Explorando a complementaridade entre estratégias para detecção de usuários influentes no Twitter*

Alan Neves , Ramon Vieira , Fernando Mourão , Leonardo Rocha

Universidade Federal de São João del Rei,
São João del Rei, Minas Gerais, Brasil

{aneves,ramonv,fhmourao,lcrocha}@ufsj.edu.br

Abstract. *The so-called influencers have an important role on the information diffusion in social media environments, since they might dictate peer-to-peer recommendation, impacting tasks such as brand evaluation, advertising, etc. Despite the growing number of works that identify influencers by exploiting distinct information, deciding on the best strategy for each domain is complex. In this work, we perform a quantitative study among some of the main strategies for identifying influencers. As main contributions, we highlight a better understanding about the selected strategies and a novel and effective meta-learning approach, based on PCA, that is able to combine linearly distinct strategies.*

Resumo. *Usuários influentes desempenham um importante papel na difusão de informação em mídias sociais, uma vez que podem ditar a publicidade par-a-par, impactando tarefas como validação de marca, propaganda, etc. Apesar do número crescente de trabalhos que visam identificar tais usuários, explorando diferentes características, decidir sobre qual a melhor estratégia para cada domínio é uma tarefa complexa. Neste trabalho realizamos um estudo quantitativo sobre algumas das principais estratégias para detecção de usuários influentes. Como principais contribuições, destacamos uma melhor compreensão sobre as estratégias selecionadas e uma nova e eficaz abordagem de meta-aprendizagem, baseada no PCA, capaz de combiná-las linearmente.*

1. Introdução

Recentemente, pesquisas relacionadas à difusão da informação [Bakshy et al. 2011] vêm recebendo grande atenção dada a popularização dos aplicativos de mídia social, tais como *microbloggings* e redes sociais. Esses aplicativos têm possibilitado um número crescente de pessoas difundir pensamentos, opiniões e comentários sobre temas distintos [Cha et al. 2010], gerando uma enorme e valiosa quantidade de dados. Diversos estudos apontam os chamados *usuários influentes* como personagens principais neste processo de difusão. Usuários influentes são pessoas com a capacidade de persuadir outras pessoas, afetando suas ações e comportamento, sendo assim capazes de ditar publicidades boca-a-boca e recomendações por pares com implicações em várias áreas, tais como recomendação, pesquisas de opinião, validação de marcas, dentre outros [Wu et al. 2011].

*Esse trabalho foi parcialmente financiado por CNPq, CAPES, FINEP, Fapemig, e INWEB.

Além de atrativa, a detecção de usuários influentes tem se mostrado uma tarefa desafiadora [Cha et al. 2010], sob a qual destacamos três principais dificuldades. Primeiramente, não está claro quais características são relevantes no processo de detecção. A segunda dificuldade é como combinar e ponderar as características selecionadas, de acordo com um objetivo específico (e.g., econômico, social, etc.) e com as informações disponíveis (e.g., conexão entre os usuários, rota das informações, etc.). Por fim, a maior dificuldade é a forma de avaliar a eficácia dos métodos propostos, tendo em vista que não há uma resposta correta sobre os usuários realmente influentes em domínios reais [Cha et al. 2010]. Além disso, não identificamos na literatura nenhum trabalho que analise e compare as principais estratégias para a detecção de usuários influentes.

Este trabalho realiza um estudo quantitativo de análise e comparação entre algumas das principais estratégias de detecção de usuários influentes. Nossa meta é entender melhor o efeito de estratégias distintas em cada domínio, auxiliando na tomada de decisão sobre qual das estratégias usar. Primeiramente, pesquisamos vários estudos na literatura e propomos uma nova taxonomia para esses estudos, baseado em características comuns que são exploradas. Foram avaliadas seis estratégias muito referenciadas: PageRank, PCC, ProfileRank, Leitores Efetivos e duas sumarizações estatísticas básicas, as quais foram agrupadas em três classes. A primeira classe inclui as estratégias que exploram a estrutura da rede (*PageRank*™ [Page et al. 1999] e PCC [Ilyas and Radha 2011]). A segunda refere-se a estratégias focadas no conteúdo e fluxo da informação (ProfileRank [Silva et al. 2013] e Leitores Efetivos [Lee et al. 2010]). A terceira classe consiste em estratégias que exploram sumarizações estatísticas das ações do usuário (número de relacionamentos [Wu et al. 2011] e número de postagens propagadas [Bakshy et al. 2011]).

Validamos nossa taxonomia medindo o nível de concordância entre as estratégias selecionadas utilizando amostras de dados do Twitter. Aplicamos cada estratégia nestas amostras, derivando uma lista ordenada decrescente dos Top-50 usuários influentes. Em seguida, foi utilizada a versão generalizada da métrica Kendall's Tau [Fagin et al. 2003] como medida de concordância entre pares de listas. Nossas análises confirmaram que as listas obtidas de estratégias que pertencem à mesma classe apresentam o maior nível de concordância. A partir destas análises, foi utilizada a Análise de Componentes Principais (PCA) para extrair informações úteis e ortogonais modeladas por elas [Pearson 1901]. Em primeiro lugar, analisamos a complementaridade da informação modelada por estratégias pertencentes à mesma classe, observando que são fortemente correlacionadas. Em seguida, avaliamos a complementaridade entre as estratégias que pertencem a classes distintas, a fim de determinar quão redundante são estratégias derivadas de campos e teorias distintas. Foi observado que as listas de usuários influentes obtidas por cada uma das técnicas, sob uma mesma amostra, divergem consideravelmente. Além disso, podemos interpretar o uso de PCA como uma estratégia de meta-aprendizagem que combina linearmente estratégias distintas.

Todas as implementações e execuções de experimentos foram realizadas pelo aluno Alan Neves, sob a orientação do professor Leonardo Rocha. A concepção da estratégia de meta-aprendizagem bem como as análises de todos os resultados foram feitas em conjunto, aluno e professor, com a colaboração do professor Fernando Mourão. Além disso, esse trabalho contou com o auxílio do aluno Ramon Vieira, que construiu o coletor de posts e informações dos usuários no Twitter.

2. Trabalhos Relacionados

Existem diversos trabalhos na literatura visando identificar usuários influentes em mídias sociais. Dentre estes trabalhos, encontramos um primeiro grupo focado em teorias sociológicas. Em [Subbian et al. 2013], os autores propõem um método baseado na noção de capital social, que mede a capacidade de ligação (i.e., similaridade) e conexão (i.e., diversidade) entre as pessoas para fins de cooperação e comunicação em uma rede. Ilyas e Radha, [Ilyas and Radha 2011] propuseram uma medida de centralidade, conhecida como Centralidade de Componentes Principais (PCC), para identificação de vizinhanças influentes em uma rede. Em [Kitsak et al. 2010], os autores demonstraram que os usuários influentes mais eficientes estão localizados no núcleo de uma rede, identificados por uma análise de decomposição de k-núcleos.

Um segundo grupo refere-se à modelagem matemática/computacional do problema, que variam desde simples análises estatísticas de registros de atividades a estratégias baseadas em grafos [Bakshy et al. 2011, Wu et al. 2011, Ilyas and Radha 2011]. Usualmente, há uma alta intersecção entre o tipo de informação explorada pelas estratégias do primeiro e segundo grupos. Entretanto, estratégias do segundo grupo tendem a combinar simples informações das estruturas sociológicas com informações adicionais, tais como o conteúdo e o tempo. De fato, alguns trabalhos argumentam que a influência dos usuários está relacionada às suas ligações e à ordem temporal da publicação do conteúdo veiculado [Lee et al. 2010, Silva et al. 2013]. Nesta linha, [Silva et al. 2013] leva em conta a ordem temporal da difusão de informação, sem considerar os relacionamentos para obter uma pontuação de influência dos usuários. Por sua vez, [Page et al. 1999] adaptou o algoritmo PageRank para derivar a pontuação de influência de cada usuário, modelado como um nó em um grafo dirigido.

Destacamos um terceiro grupo composto por trabalhos baseados em teorias econômicas focadas no marketing viral e adoção de marcas [Galeotti and Goyal 2010, Li et al. 2010]. O objetivo é entender os mecanismos que levam a uma reação em cadeia de influência em larga escala, a um custo de marketing muito baixo. Em [Li et al. 2010] os autores usaram uma rede neural artificial para encontrar revisores influentes para a publicidade boca-a-boca. Além disso, em [Galeotti and Goyal 2010] é proposta a *Lei de poucos* para explicar o papel dos usuários influentes em grupos sociais, que afirma que a maioria dos indivíduos obtêm grande parte das informações de um conjunto muito pequeno de usuários, os influentes. Diante da numerosa quantidade de trabalhos existentes, entender a complementariedade entre eles e prover uma solução adequada que combine esses trabalhos é o foco desse trabalho.

3. Uma nova taxonomia para estratégias de detecção de usuários influentes

Nesta seção, apresentamos uma nova taxonomia para esforços em detecção de usuários influentes, com base nas distintas premissas e características exploradas na literatura. Essa taxonomia foi concebida a partir de uma metodologia simples e não-automática de classificação. Em primeiro lugar, examinamos os principais trabalhos relacionados publicados ou referenciados nos últimos anos. Em seguida, extraímos desses trabalhos as principais características que eles exploraram a fim de determinar usuários influentes. Em terceiro lugar, estas características foram agrupadas de acordo com o tipo de informação modelada. Finalmente, os trabalhos foram classificados de acordo com os grupos

definidos na última etapa. É importante frisar que esta metodologia pode ser aplicada a qualquer conjunto de estratégias existentes na literatura e os grupos resultantes podem depender do conjunto de estratégias avaliadas. Além disso, destaca-se que não temos a intenção de fornecer um conjunto fechado de classes neste trabalho. Em vez disso, propomos uma maneira de estender a taxonomia com novas estratégias que surjam.

Com relação a primeira e segunda etapas da metodologia, selecionamos seis estratégias muito referenciadas na literatura as quais descrevemos detalhadamente a seguir:

- **PCC:** Centralidade de Componentes Principais (PCC) [Ilyas and Radha 2011] é baseada em centralidade, uma medida de relevância para os usuários em uma rede que leva em conta a vizinhança de cada usuário. PCC estende a métrica EVC - Centralidade de autovalores [Bonacich and Lloyd 2001], explorando as k características dominantes em um grafo. Estas características são determinadas por meio de uma Decomposição de Valores Singulares (SVD) da matriz adjacente que representa a rede social [Golub and Loan 1996]. Com base nesta decomposição e no operador de Hadamard ($\odot$) [Davis 1962], o valor do PCC é derivado de acordo com a Equação 1. Nesta fórmula, k é o número de características (ou seja, autovetores) a ser explorado e U e S são duas matrizes obtidas por meio do SVD. Aplicando o PCC em uma matriz adjacente, derivamos uma medida de centralidade para cada usuário. Quanto maior esta centralidade, maior a influência do usuário.

$$C_p = \sqrt{(U_{n\times k} \odot U_{n\times k})(S_{k\times 1} \odot S_{k\times 1})} \tag{1}$$

- **PageRank:** Algoritmo bem conhecido proposto em [Page et al. 1999] e explorado comercialmente pelo Google™ para determinar ranks globais para páginas da Web. Basicamente, ele calcula a propagação de influência entre os nós em um grafo direcionado. A fim de identificar usuários influentes, os nós deste grafo representam usuários e as arestas suas relações sociais. Assim, o valor do PageRank (PR) de cada usuário u_i é dado pela Equação 2, onde M_i é o conjunto de usuários conectados ao usuário u_i; $L(j)$ é o tamanho da vizinhança de u_j; α é o fator de amortecimento (com valor padrão de 0,85); e N é o número de usuários no grafo. O valor de PageRank representa a probabilidade de se atingir cada usuário na rede. Quanto maior for esta probabilidade, maior é a influência do usuário.

$$PR(u_i) = \alpha \sum_{j\in M_i} \frac{PR(u_j)}{L(j)} + \frac{1-\alpha}{N} \tag{2}$$

- **Leitores Efetivos:** Considera o relacionamento entre usuários e a ordem temporal no qual a informação atinge cada usuário. Os autores observaram que a informação se propaga mais rapidamente nos estágios iniciais do processo de difusão. Com base nesta observação, os autores propuseram uma pontuação de Leitores Efetivos, que é o número de usuários distintos que receberam uma dada informação pela primeira vez, a partir de um usuário específico. Supondo que os usuários leram suas mensagens cronologicamente assim que elas chegam, dada uma mensagem m publicada por um usuário u_i, os leitores efetivos $ER_0(m, u_i)$ são definidos pela equação 3, onde F é uma marcação binária de estado para cada usuário u_j. F é inicializado como o valor zero e é definido como 1 no primeiro momento que u_j recebe m de sua vizinhança. O valor de

influência IF_0 atribuído a u_i é definido como a soma dos valores ER_0 derivados para o conjunto $T(u_i)$ de todas as mensagens publicadas por u_i, como mostra a equação 4.

$$ER_0(m, u_i) = \sum_{u_j \in follower(u_i)} \overline{F(u_j, m)} \quad (3) \qquad IF_0(u_i) = \sum_{m \in T(u_i)} \| \, ER_0(m, u_i) \, \| \quad (4)$$

- **ProfileRank:** Explora uma definição cíclica de relevância e influência [Silva et al. 2013]. Usuários influentes propagam conteúdos relevantes e conteúdos relevantes são disseminados por usuários influentes. ProfileRank classifica ambos usuários e conteúdo de acordo com a pontuação de influência e relevância e modela a difusão de informação através do relacionamento entre usuários e conteúdo ao longo do tempo, por meio de duas matrizes M e L. M é uma matriz com dimensões $|U| \times |C|$ que representa o conjunto de conteúdos C criado pelo conjunto de usuários U. ProfileRank inicializa cada posição $M_{i,j}$ como $\frac{1}{q_i}$, onde q_i é o número de pedaços de conteúdo que u_i criou ou propagou. Por sua vez, a matriz L tem dimensões $|C| \times |U|$, onde $L_{i,j} = 1$ se o usuário u_j criou um pedaço de conteúdo C_i e $L_{i,j} = 0$, caso contrário. Baseado nessas matrizes, a pontuação da relevância do conteúdo (r) e influência do usuário (p) são definidas de acordo com as equações 5 e 6, respectivamente. Novamente, d representa o fator de amortecimento e, neste caso, é usado para prevenir subgrafos fortemente conectados. I é a matriz identidade e V é um vector uniforme.

$$r = (1-d)V(I - dLM)^{-1} \quad (5) \qquad p = (1-d)V(I - dML)^{-1} \quad (6)$$

- **Número de seguidores:** Projetada especificamente para o Twitter [Cha et al. 2010]. Um seguidor é um usuário que deseja ficar atualizado sobre as ações do usuário que ele está seguindo. Basicamente, alguns trabalhos contam o número de seguidores que cada usuário tem no sistema. Quanto maior for este número, maior a influência do usuário.

- **Número de retweets:** Originalmente projetada para o Twitter [Wu et al. 2011]. Ela sumariza o número de tweets publicados por um usuário que foram reencaminhados por outros usuários (ou seja, *retweets*). Quanto maior o número de *retweets* relacionados a um determinado usuário, maior a influência do usuário.

O terceiro passo da metodologia corresponde em agrupar as características de acordo com o tipo de informação modelada, e, como quarto passo, as estratégias são classificadas de acordo com os grupos previamente definidos. Com base nas estratégias avaliadas, foram identificadas três classes principais.

A primeira classe compreende as estratégias que levam em consideração apenas a estrutura da rede (ER). Encontramos neste grupo o algoritmo *PageRank*™ [Page et al. 1999], que calcula uma pontuação de influência para cada vértice em um grafo direcionado usando apenas relacionamentos e a propagação na rede e o PCC [Ilyas and Radha 2011], que utiliza uma métrica baseada em centralidade para determinar vizinhanças influentes em uma rede.

Estratégias pertencentes à segunda classe exploram conteúdo e fluxo (C&F) para determinar usuários influentes. Observam-se nesta classe o ProfileRank [Silva et al. 2013] e Leitores Efetivos [Lee et al. 2010]. Enquanto o ProfileRank

modela a difusão de informação apenas considerando a ordem temporal no qual as mensagens são propagadas em uma rede social, Leitores Efetivos avalia a difusão de informação como um efeito-cascata que tópicos têm entre os usuários.

Finalmente, a terceira classe corresponde a estratégias voltadas para Sumarizações Estatísticas (SE) de logs de atividade dos usuários. Basicamente, são estratégias que visam determinar o nível de influência dos usuários pela sumarização de alguns atributos de usuários, tais como o número de seguidores (#S) e o número de retweets (#R).

3.1. Avaliação da Taxonomia

A fim de avaliar a taxonomia proposta, realizamos experimentos nos quais medimos o nível de concordância entre as estratégias selecionadas. Realizamos todas as análises em amostras de dados do Twitter, dada a sua relevância para a difusão de informação na Web.

3.1.1. Amostras de dados

Foram coletadas duas amostras distintas do Twitter. A primeira amostra (Coleção 1) refere-se aos tweets advindos de Belo Horizonte/MG e abrangem o período de 30/10/2014 a 06/11/2014. Durante este período, os dois maiores times de futebol de Belo Horizonte estavam jogando as finais dos dois maiores campeonatos nacionais de futebol (Copa do Brasil e Campeonato Brasileiro). A segunda amostra de dados (Coleção 2) compreende os tweets oriundos da maior cidade brasileira, São Paulo, publicados entre 24/12/2014 e 01/05/2015, que corresponde ao período das celebrações de Natal e Ano Novo. Em ambas as coleções, foram coletados os tweets relacionados com os 10 temas mais discutidos (os *TopTrends*) a cada 10 minutos de coleta (utilizamos a *APIStreaming*[2]). Foram também coletadas informações públicas sobre todos os usuários relacionadas ao tweet, tais como o autor do tweet, seus/suas seguidores e amigos (utilizamos a API REST para reunir todas informações públicas dos usuários[3]. A tabela 1 detalha ambas amostras de dados.

Tabela 1. Sumário das amostras utilizadas em todos experimentos.

Coleção	# usuários	# tweets
Coleção 1	1.253	3.248
Coleção 2	3.800	13.102

3.1.2. Avaliação experimental

Para determinar o nível de concordância entre estratégias distintas, realizamos uma comparação par-a-par entre os resultados gerados por cada estratégia. Neste sentido, foi aplicada cada estratégia S_a nas amostras de dados e foi obtida uma lista ordenada decrescente dos Top-50 usuários influentes, de acordo com as pontuações definidas por S_a. Em seguida, foi utilizada a versão generalizada da métrica Kendall's Tau [Fagin et al. 2003], com parâmetro de penalidade $p = 0$, como medida de concordância entre pares de listas. Quanto maior o valor da métrica Kendall's Tau entre duas listas, mais elas concordam. Normalizamos os resultados obtidos, tal como sugerido por [McCown and Nelson 2007]. A tabela 2 mostra os resultados desta análise .

[2]https://dev.twitter.com/docs/streaming-apis

[3]https://dev.twitter.com/docs/api

Conforme destacado nas tabelas, nossa análise confirmou que as listas derivadas de estratégias que pertencem à mesma classe apresentam o mais alto nível de concordância, o que corresponde à nossa primeira contribuição para este artigo. Uma exceção refere-se à métrica de Leitores Efetivos (da classe C&F) na coleção 1, que apresentou níveis mais elevados de concordância com a estratégia Número de Seguidores (da classe SE). Este comportamento está relacionado com o tamanho da amostra e limitações inerentes a essas estratégias. Devido ao curto período de tempo relacionado à coleção 1, há poucos usuários nesta coleção com elevado número de relacionamentos. Neste tipo de cenário, Leitores Efetivos se comporta de forma semelhante ao Número de Seguidores porque não há conexões suficientes na rede para modelar a propagação de informações. Com base nestes resultados, uma questão relevante surge: é possível derivar uma única métrica de cada classe que combina informações úteis e distintas capturadas por estratégias da mesma classe? Tentaremos responder essa pergunta na próxima seção.

Tabela 2. Comparação par-a-par das listas de Top-50 usuários influentes gerada pelas estratégias avaliadas utilizando a métrica Kendall's Tau.

	PCC	PageR	LE	PR	#R	#S
PCC	1.00	0.77	0.24	0.09	0.32	0.29
PageR	0.78	1.00	0.29	0.09	0.34	0.32
ER	0.23	0.28	1.00	0.35	0.40	0.68
PR	0.09	0.09	0.36	1.00	0.05	0.08
#R	0.32	0.34	0.41	0.05	1.00	0.62
#S	0.29	0.32	0.68	0.08	0.62	1.00

Coleção 1

	PCC	PageR	LE	PR	#R	#S
PCC	1.00	0.72	0.36	0.15	0.02	0.22
PageR	0.68	1.00	0.31	0.10	0.23	0.34
ER	0.36	0.32	1.00	0.72	0.01	0.13
PR	0.15	0.11	0.73	1.00	0.00	0.02
#R	0.02	0.23	0.01	0.00	1.00	0.43
#S	0.21	0.32	0.13	0.02	0.43	1.00

Coleção 2

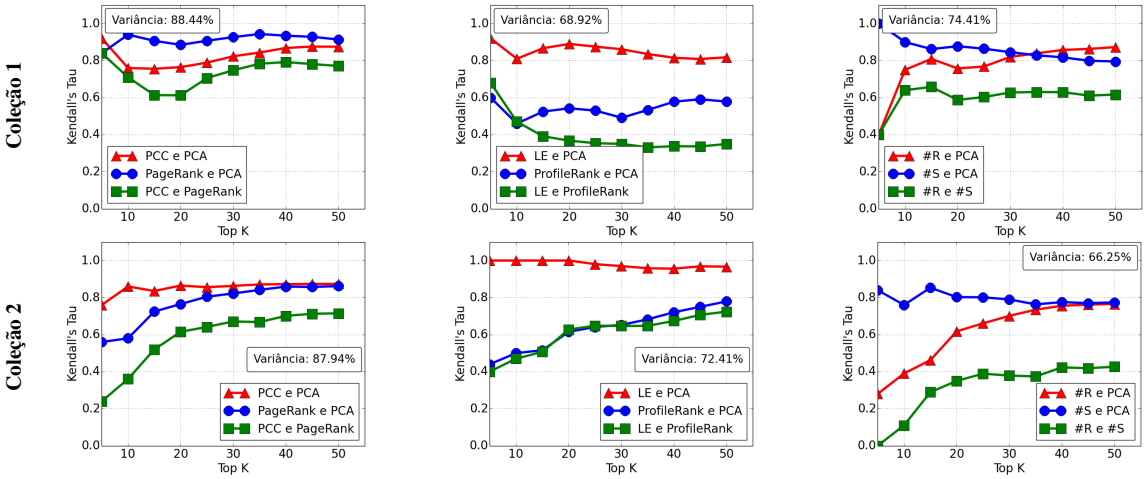
4. Análise de complementaridade intra-classe

Nesta seção, analisamos a complementaridade das informações modeladas por estratégias pertencentes à mesma classe. Os experimentos conduzidos na seção 3.1.2 demonstraram alta variância nas pontuações obtidas por cada estratégia. Esta observação nos motivou a usar a Análise de Componentes Principais (PCA) [Pearson 1901] para extrair informações úteis e ortogonais modeladas por elas.

PCA é uma técnica estatística multivariada que explora a variabilidade da estrutura dos dados [Pearson 1901]. Sua principal ideia consiste em reduzir a dimensionalidade de um conjunto de dados que apresenta um grande número de variáveis correlacionadas, enquanto captura, tanto quanto possível, a variabilidade dos dados originais. O PCA realiza esta redução pela transformação das variáveis em um novo conjunto de variáveis ortonormais, chamadas de Componentes Principais. Estes componentes são ordenados de forma decrescente pela quantidade de variabilidade que cada um modela. Em nossas análises, processamos o PCA utilizando uma matriz de entrada P de dimensões $|U| \times |S|$, onde U é o conjunto de usuários e S é o conjunto de estratégias avaliadas. Cada posição $P(u_i, S_j)$ refere-se a pontuação de influência que a estratégia S_j atribuiu ao usuário u_i. Uma vez que um dos nossos objetivos é obter uma única métrica que capta todas as informações não redundantes modeladas por cada classe, avaliamos apenas o primeiro componente principal do PCA. O primeiro componente pode ser interpretado como uma combinação linear das estratégias. Assim, derivamos uma nova pontuação de influência para cada usuário u_i que é uma combinação linear de todas as pontuações atribuídas a u_i pelas estratégias distintas de cada classe. Como a nossa taxonomia tem três classes, criamos três matrizes distintas de entrada P, cada uma contendo apenas as estratégias de cada classe (ou seja, $|S| = 2$).

Mais uma vez, para cada classe, elaborou-se uma lista ordenada decrescente dos Top-K usuários influentes, de acordo com as novas pontuações definidas pelo PCA e

comparou-se esta lista com as listas geradas por cada estratégia pertencente a mesma classe, usando a métrica Kendall's Tau. O objetivo é medir a concordância entre a nova lista gerada por PCA com as demais estratégias da mesma classe. A figura 1 apresenta os resultados desta análise, variando-se os Top-K usuários influentes de 5 a 50.



(a) ***Estrutura da Rede*** **(b)** ***Conteúdo e Fluxo*** **(c)** ***Sumário Estatístico***

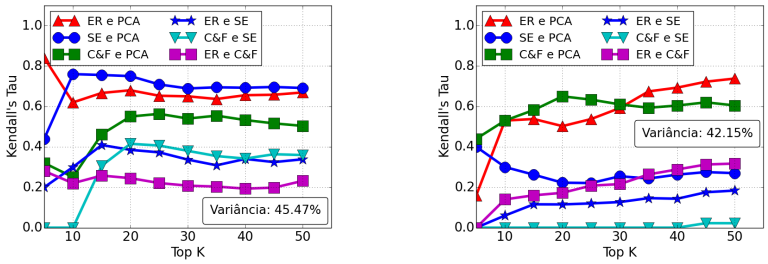
Figura 1. Análise da complementaridade da informação modelada por distintas estratégias pertencentes a mesma classe usando PCA.

Observa-se que, na maioria dos casos, as estratégias que pertencem a mesma classe (por exemplo, linhas verdes) apresentam níveis de concordância superior a 40%. Além disso, quando se compara os Top-K usuários influentes derivados do PCA com outras estratégias que pertencem a mesma classe, os níveis de concordância são ainda maiores em ambas as coleções. Para todos os conjuntos de dados e classes, encontramos pelo menos 80% de concordância entre PCA e uma das estratégias. Essas observações têm duas implicações principais. Em primeiro lugar, as estratégias que pertencem a mesma classe estão fortemente correlacionadas, mais uma vez, corroborando a taxonomia proposta. Em segundo lugar, os Top-K usuários influentes identificados pela PCA foram capazes de sintetizar corretamente as características exploradas por cada classe. O primeiro componente principal foi capaz de capturar mais de 66% da variabilidade dos dados em todos os casos (isto é, a variância apresentada nos gráficos). Além disso, podemos interpretar o uso de PCA como uma estratégia de meta-aprendizagem que combina linearmente estratégias distintas de uma mesma classe, a fim de compor uma estratégia única carregando informações úteis para identificar os usuários influentes. Na verdade, esta é uma das principais contribuições deste trabalho.

5. Análise de complementaridade inter-classe

Nesta seção, avaliamos a complementaridade entre as estratégias advindas de classes distintas e como o PCA seria capaz de combiná-las. Tal como feito na seção 4, foi criada uma matriz P de usuários por estratégias para cada amostra de dados. A posição $P(u_i, S_j)$ representa o grau de influência atribuído pela estratégia S_j ao usuário u_i. Foram usadas todas estratégias avaliadas para compor P (i.e., $|S| = 6$). Então foi aplicado o PCA sobre P, derivando um único grau de influência para cada usuário distinto, que representa uma combinação dos graus de influência originalmente definidos por cada uma das estratégias. Para cada amostra de dados, foi obtida uma lista ordenada decrescente

dos Top-50 usuários influentes, de acordo com o novo grau de influência definido pelo PCA. Contrastou-se a lista gerada pelo PCA, utilizando todas as estratégias, com as listas resultantes da aplicação do PCA sobre cada classe individualmente. Usando a métrica de Kendall's Tau para determinação dos níveis de concordância, variou-se os Top-k usuários influentes de 5 a 50. Os resultados das análises são apresentados na figura 2.



(a) Coleção 1. (b) Coleção 2.

Figura 2. Análise da complementaridade de informação modelada pelas estratégias pertencentes a classes diferentes usando PCA

Observou-se que as listas obtidas pelo PCA sobre classes distintas tem pouca concordância entre elas (i.e., menor do que 40%). Este resultado demonstra que as estratégias pertencentes a classes distintas exploram informações diferentes, tal que os Top-k usuários influentes resultantes de cada estratégia diferem uns dos outros, dificultando a tarefa de selecionar a melhor estratégia dado a falta de consenso. Por outro lado, a lista obtida pelo PCA aplicado a todas estratégias, na maioria dos casos, sintetizou melhor a informação modelada por diferentes classes, apresentando níveis de concordância superiores a 50%. A variância capturada pelo primeiro componente principal tende a ser menor (em torno de 40%), dado o número de estratégias não-alinhadas. Assim, o PCA pode ainda ser usado como uma estratégia de meta-aprendizado, para combinar linearmente estratégias de naturezas distintas. Acreditamos que esta é a principal contribuição deste trabalho, dado que o PCA pode facilitar a tarefa de escolher qual estratégia a ser adotada por pesquisadores de difusão de informação para se obter uma lista apropriada de usuários influentes.

6. Conclusões e trabalhos futuros

Neste artigo, apresentamos um estudo quantitativo de análise e comparação entre algumas das principais estratégias para identificar usuários influentes em aplicações de mídia social [Ilyas and Radha 2011, Silva et al. 2013, Lee et al. 2010, Wu et al. 2011, Bakshy et al. 2011]. Baseado nas principais características exploradas, agrupamos estas estratégias de acordo com o tipo de informação modelada, definindo uma nova taxonomia. Utilizando duas coleções reais, obtidas a partir do Twitter, avaliamos esses grupos. Foi gerada uma lista ordenada de forma decrescente dos Top-50 usuários influentes, de acordo com as pontuações definidas por cada estratégia. Comparando-se estas listas, foi confirmado que as listas derivadas de estratégias que pertencem à mesma classe apresentam o maior nível de concordância. Além disso, utilizando Análise de Componentes Principais (PCA), foi possível analisar a complementaridade das informações modeladas por estratégias distintas, demonstrando que estratégias pertencentes à mesma classe estão fortemente correlacionadas. Utilizamos também o PCA como um processo de meta-aprendizagem para combinar estratégias linearmente distintas. Acreditamos que a comparação e combinação de estratégias distintas corresponde à principal contribuição

deste trabalho, auxiliando pesquisadores da área na tomada de decisão sobre qual das estratégias usar. Como trabalho futuro, pretendemos estender a taxonomia proposta inspecionando outras estratégias existentes na literatura.

Referências

Bakshy, E., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Everyone's an influencer: Quantifying influence on twitter. In *Proc. 4th ACM WSDM*. ACM.

Bonacich, P. and Lloyd, P. (2001). Eigenvector-like measures of centrality for asymmetric relations. *Social Networks*, 23(3):191–201.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*.

Davis, C. (1962). The norm of the schur product operation. *Numerische Mathematik*.

Fagin, R., Kumar, R., and Sivakumar, D. (2003). Comparing top k lists. In *Proc. 14th ACM-SIAM 2003*.

Galeotti, A. and Goyal, S. (2010). The law of the few. *The American Economic Review*.

Golub, G. H. and Loan, C. F. V. (1996). *Matrix Computations*. Johns Hopkins University Press.

Ilyas, M. U. and Radha, H. (2011). Identifying influential nodes in online social networks using principal component centrality. In *Communications (ICC), IEEE International Conference*. IEEE.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*.

Lee, C., Kwak, H., Park, H., and Moon, S. (2010). Finding influentials based on the temporal order of information adoption in twitter. In *Proceedings of the 19th International Conference on World Wide Web*. ACM.

Li, Y.-M., Lin, C.-H., and Lai, C.-Y. (2010). Identifying influential reviewers for word-of-mouth marketing. *Electronic Commerce Research and Applications*, (4).

McCown, F. and Nelson, M. L. (2007). Agreeing to disagree: Search engines and their public interfaces. In *Proc. 7th ACM/IEEE-CS JCDL 2007*. ACM.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical report.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*.

Silva, A., Guimarães, S., Meira, Jr., W., and Zaki, M. (2013). Profilerank: Finding relevant content and influential users based on information diffusion. In *Proceedings of the 7th Workshop on Social Network Mining and Analysis*. ACM.

Subbian, K., Sharma, D., Wen, Z., and Srivastava, J. (2013). Finding influencers in networks using social capital. In *Proc. of 2013 IEEE/ACM ASONAM*.

Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proc. the 20th WWW 2011*.