Detecção de Conteúdo Relevante e Usuários Influentes no Twitter

Hérico Valiati[1], **Arlei Silva**[1], **Sara Guimarães**[1], **Wagner Meira Jr.**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)

{herico,arlei,sara,meira}@dcc.ufmg.br

Abstract. *Social networks are an increasingly important media for information and member influence. The main target of this paper is to determine which users are influential and to identify relevant content - i.e. ranking user groups and the content spread by them. We propose a novel technique that is based on an intuitive and circular definition of relevance and influence. We describe the proposed technique in detail, as well as its efficient implementation. In order to validate it, we considered the task of recommending users and content to users. We used two real data sets extracted from Twitter and our results show that our technique outperforms by 37% a collaborative filtering strategy, while both influential users and relevant content are qualitatively better.*

Resumo. *Redes sociais têm desempenhado um papel cada vez mais fundamental como um meio para a disseminação de informações, idéias e influências entre os seus membros. O problema alvo deste artigo é determinar tanto usuários influentes quanto conteúdo relevante, ou seja, ordenar tanto grupos de usuários quanto o conteúdo por eles disseminado. Propomos uma nova técnica que se baseia em uma definição intuitiva e circular de relevância e influência. Essa técnica é descrita em detalhes, assim como a sua implementação eficiente. Ela foi validada utilizando duas bases de dados reais do Twitter para fins de recomendação. Os resultados obtidos mostram que a técnica proposta apresenta ganhos de 37% quando comparada a um método de filtragem colaborativa, ao mesmo tempo que tanto usuários influentes quanto conteúdo relevante se mostram superiores qualitativamente.*

1. Introdução

Redes sociais têm desempenhado um papel cada vez mais fundamental como meio para a disseminação de informação, idéias e influência entre os seus membros e mesmo além das redes, tendo em vista a crescente convergência que pode ser observada entre as várias mídias. Em suma, redes sociais têm se mostrado um importante mecanismo para atingir grandes parcelas da população, influenciando a opinião pública, a adoção de inovações, a publicidade de novos produtos ou marcas. Entretanto, identificar as características do conteúdo relevante, assim como dos usuários que os tornam populares ou impactantes é uma questão de pesquisa que vem sendo investigada sob várias perspectivas. Por exemplo, o problema de otimizar a disseminação de conteúdo e explorá-lo para fins de marketing já foi provado como NP-árduo, assim como alguns de seus variantes. Vários autores também se dedicaram a identificar e caracterizar as redes de difusão de informação.

O problema alvo deste artigo é determinar tanto usuários influentes quanto conteúdo relevante, ou seja, ordenar tanto grupos de usuários quanto o conteúdo por eles disseminado. Há vários desafios ao realizar essas tarefas: (i) a rede de disseminação de informação é pouco observável, dificultando a detecção exata do processo de disseminação e, portanto, da relevância do conteúdo e da influência dos usuários; (ii) quando é possível observar essa difusão e identificar os principais fatores que a explicam, em geral falamos de eventos de sucesso, de alto impacto, que são raros e cujas características podem não se generalizar para outros tipos de influência e/ou relevância de interesse, como por exemplo eventos que se limitam a grupos ou contextos específicos; (iii) o comportamento dos usuários muda ao longo do tempo, e mais uma vez de forma pouco previsível e observável. Nesse caso, temos que levar em conta essa evolução e sermos capazes de lidar com ela.

Neste artigo propomos uma nova técnica para ordenar usuários e conteúdo, de acordo com a sua influência sobre outros usuários e com sua relevância, respectivamente. Essa nova técnica se baseia em uma definição intuitiva e circular de relevância e influência, ou seja, usuários influentes tendem a disseminar conteúdo relevante e conteúdo relevante é em geral disseminado por usuários influentes. É interessante notar que não nos concentramos em um grupo restrito e pequeno de conteúdos, mas em todo o tráfego disseminado por uma rede social. Essa definição foi modelada como uma extensão do problema de *PageRank* [Brin and Page 1998] e implementada eficientemente, uma vez que não requer a obtenção da rede de seguidores e determina as ordenações através de uma estratégia randomizada. A técnica proposta foi avaliada no contexto de dois cenários amplamente discutidos na rede social Twitter: política e automóveis. Finalmente, é também interessante notar que a nossa metodologia de avaliação é baseada em recomendação, que é uma forma indireta, mas que se mostrou efetiva de analisar a qualidade dos resultados providos pelo método proposto.

2. Trabalhos Relacionados

Nesta seção, revisamos brevemente as áreas que tratam de influência, recomendação e abordagens relacionadas ao algoritmo Pagerank, que são fundamentos do nosso trabalho.

Influência A noção de influência é um conceito muito importante nas áreas de sociologia [Katz and Paul 2005], comunicação e marketing. Em especial, com o surgimento e popularização das redes sociais, influência nesse contexto tem sido um tópico de muita atenção e pesquisa. Identificar quem são os usuários influentes, e como a informação relevante se propaga nas redes sociais são tarefas que, se bem resolvidas, trazem conhecimento estratégico para empresas de marketing, campanhas políticas e estudos sociológicos. Alguns exemplos são: identificar o melhor "ponto de partida" para campanhas de marketing no Facebook, ou caracterizar o surgimento de "fenômenos nacionais" em redes sociais, como a crítica ao comentarista Galvão Bueno espalhada no twitter, com a *hashtag* #foragalvao. No âmbito político, é importante discernir quem são os formadores de opinião, de forma a direcionar campanhas a esses indivíduos. Apesar da importância da noção de influência, não há um consenso na literatura sobre qual é a melhor forma de medir a influência de um determinado usuário. Várias métricas foram propostas: número de seguidores ou amigos, número de retweets ou citações, ou uma combinação destes [Cha et al. 2010, Leavitt et al. 2009]. Aparentemente, a maioria dos

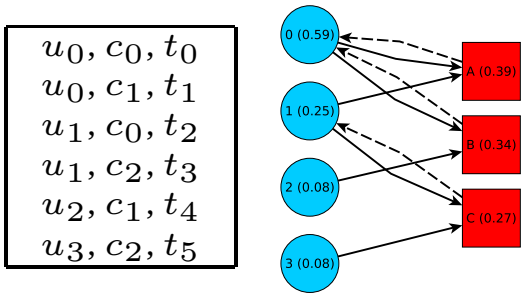
trabalhos conclui que popularidade (i.e. número de seguidores no twitter, número de amigos no Facebook) não necessariamente implica em influência [Cha et al. 2010]. Um exemplo é o trabalho de [Romero et al. 2011], onde os autores utilizam a noção de passividade para ajudar na determinação da influência de um usuário. A maioria dos usuários do Twitter atua como consumidor passivo, ou seja, não disseminam nenhum conteúdo para a rede. Outros usuários, por sua vez, acrescentam muito conteúdo, mas poucos usuários reagem ao conteúdo publicado. Tais usuários não são considerados influentes. Os usuários considerados influentes são aqueles que conseguem "romper a passividade" de outros usuários, ou seja, que publicam conteúdo que é propagado por seus seguidores. Outras métricas de influência utilizam o algoritmo PageRank, discutido logo a frente. O trabalho de [Iribarren and Moro 2007] propõe uma forma de predizer, quantitativamente, *como* e *quando* a informação se espalha nas redes sociais. Ele assume que a difusão de informação ocorre baseada em um modelo viral, mas tal ação depende da vontade individual humana. Experimentos realizados com dados provenientes de campanhas de marketing viral mostraram que o comportamento das pessoas é altamente heterogêneo, mesmo diante das mesmas decisões a serem tomadas. Portanto, modelos que predizem o comportamento de usuários levando em consideração a média da população falham em explicar a dinâmica das redes sociais. Em [Silva et al. 2011], nós estudamos o problema de predizer relações de influência a partir de dados difundidos na rede. Tal trabalho serviu como maior motivação para o este estudo.

Recomendação Com o grande crescimento do volume de informação disponível ao usuário, tornou-se difícil a tarefa de encontrar conteúdo relevante e novo. Nesse contexto surgiram os sistemas de recomendação [Resnick and Varian 1997]. Dentre eles, se destacam os algoritmos baseados em filtragem colaborativa [Koren and Bell 2011], que geram recomendações utilizando padrões de uso (exemplo: compras, avaliações), sem a necessidade de informação sobre o domínio ou os itens recomendados. No contexto específico do Twitter, vários trabalhos exploram recomendação, tanto de usuários quanto de tweets. O trabalho de [Uysal and Croft 2011], por exemplo, propõe um método personalizado por usuário, onde as recomendações feitas para um determinado usuário são escolhidas de forma a maximizar a probabilidade de que o usuário propague a informação recomendada. Outro exemplo é o trabalho de [Hannon et al. 2011], em que o os autores propõem técnicas baseadas em conteúdo e em filtros colaborativos a fim de recomendar usuários a serem seguidos.

PageRank O algoritmo PageRank [Brin and Page 1998] assinala um valor de importância a páginas da Web, de forma que uma página p tem um peso proporcional ao número e importância das páginas que apontam (através de *hyperlinks*) para p. Outro algoritmo que leva em conta a relação entre as páginas é o HITS (*Hypertext Induced Topic Search*), proposto em [Kleinberg 1998], que utiliza o conceito de autoridades e *hubs*. Intuitivamente, *hubs* são páginas que não são autoridades por si só, mas direcionam os usuários a páginas importantes. Páginas importantes (autoridades), por sua vez, são páginas que são apontadas por vários *hubs* diferentes. O autor de [Franceschet 2011] revisita os conceitos por trás do PageRank, ressaltando que a Web foi revolucionada pela idéia de introduzir a noção de um "índice de importância", que calibra o status de uma página utilizando apenas a topologia do grafo da Web. O trabalho revisita os conceitos sobre os quais o PageRank está fundamentado. Franceschet destaca, por exemplo, a so-

ciometria como um antecessor notavelmente antigo, uma vez que sociologistas foram os primeiros a utilizar abordagens de rede para identificar propriedades de grupos relacionados. O conceito por trás era a mesma premissa do Pagerank: *Uma pessoa tem prestígio se ela é endossada por pessoas prestigiosas*. Com a popularização e sucesso dos conceitos utilizados pelo PageRank, vários trabalhos passaram a explorar a circularidade da rede a fim de calcular a importância dos seus nós. [Baluja et al. 2008] utilizam *Random Walks* a fim de prover um método simples para propagar informações de preferência através de uma variedade de grafos, utilizando um estudo de caso que é a recomendação de vídeos para usuários no Youtube. [Tong et al. 2006] propõe um método dinâmico para monitorar a proximidade entre autores e conferências, também baseado em *Random Walks*. Por fim, os autores de [Weng et al. 2010] propõem o *TwitterRank*, que utiliza o PageRank para quantificar a influência de usuários no Twitter. Os autores utilizaram tanto a similaridade do conteúdo postado pelos usuários quanto a rede de seguidores, e conseguiram caracterizar a presença de homofilia no Twitter, além de propor uma nova forma de medir a influência de um usuário. Entretanto, o trabalho leva em conta apenas o tópico do conteúdo postado, desconsiderando os retweets, assim como os links formados entre usuários e conteúdo, como fazemos neste presente trabalho.

3. Detecção de Conteúdo Relevante e Usuários Influentes



(a) Dados (b) Modelo de difusão

Figura 1. Modelagem de dados de difusão: Visão geral. Os dados em 1(a) representam usuários u que postaram conteúdo c no tempo t. Esses dados são utilizados para gerar o grafo bipartido 1(b), que conecta cada usuário ao conteúdo postado por ele, e cada conteúdo é conectado de volta ao usuário que o criou.

3.1. Relevância e influência

Esta seção apresenta alguns conceitos importantes relacionados à identificação de conteúdos relevantes e usuários influentes com base em dados de difusão de informação. A idéia é associar os usuários influentes e conteúdo relevante através de uma definição circular. Além disso, esses conceitos consideraram a influência e relevância sob uma perspectiva da recomendação.

Seja C o conjunto de conteúdos e U o conjunto de usuários. Nós definimos a relevância global de um conteúdo $c \in C$ como uma função $r(c)$. Além disso, definimos a influência global de um usuário $u \in U$ como uma função $p(u)$. Como se trata de uma métrica orientada ao comportamento do usuário, a importância global de $r(c)$ depende da relevância personalizada $r(c, u)$, que dá a relevância de c para u. No entanto,

$r(c)$ também é afetada pela influencia dos usuários, ou seja quanto mais influentes forem os usuários para os quais c é relevante, mais relevante será c. Portanto, a relevância do conteúdo é baseado na influência dos usuários. Da mesma forma, definimos a influência $p(u)$ de um usuário u com base na relevância do conteúdo que ele produz. A função de influência personalizada $p(u_i, u_j)$ dá a influência de um usuário u_i para um usuário u_j. Essas definições circulares são formalizadas na Seção 3.3. É interessante entender o problema de identificar os usuários influentes e conteúdo relevante sob uma perspectiva de recomendação. Um conteúdo que é relevante para alguns usuários deveria ser recomendado para tais usuários. Portanto, podemos aplicar as funções $r(c, u)$ e $p(u_i, u_j)$ em um contexto de recomendação. Ao avaliar a eficácia dessas funções, podemos avaliar a qualidade das funções $r(c)$ e $p(u)$. Esta abordagem torna-se especialmente útil quando não há informação sobre a relevância do conteúdo e a influência dos usuários - que é um cenário muito frequente.

3.2. Dados de difusão de informação

Nós chamamos de dados de difusão de informação o conjunto de ocorrências de um item de informação. Cada ocorrência de um item é definida como uma tupla na forma $< u, c, t >$, onde u é um usuário do conjunto de usuários U, c é um conteúdo do conjunto de conteúdos C , e t é um instante do tempo. Portanto, os dados de difusão de informação descrevem a associação entre usuários e conteúdo ao longo do tempo. A Figura 1(a) mostra um exemplo ilustrativo de um conjunto de dados de difusão de informação onde $U = \{u_0, u_1, u_2, u_3\}$, $C = \{c_0, c_1, c_2\}$ durante o intervalo de tempo $[t_0, t_5]$. Considerando-se o Twitter como exemplo, U representa o conjunto de usuários, C representa o conteúdo postado pelos usuários (tweets, URLs, hashtags), e os instantes de tempo são definidos de acordo com o momento das postagens. Da mesma forma, U pode ser um conjunto de blogueiros que postam conteúdo (URLs, palavras-chave, temas) ao longo do tempo. Dados de difusão de informação aparecem em muitos outros cenários da vida real, especialmente em aplicações de mídia social. É importante notar que a nossa definição de dados de difusão de informação não leva a rede social em consideração. Em outras palavras, não temos qualquer informação sobre amizades, seguidores ou qualquer outro tipo de relação que pode ser considerado como um meio para a difusão da informação. Dada a dificuldade de conseguir esses dados em larga escala, obter essas informações pode ser uma limitação para a aplicação de trabalhos que utilizem a rede social em situações reais, situação que não ocorre no presente trabalho.

3.3. Modelo de Difusão de Informação

Esta seção dá uma definição formal do nosso modelo de difusão da informação. Nós definimos o modelo de acordo com os conceitos de relevância do conteúdo e influência do usuário descritos na Seção 3.1. Esse modelo pode ser descrito intuitivamente baseado na ideia de um *random surfer* (usuário que navega de forma aleatória pela rede), semelhante ao algoritmo PageRank. Além disso, apresentamos uma formulação algébrica para este modelo e um exemplo ilustrativo de sua aplicação. Nosso modelo é baseado em um grafo bipartido $G(U, C, F, E)$ que associa os usuários à conteúdos através de dois conjuntos de arestas, F e E. Para cada usuário $u \in U$ e conteúdo $c \in C$, existe uma aresta direcionada $(u, c) \in F$ se o usuário u propaga o conteúdo c. As arestas em F dão relevância ao conteúdo com base na influência do usuário. Além disso, existe uma aresta direcionada $(c, u) \in E$ que parte de um conteúdo para o usuário que o propagou. Arestas em E dão

crédito aos usuários de acordo com a relevância do conteúdo que eles criam. A Figura 1(b) apresenta o grafo bipartido construído a partir dos dados mostrados na Figura 1(a). Definimos a relevância do conteúdo $r(c)$ como a frequência relativa em que um *random surfer* que começa a partir de um nó de usuário arbitrário e, navegando através do gráfico bipartido $G(U, C, F, E)$, atinge um determinado conteúdo c. Já a relevância personalizada $r(c, u)$ começa a partir de um determinado usuário u em vez de um usuário arbitrário. De um modo semelhante, a influência de um usuário $p(u)$ é a frequência relativa que o *random surfer* visita um determinado usuário e pode ser personalizado, iniciando a partir de um usuário particular. Para dar uma visão mais realista sobre o nosso modelo, vamos considerar o Twitter como um cenário de exemplo. O algoritmo seguido pela nosso *random surfer*, com base em dados do Twitter, pode ser descrito como se segue: (i) Seleciona um perfil de usuário arbitrário; (ii) Escolhe um tweet aleatório ou retweet do usuário atual; (iii) Seleciona o perfil do autor do tweet dado; e (iv) Volta para o passo 2.

O gráfico bipartido $G(U, C, F, E)$ pode ser representado por duas matrizes, M e L. A matriz $M = (m_{i,j})$ é $|U| \times |C|$ e $m_{i,j} = 1/q_i$, onde q_i é a quantidade de conteúdo que u_i criou ou propagou. Além disso, $L = (l_{i,j})$ é $|C| \times |U|$ e $l_{i,j} = 1$ se o usuário u_j criou o conteúdo c_i ou $l_{i,j} = 0$, caso contrário. Com base em M e L, a função de relevância do conteúdo $r(c)$ e a função de influência do usuário $p(u)$ são definidas assim: $r = pM$ e $p = rL$, onde r é um vetor de relevância do conteúdo (ou seja, r_i é a relevância do conteúdo c_i) e p é um vetor de relevância do usuário (ou seja, p_j é a influência do usuário u_j). Nessa definição, assumimos que já temos um dos vetores (r ou p), a fim de calcular um a partir do outro, o que não acontece na realidade. Contudo, r e p podem ser calculados recursivamente: $r^{(k)} = r^{(k-1)}LM$ e $p^{(k)} = p^{(k-1)}ML$, onde $k \geq 0$ e $r^{(0)}$ e $p^{(0)}$ são vetores uniformes[1]. Esse modelo apresenta dois problemas importantes: (1) A possível presença de usuários *dangling* e (2) a possível existência de *buckets*. Um usuário *dangling* é um usuário que nunca propaga conteúdo de outros usuários. Considerando a metáfora do *random surfer*, o *surfer* ficará preso sempre que um usuário *dangling* u é atingido pois ele sempre seguirá arestas para o conteúdo gerado por u e depois consequentemente voltará para u. O PageRank também precisa lidar com páginas *dangling* e nós aplicamos uma solução semelhante aqui. Criamos uma aresta (u, c) de cada usuário *dangling* para um conteúdo "fantasma"c e adicionamos uma aresta (c, u) a partir do conteúdo fantasma para cada usuário $u \in U$. Como consequência, garantimos que o *random surfer* conseguirá, a partir de um usuário *dangling*, chegar a qualquer outro usuário em U. No gráfico mostrado na Figura 1(b), u_0 é um usuário *dangling*. Um *bucket* é um subgrafo fortemente conexo do grafo bipartido. Quando o *random surfer* atinge um *bucket*, ele não é capaz de deixá-lo. Podemos ver um usuário *dangling* como se fosse um *bucket* de tamanho 1. A fim de evitar que o *random surfer* fique preso em *buckets*, podemos adicionar um mecanismo de amortecimento ao nosso modelo. Esse mecanismo determina uma pequena probabilidade d do *random surfer* pular do usuário atual para um conteúdo aleatório ou vice-versa. Nós adicionamos esse mecanismo na definição de r e p da seguinte forma:

$$r^{(k)} = dr^{(k-1)}LM + (1-d)u \qquad e \qquad p^{(k)} = dp^{(k-1)}ML + (1-d)u$$

em que u é um vetor uniforme. Podemos reformular as equações acima algebricamente a fim de obter as suas soluções exatas de uma forma não recursiva:

$$r = (1-d)u(I - dLM)^{-1} \tag{1}$$

[1]Em um vetor uniforme, todos os valores são iguais e a soma deles é 1

$$p = (1 - d)u(I - dML)^{-1} \quad (2)$$

Na próxima seção, vamos discutir por que essa formulação algébrica não é computacionalmente eficiente. Duas questões mais importantes neste momento são: (1) Será que essas equações têm uma solução? e (2) Essas soluções são únicas? A resposta afirmativa para a primeira pergunta vem do fato de que as matrizes ML e LM são estocásticas. De fato, sabe-se que o produto de duas matrizes estocásticas é sempre uma matriz estocástica. Além disso, uma combinação linear de duas matrizes estocásticas é também estocástica. Em relação à pergunta 2, podemos mostrar que as nossas equações têm uma solução única, baseada no *teorema de Perron-Frobenius* [Frobenius 1912, Franceschet 2011]. O teorema de Perron-Frobenius diz que se uma matriz A é irredutível (ou seja, se seu gráfico associado é fortemente conectado) e também quadrada não negativa, então a equação $xA = rx$, onde $x > 0$ e $\sum_i x_i = 1$, tem uma única solução. Como M, L, e u são não-negativos, nossas equações tem matrizes não negativas. Além disso, a remoção de usuários *dangling* e de *buckets* garante que ML e LM são irredutíveis. Na Figura 1(b), calculamos os valores da influência do usuário e da relevância do conteúdo usando d igual a 0,85. Podemos notar que o usuário mais influente é u_0 ($p(u_0) = 0.59$), pois os dois conteúdos produzidos por u_0 (c_0 e c_1) são propagados por dois usuários (u_1 e u_2). Os conteúdos produzidos por u_1 também são propagados por dois usuários, mas esses usuários são menos influentes do que os usuários que propagam o conteúdo a partir de u_0. Portanto, u_1 é menos influente que u_0. O conteúdo mais relevante é c_0 porque ele foi difundido por dois usuários influentes (u_0 e u_1). Embora c_2 também seja difundido por dois usuários, esses usuários não são tão influentes como os associados a c_0. A elaboração de valores personalizados de relevância de conteúdo ($r(c, u)$) e influência de usuário ($p(u_i, u_j)$) em nosso modelo é simples. Nestes cenários, em vez de iniciar a partir de um usuário arbitrário, vamos supor que o *random surfer* começa a partir de um usuário específico para qual o modelo está sendo personalizado. Da mesma forma, em vez de saltar para um conteúdo aleatório com uma probabilidade não-zero, o *random surfer* sempre salta de volta para esse nó específico. Esse comportamento pode ser induzido substituindo o vetor uniforme u por um vetor 1_i, que é um vetor com todos os elementos iguais a 0, com exceção da posição i igual a 1, onde u_i é o usuário para o qual o modelo está sendo personalizado.

3.4. Solução Eficiente

Na seção anterior, descrevemos as equações que definem influência do usuário e relevância do conteúdo no nosso modelo. Para aplicar esse modelo em cenários reais de mídia social, com grande volume de usuários e conteúdo, precisamos resolver tais equações de forma eficiente. Em situações reais, as matrizes M e L tendem a ser muito grande e esparsas. Portanto, uma solução eficiente para nosso modelo deve levar em consideração essas propriedades. Como mostrado nas Equações 1 e 2, podemos calcular os vetores r e p invertendo uma matriz $|U| \times |U|$ e uma matriz $|C| \times |C|$. Como a inversão de uma matriz $n \times n$ tem custo $O(n^3)$, calcular os valores exatos de r e p não é viável em situações reais. No entanto, o método da potências [Mises and Pollaczek-Geiringer 1929, Franceschet 2011], que é um método de iteração rápido para calcular o autovalor e autovetor dominante de uma matriz, pode ser aplicado no cálculo de r e p. O Algoritmo 1 descreve o método da potências. Ele recebe duas matrizes (Z_1 e Z_2) e repetidamente itera sobre a solução g, que é iniciada como uniforme,

até que um determinado número de iterações k seja atingido. Se $Z_1 = M$ e $Z_2 = L$, o método nos dá o vetor de influência (p). Por outro lado, se $Z_1 = L$ e $Z_2 = M$, ele nos dá o vetor de relevância (r). Conforme descrito na seção anterior, podemos calcular os valores personalizados de influência e relevância para um usuário u_i, substituindo o vetor uniforme U por um vetor 1_i que tem 1 na i-ésima posição e 0 nas posições restantes. Além de aplicar o método de potências para calcular a influência e relevância, fazemos uso de representações esparsas das matrizes M e L, com o objetivo de reduzir a quantidade de memória e o tempo de execução para computar r e p. Mais especificamente, representamos matrizes no *formato de coordenadas*. Valores são armazenados numa lista de tuplas (linha, coluna, valor), onde apenas tuplas com valores diferentes de zero são inseridas.

Algoritmo 1: Método da Potências

Input: Z_1, Z_2, k, d
Output: g
1 $u \leftarrow$ vetor uniforme;
2 $g \leftarrow u$;
3 $i \leftarrow 0$;
4 **while** $i < k$ **do**
5 $\quad g \leftarrow dgZ_1Z_2 + (1.0 - d)u$

4. Resultados Experimentais

Nesta seção, iremos apresentar os principais resultados experimentais deste trabalho. Nosso objetivo é avaliar o modelo para a análise de influência de usuários e relevância de conteúdo utilizando dados reais.

4.1. Bases de dados

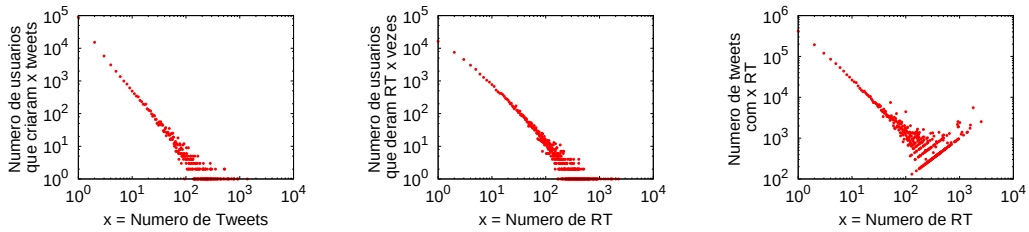
Nós utilizamos duas bases de dados obtidas a partir do *Twitter*[2]. A coleta de dados foi realizada através da API de coleta em modo *streaming* do Twitter[3]. Para cada base de dados, conjuntos de palavras-chave relevantes em um dado contexto foram selecionadas manualmente. Os contextos utilizados neste trabalho são *automóveis* e as eleições americanas para presidente. A Tabela 1 sumariza as principais propriedades das bases de dados utilizadas. De maneira geral, nós consideramos tweets como conteúdo e retweets como forma de propagação de conteúdo através de usuários.

nome	#usuários	#tweets	#RTs	período de coleta
Eleições	529.630	369.287	1.368.080	31/12/2011 - 31/01/2012
Carros	127.106	53.670	138.352	10/01/2012 - 05/03/2012

Tabela 1. Bases de dados

A Figura 2 mostra uma breve caracterização de importantes aspectos da base de dados Eleições. Mais especificamente, são apresentadas as distribuições do número de tweets por usuário, de retweets (RTs) por usuário e de retweets por tweet. Podemos notar que, como era esperado, a maior parte das distribuições aparenta seguir uma lei de potências, ou seja, grande concentração de popularidade e atividade de usuários. Realizamos a mesma caracterização para a base de dados de Carros, a qual apresentou propriedades muito semelhantes.

[2] `http://twitter.com/`
[3] `https://dev.twitter.com/docs/streaming-api`



(a) Distribuição de tweets por usuário (b) Distribuição de RTs por usuário (c) Distribuição de RTs por tweet

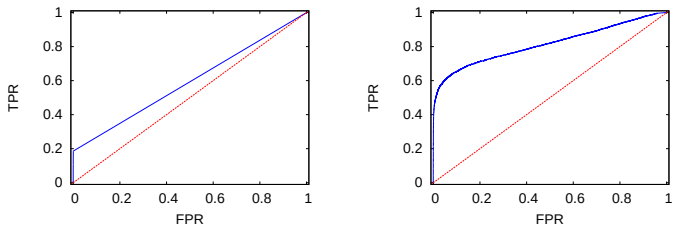
Figura 2. Caracterização da Base de dados Eleições

4.2. Recomendação de conteúdo

Apesar de se tratar de um problema bastante estudado na literatura, não existem maneiras efetivas de avaliar a qualidade de um método que identifica usuários influentes. Além disso, como nossa definição de relevância esta fortemente associada à de influência, analisar a eficácia do modelo proposto se torna um desafio. Neste trabalho, nós propomos uma interpretação do conceito de relevância com base no problema de recomendação. Dessa forma, um conteúdo é considerado relevante em geral, se ele é relevante para muitos usuários, especialmente os mais influentes. Isso permite que a avaliação do modelo proposto no contexto de recomendação de conteúdo gere evidências da sua qualidade na identificação de conteúdo relevante e usuários influentes. Em outras palavras, se a versão personalizada do modelo é capaz de recomendar conteúdo para usuários de forma acurada, então podemos inferir que ele também é capaz de identificar conteúdo relevante em geral e, por consequência das premissas do modelo, avaliar a influência de usuários.

Os experimentos para esta avaliação foram realizados da seguinte forma: Para cada tweet da base de dados, seus retweets foram divididos entre dados de treino e teste numa razão 50%/50%, considerando a ordem de ocorrência. As primeiras ocorrências de cada tweet foram inserídas na base de treino e o restante na base de teste. Os dados de treino foram utilizados na construção de modelos de recomendação a serem avaliados utilizando os dados de teste. Nós consideramos como baseline uma técnica tradicional de recomendação de conteúdo denominada *Filtragem Colaborativa*. A idéia básica desta técnica é se basear na semelhança entre interesses de usuários para recomendar novos itens. Dois usuários u_1 e u_2 são considerados semelhantes se eles compartilham interesses em comum e o modelo de recomendação identifica potenciais itens consumidos por u_1 a serem recomendados a u_2 e vice-versa. Existem diversas variações de filtragem colaborativa. Neste trabalho, nós empregamos uma estratégia denominada *recomendação de itens baseada em vizinhos mais próximos com pesos* (do inglês *weighted item KNN*). A implementação empregada foi obtida a partir da biblioteca de recomendação *MyMediaLite*[4]. Todos os parâmetros foram utilizados em sua configuração *default*. A métrica de avaliação utilizada neste trabalho é a *curva de ROC* (do inglês *Receiver Operating Characteristic*). Uma curva de ROC mostra como a taxa de verdadeiros positivos (TPR) e falsos positivos (FPR) varia ao longo do intervalo de *scores* gerado por um dado modelo. Em geral, é esperado que modelos sejam mais efetivos no topo das previsões, ou seja, no início da curva de ROC. A medida da qualidade de um modelo é obtida através do cálculo da área da curva de ROC, denominada AUC (do inglês *Area Under the Curve*).

[4] `http://www.ismll.uni-hildesheim.de/mymedialite/index.html`



(a) WIKNN, AUC = 0.59 (b) MDC, AUC = 0.81

Figura 3. Curva de ROC para uma técnica de filtragem colaborativa (WIKNN) e para nosso modelo baseado em difusão de conteúdo (MDC)

tweet	relevância	usuário
Dia sem Globo' faz emissora registrar o dobro da Record e o triplo do SBT em audiencia - Parabens aos envolvidos.	0.003	kibeloco
Quanto a Fiat pagou para mostrar que esse carro de bosta pequeno e que o porta-malas é difícil de abrir? #BBB12	0.002	kibeloco
Delegado passa 4 horas ouvindo BBBs - Quatro horas? Coitado! Fez prova de resisteancia e nao levou nem um Fiat...	0.002	kibeloco
"Tem uns professores que acham que os alunos tem problemas de audicao, so pode.	0.002	PiadasDeAluno
FALA: GATA VOCE NAO é DONA DA FIAT, MAIS VOCE FAZ MEU STILO AS MINA PIRA E FICA SEM JEITO D +	0.002	AsMinaaPira

Tabela 2. Tweets mais relevantes da base de dados Carros

Nesta avaliação foi considerada apenas a base de dados Carros. A Figura 3 mostra uma comparação entre o modelo baseado em difusão de conteúdo e uma técnica de filtragem colaborativa. Podemos notar que nosso modelo apresenta uma acurácia 37% superior ao baseline. A explicação para a superioridade do nosso modelo é a sua capacidade de generalizar relacionamentos entre usuários que não compartilham tweets em comum através de *random walk*. Dessa forma, nosso modelo é mais efetivo em lidar com a esparsidade dos dados do que técnicas tradicionais. No modelo, o número de iterações e o valor do *damping factor* (d) foram definidos como 10 e 0.85, respectivamente.

4.3. Conteúdo relevante e usuários influentes

Esta seção apresenta os tweets relevantes e usuários influentes identificados pelo nosso modelo a partir das bases de dados Carros e Eleições. O objetivo é mostrar que tal modelo obtém resultados que são semanticamente válidos dentro dos contextos considerados.

A Tabela 2 mostra os tweets mais relevantes identificados na base Carros. Os tweets são mostrados da forma que aparecem na base de dados, apenas alguns caracteres especiais foram removidos. Devido a uma falha na coleta, alguns dos tweets citados não são sobre automóveis. O tweet mais relevante, por exemplo, foi coletado devido a presença do termo 'audiência', que a API de coleta confunde com o termo 'audi'. Em geral, podemos notar que os tweets mais relevantes têm papel humorístico e foram criados por usuários relacionados a esse gênero, como é o caso do usuário *kibeloco*. Esses tweets alcançam um grande número de retweets, o que explica a identificação dos mesmos como relevantes. Resultados semelhantes foram encontrados na análise da base de dados Eleições, como mostrado na Tabela 3.

A Figura 4 mostra os usuários mais influentes na base Carros (Figura 4(a)) e Eleições (Figura 4(b)). De maneira geral, os usuários mais influentes são aqueles que obtiveram um maior número de retweets, especialmente retweets de outros usuários in-

tweet	relevância	usuário
Rosa Parks sat, so that Dr. Martin Luther King, Jr. could walk, so that Barack Obama could run, so that the next generation, us, could fly.	0.004	mind
Rick Santorum's stance on homosexuality is so fucking gay.	0.003	SethMacFarlane
I feel like Newt Gingrich is what Justin Bieber will look like old.	0.002	SethMacFarlane
Rick Santorum seems so homophobic that I'm surprised he even allows another man to vote for him.	0.002	GarryShandling
It's illegal for prisoners to vote, but they can run for President of the United States!	0.002	WTFuckFacts

Tabela 3. Tweets mais relevantes da base de dados Eleições

usuário	relevância
Estadao	0.008
VEJA	0.008
viacertanatal	0.007
kibeloco	0.007
aguinaldaosilva	0.006
rd1oficial	0.006
viacertaRN	0.003
JovemPanBH	0.003
kiamotorsbrasil	0.003

(a) Carros

usuário	influência
BorowitzReport	0.020
LOLGOP	0.012
BreakingNews	0.008
RonPaul	0.008
robdelaney	0.007
thinkprogress	0.007
SethMacFarlane	0.006
AP	0.06
rationalists	0.005

(b) Eleições

Figura 4. Usuários mais relevantes identificados

fluentes, e compreendem perfis de agências de notícias (e.g., Estadao, AP), perfis humorísticos(e.g., kibeloco, LOLGOP), e personagens atuantes (e.g., aguinaldaosilva, RonPaul). É interessante notar que tais resultados são diferentes dos de técnicas simples como a contagem do número de retweets ou do número de seguidores dos usuários.

5. Conclusões

Neste artigo, propusemos uma nova técnica para ordenar usuários e conteúdos, de acordo com a sua influência e sua relevância, respectivamente. Essa nova técnica se baseia em uma definição intuitiva e circular de relevância e influência, ou seja, usuários influentes tendem a disseminar conteúdos relevantes e conteúdos relevantes são em geral disseminados por usuários influentes. Nossa técnica foi avaliada utilizando duas bases de dados reais do Twitter. Os resultados obtidos mostram que a técnica proposta apresenta ganhos de 37% quando comparada a um método de filtragem colaborativa. Nossa técnica aponta ainda como usuários influentes aqueles que realmente são influentes na rede (agências de notícias, humoristas e personagens famosos). Além disso, classificamos como relevantes conteúdos que são muito disseminados na rede, especialmente por usuários influentes. Esse tipo de resultado não é obtido com a utilização de técnicas mais simples (e.g. contagem do número de seguidores ou do número de retweets). Em termos de trabalhos futuros, pretendemos construir um modelo formal, avaliar a tarefa de recomendação de usuários a seguir em redes sociais como o Twitter, além de comparar o nosso método com outros da literatura, assim como a nossa medida de influência com outras medidas já usadas na literatura. Todas essas tarefas devem considerar outras bases de dados reais.

Referências

Baluja, S., Seth, R., Sivakumar, D., Jing, Y., Yagnik, J., Kumar, S., Ravichandran, D., and Aly, M. (2008). Video suggestion and discovery for youtube: taking random walks through the view graph. In *Proceedings of the 17th international conference on World Wide Web*, pages 895–904, New York, NY, USA. ACM.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K. P. (2010). Measuring User Influence in Twitter: The Million Follower Fallacy. In *In Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, Washington DC, USA.

Franceschet, M. (2011). Pagerank: standing on the shoulders of giants. *Commun. ACM*, 54(6):92–101.

Frobenius, G. (1912). Über Matrizen aus nicht Negativen Elementen.

Hannon, J., McCarthy, K., and Smyth, B. (2011). Finding useful users on twitter: twittomender the followee recommender. In *Proc. of the 33rd European conf. on Advances in information retrieval*, pages 784–787, Berlin, Heidelberg. Springer-Verlag.

Iribarren, J. L. and Moro, E. (2007). Information diffusion epidemics in social networks. *ArXiv e-prints*. http://arxiv.org/abs/0706.0641.

Katz, E. and Paul, F. (2005). *Personal Influence: The Part Played by People in the Flow of Mass Communications*. Transaction Publishers.

Kleinberg, J. M. (1998). Authoritative sources in a hyperlinked environment. In *Proc. of the 9th annual ACM-SIAM symposium on Discrete algorithms*, pages 668–677, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

Koren, Y. and Bell, R. M. (2011). Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186.

Leavitt, A., Burchard, E., Fisher, D., and Gilbert, S. (2009). The Influentials: New Approaches for Analyzing Influence on Twitter. *Webecology Project*.

Mises, R. V. and Pollaczek-Geiringer, H. (1929). Praktische Verfahren der Gleichungsauflösung. *Zamm-zeitschrift Fur Angewandte Mathematik Und Mechanik*, 9:58–77.

Resnick, P. and Varian, H. (1997). Recommender systems. *Commun. ACM*, 40(3):56–58.

Romero, D. M., Galuba, W., Asur, S., and Huberman, B. A. (2011). Influence and passivity in social media. In *Proceedings of the 20th international conference companion on World wide web*, pages 113–114, New York, NY, USA.

Silva, A., Valiati, H., Guimarães, S., and Meira Jr, W. (2011). From individual behavior to influence networks: A case study on twitter. In *Proc. of the 17th Brazilian Symposium on Multimedia, Hypermedia and Web*, Porto Alegre, RS, Brazil. SBC.

Tong, H., Faloutsos, C., and Pan, J.-Y. (2006). Fast random walk with restart and its applications. In *Proceedings of the Sixth International Conference on Data Mining*, pages 613–622, Washington, DC, USA. IEEE Computer Society.

Uysal, I. and Croft, W. B. (2011). User oriented tweet ranking: a filtering approach to microblogs. In *Proc. of the 20th ACM international conference on Information and knowledge management*, pages 2261–2264, NY, NY, USA. ACM.

Weng, J., Lim, E.-P., Jiang, J., and He, Q. (2010). Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the 3rd ACM international conference on Web search and data mining*, pages 261–270. ACM.