Dinâmica de Temas Abordados no Twitter Via Evolução de Clusters

Priscila R. F. Rodrigues[1], Ticiana Coelho da Silva[1], Flávio R. C. Sousa[2], Regis P. Magalhães[1], Jose A. F. de Macêdo[2]

Universidade Federal do Ceará (UFC)

[1]Quixadá – CE – Brasil
[2] Fortaleza – CE – Brasil

priscila.rfr@alu.ufc.com, {ticianalc, sousa, regismagalhaes}@ufc.br, jose.macedo@lia.ufc.br

Abstract. *By monitoring and analyzing the subjects' evolution of social network along time is of key importance for users or organizations responsible for decision making. This work focus on investigating the subjects transitions in social medias over time, aiming at achieving an overview and understanding of the motivations of such evolutions. Therefore, this paper proposes to monitor and analyze postings during time windows via clusters evolution. The experiments were performed using data obtained from the Twitter and demonstrate that the proposal is a promising solution to monitor subjects evolution patterns over time.*

Resumo. *Ao monitorar e analisar a evolução de assuntos da rede social ao longo do tempo é de fundamental importância para usuários ou organizações responsáveis por tomadas de decisão. Este trabalho foca na investigação das transições de assuntos em redes sociais ao longo do tempo, objetivando alcançar uma visão panorâmica e compreender as motivações de tais evoluções. Para isso, este trabalho propõe monitorar e analisar postagens em janelas de tempo por meio da evolução de clusters. Os experimentos foram realizados utilizando dados obtidos do Twitter e demonstram que a solução proposta é promissora para acompanhar os padrões de evolução de assuntos ao longo do tempo.*

1. Introdução

Devido à grande quantidade de conteúdo produzido nas redes sociais, o monitoramento e análise desses dados de forma não automatizada consistem em um problema não trivial. Dentre as técnicas que podem ser utilizadas para a análise desses dados, destaca-se a clusterização, que consiste no agrupamento de dados multidimensionais de um conjunto de classes, denominadas *clusters*, com base em uma função de similaridade [Jain et al 1999]. Um *cluster* pode sofrer alterações ao longo do tempo em virtude da constante evolução e consequentes impactos nos mecanismos de geração dos dados [Kaur et al. 2009]. A constatação da volatilidade dos dados ao longo do tempo, aliada à consciência de que é mais relevante compreender a dinâmica e a natureza da evolução, do que meramente identificá-la [Spiliopoulou et al 2006], foi a principal razão que motivou à escolha do tema em estudo.

Dentre outras abordagens, o estudo da evolução de *clusters* pode ser direcionado para a descoberta e observação do modo como os grupos sociais tendem a evoluir por meio de uma dimensão temporal, favorecendo tarefas como a publicidade dirigida e a personalização de conteúdos e serviços, adequando-os às necessidades e preferências dos consumidores ao longo do tempo.

A fim de demonstrar a aplicabilidade da utilização de técnicas de evolução de *clusters* em bases de dados altamente dinâmicas como as de redes sociais, neste artigo aplicam-se técnicas de clusterização em *tweets* coletados em 2015 relacionados a presidenta do Brasil, Dilma Rousseff. Esse assunto foi escolhido, em função da grande repercussão, tendo em vista as manifestações que estavam ocorrendo no país neste período, a fim de verificar a popularidade da presidenta ao longo dos dias. Por meio do monitoramento da evolução dos *clusters*, apresenta-se a dinâmica e transição dos assuntos detectados em torno desse tema, objetivando alcançar uma visão panorâmica e compreender as motivações de tais evoluções. O algoritmo de evolução de *clusters* empregado neste trabalho foi proposto por [Coelho Silva et al 2014], e detecta as seguintes transições: Criação, Sobrevivência, Desaparecimento, União, Divisão, Expansão e Retração de *clusters.*

Este artigo está organizado da seguinte forma: Na Seção 2 é apresentada a estratégia de Evolução de *Clusters* empregada nos dados do Twitter. Em seguida, na Seção 3, é descrita a aplicação da estratégia e realizada a análise dos resultados. Na Seção 4 são apresentados os trabalhos relacionados. Finalmente, a Seção 5 apresenta as conclusões obtidas até o momento e as direções futuras desta pesquisa.

2. Estratégia de Evolução de *Clusters* em dados do Twitter

Para a aplicação da técnica de clusterização, o algoritmo DBSCAN [Ester et al 1996] foi implementado e aplicado utilizando como base a medida de similaridade *Fading* [Lee et al 2014]. Tal medida considera o atributo de tempo para o cálculo da similaridade, tendo em vista quais publicações mais próximas no tempo têm maior possibilidade de versarem sobre o mesmo assunto. Formalizando p_i^L como uma lista de palavras em um momento i, p_j^L como uma lista de palavras em um momento j no tempo, [Lee et al 2014] usa uma função exponencial para incorporar o efeito do tempo decorrido entre as publicações, sendo a similaridade *Fading* definida como:

$$SF\left(p_{i,}\, p_j\right) = \frac{|\, p_i^L \cap\, p_j^L|}{\left|\, p_i^L \cup\, p_j^L\right|.\; e^{|p_i^\tau - p_j^\tau|}}$$

Para detecção das evoluções sofridas pelos *clusters* ao longo dos dias, foi implementado e executado o algoritmo de evolução proposto por [Coelho Silva et al 2014]. O algoritmo recebe como entrada uma matriz M onde foram armazenados em cada célula o valor de similaridade entre dois *clusters* em diferentes momentos do tempo. O algoritmo recebe também como entrada C_t, que são todos os *clusters* encontrados no tempo t, e $C_{t+\delta t}$, todos os *clusters* encontrados no tempo t + δt. O outro parâmetro passado é o τ que consiste no *threshold* de similidaridade, ou seja, o limiar que define o número mínimo de elementos em comum que dois *clusters* (um do tempo t e outro do tempo t + δt) devem possuir para serem considerados similares. Por meio de comparações entre os *clusters*, o algoritmo detecta criação, sobrevivência, desaparecimento, união, divisão, expansão e retração de *clusters.*

3. Aplicação da estratégia e Análise dos Resultados

3.1. Coleta e Pré-processamento dos Dados

Os dados do Twitter foram coletados por meio da implementação de um *crawler*. Este recebeu como parâmetro a palavra "Dilma" e retornou *tweets* que continham essa palavra. A análise dos dados foi referente à coleta realizada entre os dias 08/03/2015 e 16/03/2015. Nesse período foram coletados 672.551 *tweets* coletados, o que resultou em aproximadamente 512Mb de dados.

Inicialmente foi aplicada a remoção dos *stopwords,* uma técnica de Processamento de Linguagem Natural para remoção de palavras dotadas de pouco valor semântico e com elevada recorrência em qualquer texto, não expressando conteúdo significativo dentro do *tweet*. Para aplicação desta técnica foi utilizada a biblioteca *Natural Language Toolkit* (NLTK)[1]. Foi removido também dos *tweets* a palavra "Dilma", em função de ter sido utilizada para a captura dos dados e estar presente em todos os *tweets* coletados. Desta forma, após o processamento restaram apenas as palavras chaves na composição do *tweet*.

3.1. Clusterização e Evolução de *Clusters*

Para avaliar a evolução, os *tweets* coletados foram decompostos em uma série de *snapshots*, onde cada *snapshot* corresponde a um dia de coleta. Com base nos experimentos e valores adotados em [Lee et al 2014], foram repassados como parâmetros para o algoritmo de clusterização *eps*: 0.3 e *minPts*: 0.25 * (quantidade de *tweets* do *snapshot)*. No contexto deste trabalho o *eps* é um valor limite para a similaridade de dois *tweets*. Logo, o resultado obtido após a aplicação da medida de similaridade entre dois *tweets* deve ser igual ou maior que 0.3 para que sejam considerados similares. Já *minPts* é o parâmetro que define a densidade mínima de um *cluster.* Dessa forma, um *cluster* agrupa pelo menos 25% dos *tweets* do *snapshot*.

A Figura 1 apresenta, em linha temporal, os *clusters* detectados ao longo dos dias de análise. É possível notar que a quantidade de *tweets* sobre os temas tem grandes variações ao longo dos dias.

Realizada a clusterização, foi executado o algoritmo de evolução de *clusters,* onde o valor do parâmetro τ (*threshold* de similaridade) foi de 0.3, ou seja, dois *clusters* são similares se possuírem pelo menos 30% de elementos em comuns, valor adotado com base nos experimentos de [Lee et al 2014]. Após a execução do algoritmo, foram identificadas as transições sofridas pelos *clusters* ao longo do tempo, sendo possível observar como os assuntos repercutiram e se relacionaram, conforme descrito a seguir e apresentado na Figura 2.

No dia 08 de março foram detectados dois assuntos: um sobre o Dia da Mulher, vaia que ocorreu no momento do pronunciamento realizado pela presidenta para as mulheres naquele mesmo dia[2]. Tais assuntos polarizaram apoiadores e críticos do governo petista. No dia seguinte, os dois assuntos passaram a repercutir juntos em um mesmo *cluster* que se referia ao "panelaço"(como ficou conhecido o protesto realizado

[1] http://nltk.org/
[2] http://g1.globo.com/politica/noticia/2015/03/pessoas-protestam-durante-pronunciamento-de-dilma.html

no momento do pronunciamento no Dia da Mulher)[3], sendo detectada pelo algoritmo **uma união de *clusters***. O assunto marcado pela *hashtag* "#NãoQueroMorarNoBrasilpq" também foi **um *cluster* que apareceu** dia 09 de março no Twitter, mas no dia seguinte **desapareceu**. Nos dias 10 e 11 de março o *cluster* sobre o Panelaço passou a sofrer **retrações**, até passar por uma **divisão** no dia 12 que distinguiu a repercução entre apoiadores e opositores do governo marcados pela hashtag "#menosOdioMaisDemocracia" e pedidos de impeachment. O *cluster* que repercutia favorável ao governo possuía maior densidade de *tweets,* por fazer menção a uma manifestação que estava sendo organizada para o dia seguinte por apoiadores do governo Dilma. No dia 13/03, a discussão que repercutia a favor do governo **sofreu ainda uma maior expansão** repercutindo a *hashtag* "#Dia13Diadeluta" e "#GloboGolpista", tendo em vista que nesse mesmo dia estava ocorrendo em todo o país uma manifestação pró-Dilma organizada pela CUT (Central Única dos Trabalhadores)[4]. Ainda no dia 13, o *cluster* que repercutia com pedidos de impeachment sofreu uma pequena **expansão.**

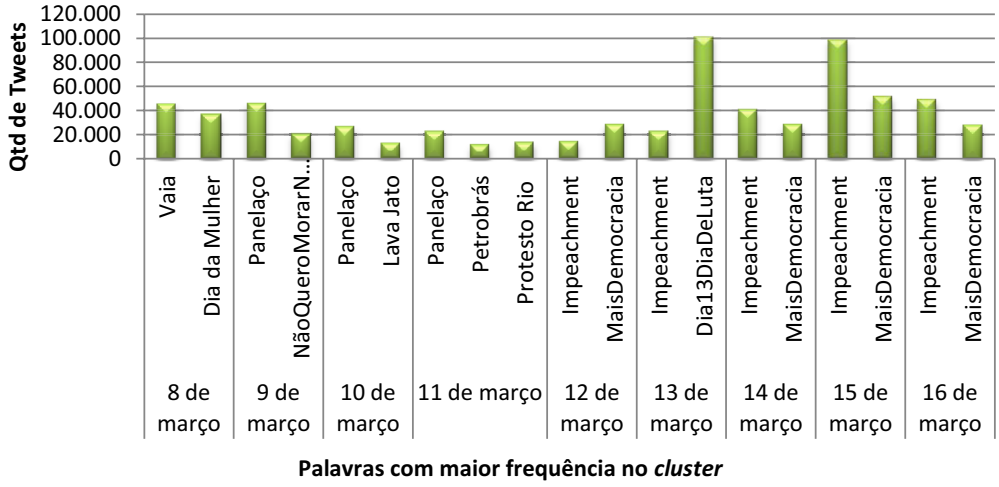

Figura 1 - Clusters gerados de 08/03/15 à 16/03/15

Já no dia 14/03, o *cluster* favorável ao impeachment **continuou a expandir**, ao passo que o *cluster* positivo ao governo que repercutia a hashtag "#MenosOdioMaisDemocracia" sofreu retração. No dia 15/03, dia do protesto realizado nacionalmente contra o governo Dilma[5], a repercussão em torno das manifestações a favor do impeachment sofreu **grande expansão**. O *cluster* favorável ao governo, embora tenha expandido em comparação com o dia anterior, não alcançou a expansão do *cluster* a favor do impeachment. Nos dias seguintes a este evento, os dois assuntos **sofreram retração** e continuaram a evoluir de forma independente.

[3] http://www1.folha.uol.com.br/poder/2015/03/1600073-em-cidades-com-panelaco-internautas-tambem-defendem-dilma.shtml
[4] http://www.cut.org.br/noticias/cut-mobilizada-para-o-dia-13-de-marco-664c/
[5] http://www1.folha.uol.com.br/especial/2015/protestos-15-de-marco/

Figura 2 - Gráfico de Transições dos assuntos sobre "Dilma"- 08/03/15 à 15/03/15

4. Trabalhos Relacionados

[Kim e Han 2009] propõe decompor uma base de dados em uma série de *snapshots* (conjunto de dados em um ponto específico no tempo), depois aplicar algoritmos de mineração em cada *snapshot* para encontrar padrões úteis, combinando esses padrões para gerar uma sequência de padrões dinâmicos. Entretanto, a proposta não oferece suporte à atividades como divisão e união de *clusters*, transições comuns em dados dinâmicos [Lee et al 2014] e de elevada importância para a completa compreensão das evoluções sofridas pelos *clusters*.

Em [Tang et al 2013] a abordagem de evolução de *clusters,* compreendendo todas as transições, é aplicada em dados dinâmicos de trajetória produzidos por meio de tecnologias móveis, com o intuito de descobrir objetos que se movimentam juntos. Finalmente, [Lee et al 2014] propõe uma abordagem para detectar a evolução de assuntos em dados dinâmicos de redes sociais, observando todas as evoluções de *clusters*. Porém, sua estratégia aplica a característica *fading time window* para remoção de dados. O fato deste trabalho não empregar essa característica e fazer uma análise de dados correntes, permite a flexibilidade de ter uma estratégia que define a janela de tempo de acordo com a aplicação, como no experimento realizado onde foi adotado janelas de tempo de um dia para cada *snapshot,* considerando ser este um período razoável para analisar as transições das opiniões do eleitores e os eventos repercutidos no período. Além disso, este trabalho faz uso de Processamento de Linguagem Natural

(PLN) sobre os dados coletados. Tal técnica não foi utilizada na proposta de [Lee et al 2014], mas é de extrema relevância por possibilitar um maior refinamento dos dados, removendo impurezas comuns em dados textuais de escrita livre como os de redes sociais.

5. Conclusões e Direções Futuras

Os resultados apresentados reforçam a aplicabilidade do uso das técnicas de evolução de *cluster* como um recurso promissor para monitorar as transições de assuntos nas redes sociais. Vale ressaltar que este tipo de aplicação é recente no contexto de redes sociais. No entanto, apresenta-se bastante útil para usuários interessados em acompanhar a evolução de determinados acontecimentos ao longo do tempo, sobretudo para aqueles usuários ou órgãos responsáveis por tomadas de decisão em relação ao assunto sobre o qual as evoluções estão sendo monitoradas.

Como etapas subsequentes, serão realizados outros estudos de caso para melhor validar o trabalho proposto. Comparar a abordagem proposta com outras presentes na literatura. Considerando uma melhor exatidão na detecção e evolução de *clusters*, pretende-se propor uma medida de similaridade direcionada ao contexto de redes sociais. Uma medida que não avalie apenas a variável de tempo e similaridade sintática entre os textos, mas informações do contexto da publicação (localidade, usuário que a postou). Além disso, pretende-se realizar um estudo mais profundo da variação dos parâmetros utilizados no DBSCAN para este contexto de clusterização em redes sociais, bem como, experimentar diferentes granularidades no tamanho do *snapshot*. Por limitação de espaço, não foi possível apresentar os resultados para outros *datasets*.

Referências

Ester, Martin et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd. 1996. p. 226-231.

Jain, Anil K.; Murty, M. Narasimha; Flynn, Patrick J. Data clustering: a review. ACM computing surveys (CSUR), v.31, n. 3, p. 264-323, 1999.

Kaur, S. et al. Concept drift in unlabeled data stream. Technical Report, University of Delhi, 2009

Kim, Min-Soo; Han, Jiawei. A particle-and-density based evolutionary clustering method for dynamic networks. Proceedings of the VLDB Endowment, v. 2, n. 1, p. 622-633, 2009.

Lee, Pei et al. Incremental cluster evolution tracking from highly dynamic network data.In: Data Engineering (ICDE). IEEE, 2014. p. 3-14.

Coelho da Silva, Ticiana L., José AF de Macêdo, and Marco A. Casanova. "Discovering frequent mobility patterns on moving object data. "Proceedings of the Third ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems. ACM, 2014.

Spiliopoulou, Myra et al. Monic: modeling and monitoring cluster transitions. In: Proceedingsof the 12th ACM SIGKDD. ACM, 2006. p. 706-711.

Tang, Lu-An et al. A framework of traveling companion discovery on trajectory data streams. ACM Transactions on Intelligent Systems and Technology (TIST), v. 5, n. 1, p. 3, 2013.