.**, New York, NY, USA, v.5, n.1, p.3–55, Jan. 2001.

SHEN, D. et al. Latent Friend Mining from Blog Data. In: SIXTH INTERNATIONAL CONFERENCE ON DATA MINING, Washington, DC, USA. **Proceedings...** IEEE Computer Society, 2006. p.552–561. (ICDM '06).

SONG, X. et al. Identifying opinion leaders in the blogosphere. In: ACM CONFERENCE ON CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, New York, NY, USA. **Proceedings...** ACM, 2007. p.971–974. (CIKM '07).

SPEARMAN, C. The proof and measurement of association between two things. **The American journal of psychology**, [S.l.], v.15, p.72–101, 1904.

TSAI, F. S. A tag-topic model for blog mining. **Expert Syst. Appl.**, Tarrytown, NY, USA, v.38, n.5, p.5330–5335, May 2011.

WASSERMAN, S.; FAUST, K. **Social Network Analysis**: methods and applications. 1.ed. [S.l.]: Cambridge University Press, 1994. n.8. (Structural analysis in the social sciences).

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, [S.l.], v.393, n.6684, p.440–442, June 1998.

XU, W.; LIU, X.; GONG, Y. Document clustering based on non-negative matrix factorization. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMAION RETRIEVAL, 26., New York, NY, USA. **Proceedings...** ACM, 2003. p.267–273. (SIGIR '03).

ZHANG, Q. et al. Template-independent wrapper for web forums. In: ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 32., New York, NY, USA. **Proceedings...** ACM, 2009. p.794–795. (SIGIR '09).

APÊNDICE A <INFRAESTRUTURA>

A infraestrutura desenvolvida para processar a coleta e armazenar os dados é composta por um servidor Dell Xeon 8 cores com 24GB de RAM. Nele foram criadas sete máquinas virtuais de tamanhos diferentes, como é mostrado na Figura 5.1. Foram criadas quatro máquinas mais leves que servem somente para coleta de dados: 1 processador virtual e 512MB de RAM. Essas máquinas executam o processa de busca de autores e consulta do perfil dos autores.

Nas máquinas mais pesadas, as três que mantêm as informações coletadas no banco de dados, foi alocado mais processamento e memória: 4 processadores virtuais e 8GB de RAM. Essas máquinas armazenam as informações e fazem o processo de coleta de postagens na API do Blogger.

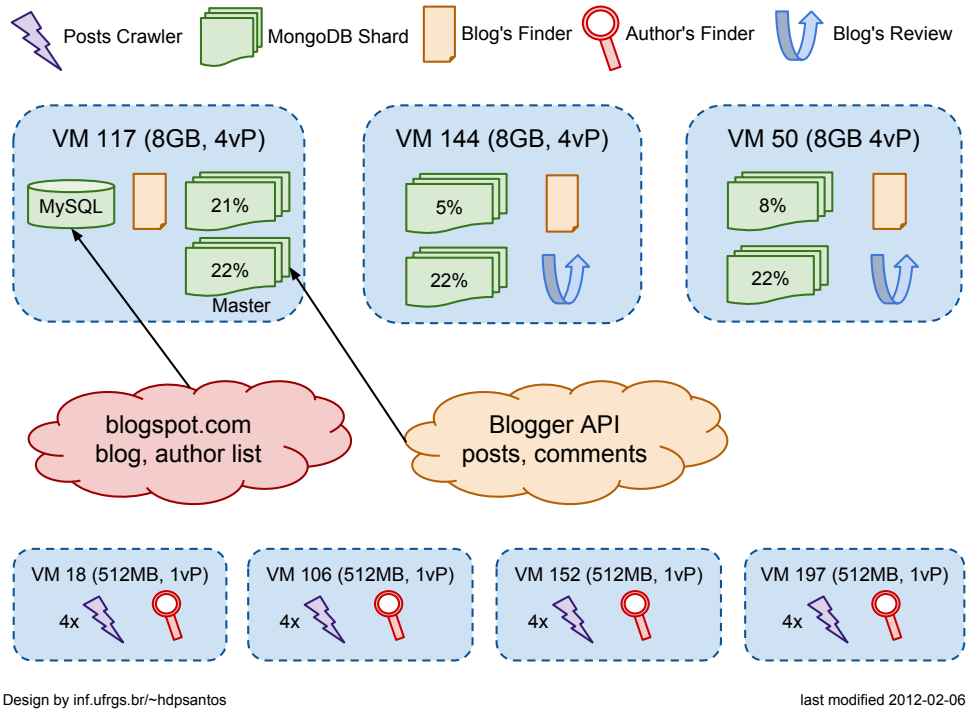


Figura 5.1: Infraestrutura construída e utilizada pelos coletores.
Fonte: figura elaborada pelo autor

Abaixo segue a explicação de cada processo de coleta:

- *Posts Crawler*: Coletor sobre o Blogger API de postagens e comentários sobre a lista de blogs encontrados;

- *MongoDB Shard*: Instâncias do banco de dados que contém parte das postagens;
- *Blog's Finder*: Coletor sobre as páginas dos autores descobertos pelo processo seguinte;
- *Author's Finder*: Coletor que percorre a lista de blogueiros brasileiros no Blogspot;
- *Blog's Review*: Processo que verifica se todas as postagens dos blogs foram coletadas corretamente.

Como foi destacado na seção 3.4.1, os dados das postagens foram armazenados em um banco de dados orientado a documentos, com possibilidade de programação paralela e distribuída pelas diversas partes, *shards*, das máquinas virtuais. As informações estáticas de perfil dos autores e endereços dos blogs foram armazenadas em banco de dados relacional. Nas máquinas mais leves, o processo *Posts Crawler* é executado em quatro *threads*.

APÊNDICE B <GRAFO DE COMUNIDADES>

Abaixo segue um exemplo de como se distribuiu os usuários da blogosfera considerando os comentários como arestas. Nesse grafo as diversas subcomunidades existentes entre os usuários foram coloridas aleatoriamente e os nodos dimensionados de acordo com o *indegree* do vértice. Esse grafo contém 9.755 usuários e 13.547 comentários ligando-os. Nele, está representada a comunidade de usuários que interagiram sobre a *tag* filme. Nesse caso foram detectadas 468 subcomunidades diferentes, de acordo com o algoritmo *Fast unfolding of communities in large networks* (BLONDEL et al., 2008).

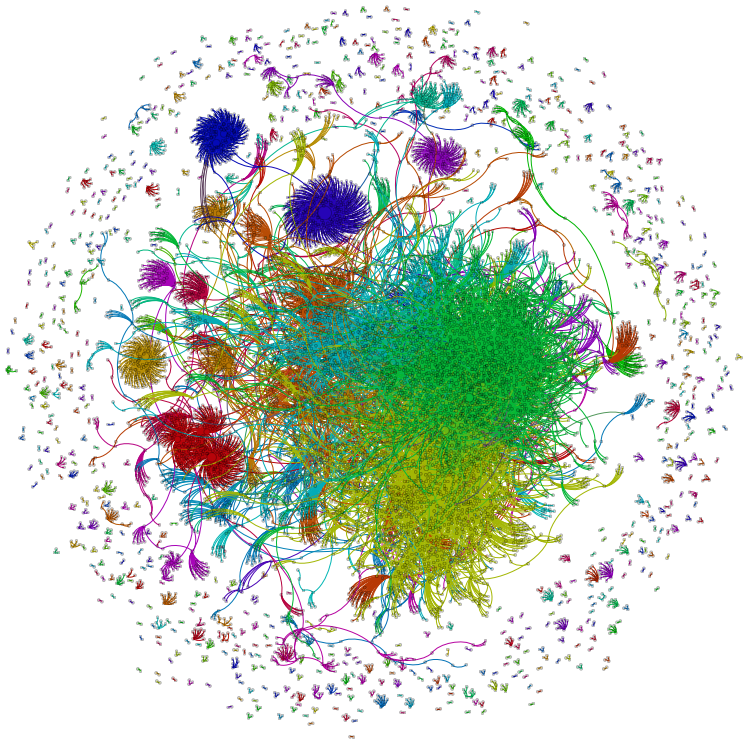


Figura 5.2: Grafo da comunidade de usuários da *tag* filme.
Fonte: figura elaborada pelo autor utilizando o algoritmo Force Atlas 2 de Mathieu Jacomy no programa Gephi

Na figura 5.2 pode ser visto que os usuários mais comentados estão envolvidos por diversas comunidades sobre o tema, enquanto existem usuários que pouco interagem com outros, os que estão orbitando o grafo. Outras estatísticas interessantes do grafo são o diâmetro de 14 e a distância média entre os nodos de 5. Isso mostra que esse grafo pertence à classificação de pequeno-mundo, pois os nodos, em geral, estão próximos.

APÊNDICE C <TOPIC BOOTSTRAP>

Outra abordagem de personalização do PageRank que foi tentada para melhor o desempenho do algoritmo na descoberta de autoridades em tópicos foi a inicialização do valor de cada nodo com heurísticas diferentes. Foram inicializados os nodos com os seus respectivos números de postagens sobre o tópico e comentários sobre o tópico. A validação foi dada sobre as visualizações do perfil dos primeiros 10 autores ordenados como autoridades.

Esse é o resultado da soma de visualizações dos usuários presentes na lista Top10 com cada uma das inicializações:

- Rb: *Rank Baseline* (PageRank)
- R10: *Rank* do Algoritmo Proposto com peso 1 nas Postagens e 0 em Comentários
- R55: *Rank* do Algoritmo Proposto com peso 0,5 nas Postagens e 0,5 em Comentários

Tabela 5.1: Resultado do PageRank com Topic Bootstrap

Tag	**Rb**	**R10**	**R50**
cinema	**72,070**	41,165	40,269
moda	**55,191**	51,201	53,059
política	**25,857**	10,976	25,663
educação	**90,554**	12,022	15,469
futebol	**11,638**	8,942	8,705
saúde	**69,142**	11,020	68,810

Fonte: tabela elaborada pelo autor com base nos experimentos realizados.

Como pode ser observado, nos dados em negrito, o PageRank consegue um desempenho melhor, na maioria dos casos, quando iniciado normalmente, sem ponderar os valores de conteúdo publicado pelos autores. Portanto, essa tentativa foi descartada na dissertação.

APÊNDICE D <ENTREVISTA COM AUTORES>

Para fazer uma validação dos perfis de usuários coletados e entender melhor esses usuários da plataforma Blogspot, foi realizada uma entrevista com 24 mil usuários, escolhidos por quantidade de postagens. Até agora houve 4.300 respostas (i.e., 17%). Se considerarmos a lista de 700 mil autores coletados no conjunto de dados como uma população infinita e tendo uma distribuição binomial, esse questionário dá 99% de nível de confiança e 2,04% de margem de erro. Algumas questões propostas no questionário foram criadas para certificar a autenticidade dos usuários. Campos que já existiam no perfil do usuário na plataforma, como sexo, cidade e ocupação, foram utilizados para comparar se as respostas eram verídicas. Apenas 1% dos usuários não respondeu de acordo com o seu próprio perfil nos três campos e esses foram desconsiderados nas estatísticas das respostas. A principal motivação da entrevista era verificar se os tópicos escritos pelos usuários, identificados na coleta, eram os mesmos definidos pelos autores, pois alguns usuários definem em seus perfis escreverem sobre um tópico e escrevem sobre outro.

A análise dos resultados mostra que apenas 43% dos usuários escreve em seus blogs conteúdo relacionado com o interesse definido em seu perfil. Isso significa que os perfis dos usuários não podem ser usados como fonte para deduzir o conteúdo publicado em seus blogs. Outra importante questão da entrevista era sobre a frequência de comentários realizados pelos usuários. Como é mostrado na Figura 5.3, a maioria dos usuários geralmente comenta nas postagens que leem, mas os dados coletados mostram uma proporção inversa. Esse resultado pode representar uma falsa visão sobre o comportamento que o autor tem sobre si ou uma distribuição mal obtida pelo coletor de blogs.

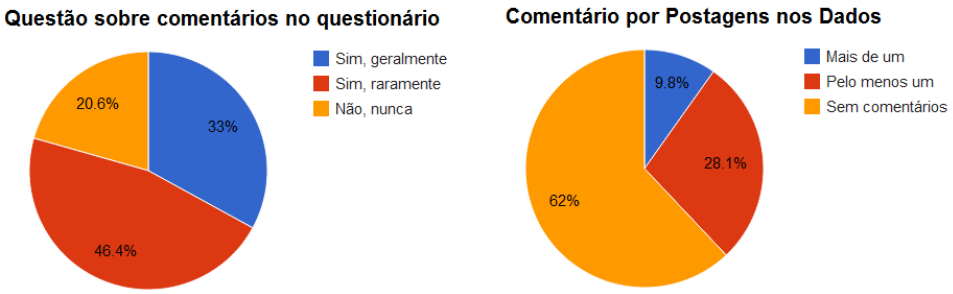
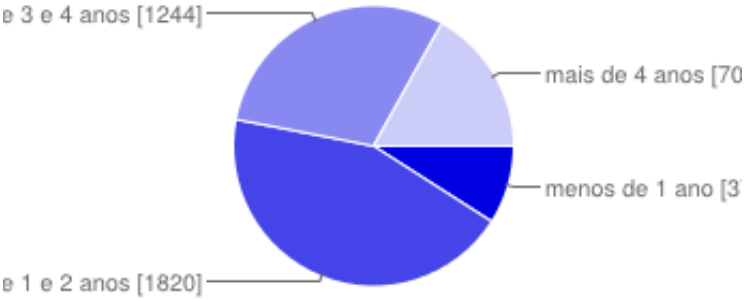


Figura 5.3: Frequência de comentários respondida no questionário e a encontrada nos dados coletados.
Fonte: figura elaborada pelo autor

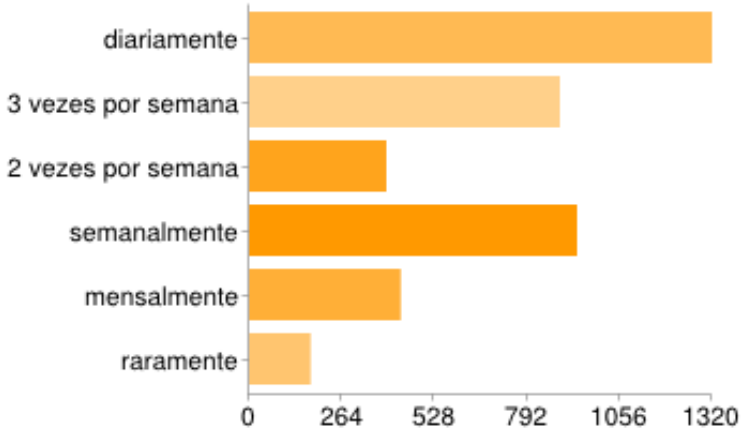
O restante das perguntas e a estatística das respostas estão nas páginas seguintes.

Há quanto tempo você mantém esse blog?



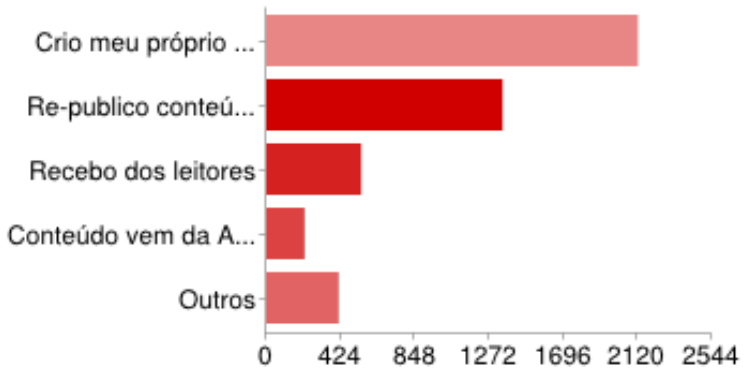
menos de 1 ano	374	9%
entre 1 e 2 anos	1820	44%
entre 3 e 4 anos	1244	30%
mais de 4 anos	700	17%

Qual o período de atualização do blog?



diariamente	1320	32%
3 vezes por semana	886	21%
2 vezes por semana	391	9%
semanalmente	935	23%
mensalmente	432	10%
raramente	175	4%

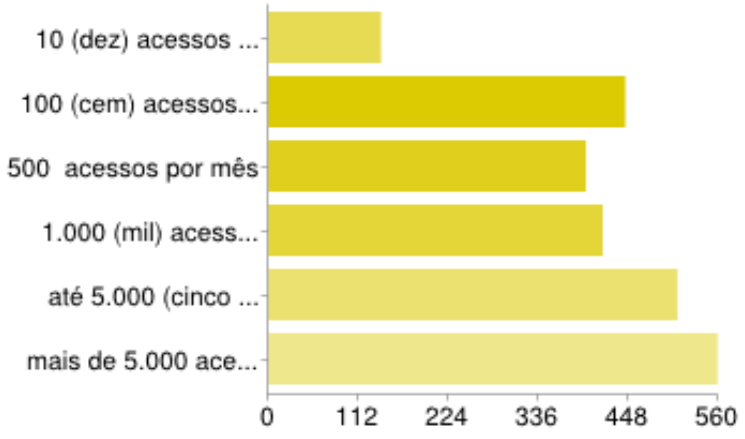
Com relação a autoria do conteúdo de seu blog:



Crio meu próprio conteúdo	2121	86%
Re-publico conteúdo de terceiros (outros blogs, jornais, etc)	1348	55%
Recebo dos leitores	541	22%
Conteúdo vem da Assessoria de imprensa	218	9%
Outros	413	17%

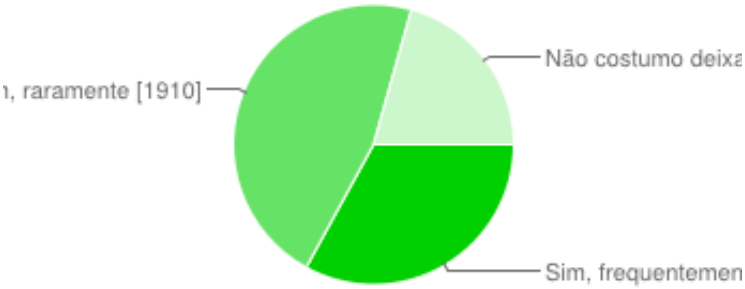
As pessoas podem selecionar mais de uma opção, assim percentagens podem somar mais de 100%.

Quantidade média de acessos do seu blog por mês?



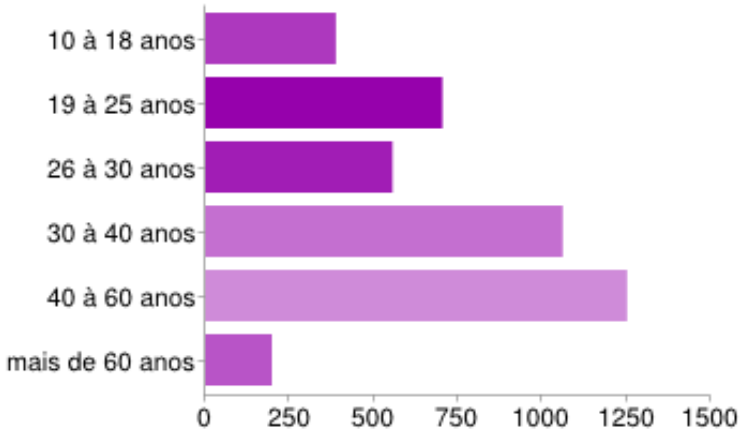
10 (dez) acessos por mês	140	3%
100 (cem) acessos por mês	444	11%
500 acessos por mês	395	10%
1.000 (mil) acessos por mês	416	10%
até 5.000 (cinco mil) acessos por mês	509	12%
mais de 5.000 acesso por mês	559	13%

Costuma deixar comentários nos blogs que segue?



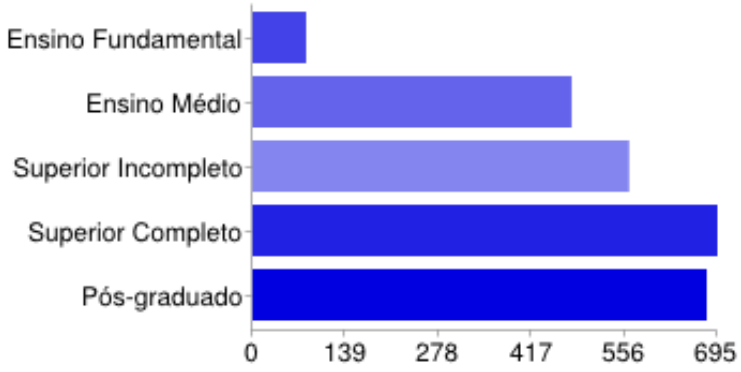
Sim, frequentemente	1358	33%
Sim, raramente	1910	46%
Não costumo deixar comentários	850	20%

Qual sua idade?



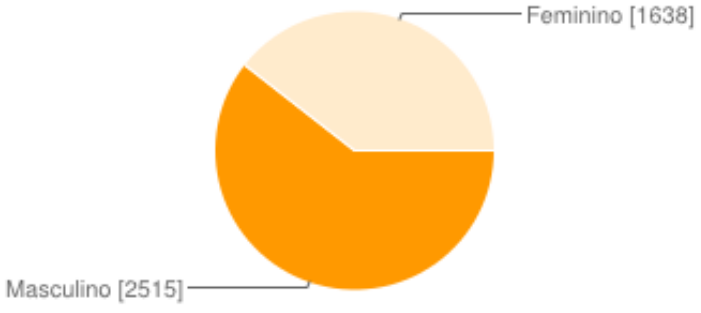
10 à 18 anos	387	9%
19 à 25 anos	703	17%
26 à 30 anos	555	13%
30 à 40 anos	1060	26%
40 à 60 anos	1251	30%
mais de 60 anos	197	5%

Nivel de Escolaridade



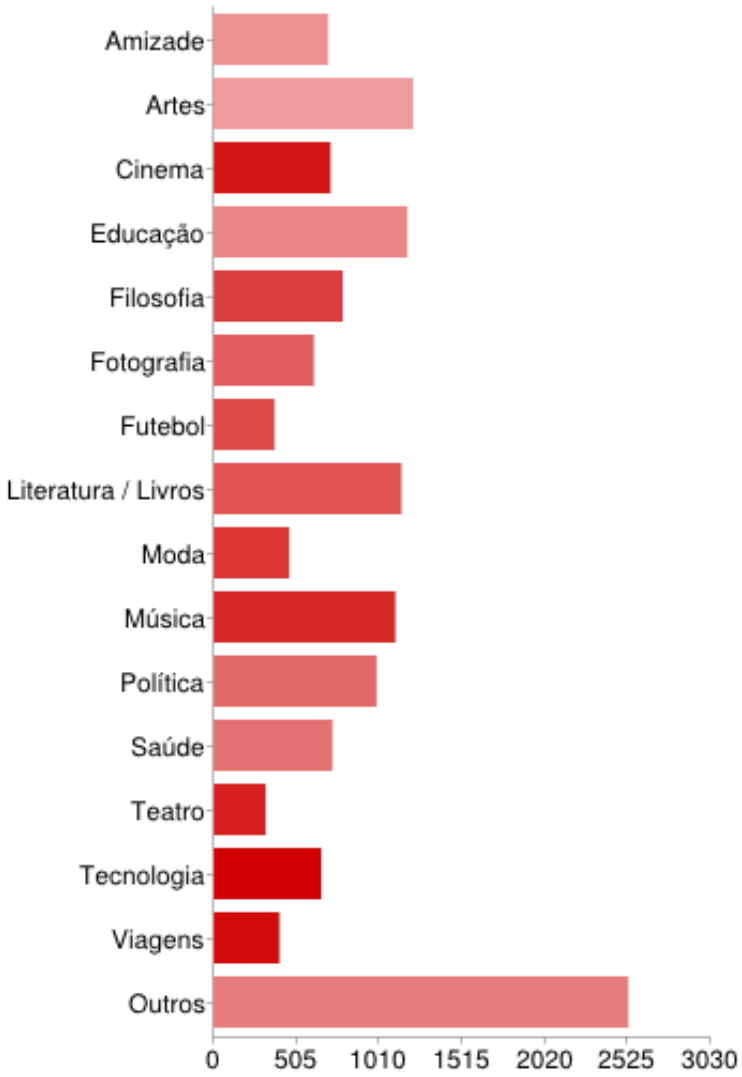
Ensino Fundamental	80	2%
Ensino Médio	477	11%
Superior Incompleto	563	14%
Superior Completo	695	17%
Pós-graduado	679	16%

Qual seu sexo?



Masculino	2515	61%
Feminino	1638	39%

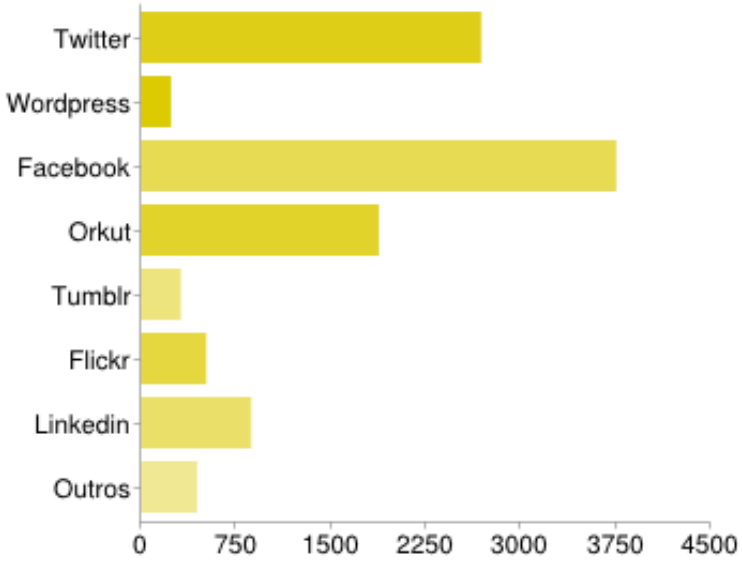
Sobre qual(is) tópico(s) você mais escreve?



Amizade	696	17%
Artes	1214	29%
Cinema	710	17%
Educação	1179	28%
Filosofia	785	19%
Fotografia	609	15%
Futebol	369	9%
Literatura / Livros	1143	28%
Moda	459	11%
Música	1106	27%
Política	991	24%
Saúde	723	17%
Teatro	315	8%
Tecnologia	654	16%
Viagens	399	10%
Outros	2526	61%

As pessoas podem selecionar mais de uma opção, assim percentagens podem somar mais de 100%.

Utiliza outra rede social?



Twitter	2685	67%
Wordpress	235	6%
Facebook	3752	94%
Orkut	1877	47%
Tumblr	313	8%
Flickr	514	13%
Linkedin	869	22%
Outros	442	11%

As pessoas podem selecionar mais de uma opção, assim percentagens podem somar mais de 100%.