Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Jairo Ataide

Explorando análises descritivas do Twitter

Curitiba
2019

Jairo Ataide

Explorando análises descritivas do Twitter

Monografia apresentada ao Programa de Especialização em *Data Science* e *Big Data* da Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.

Orientador: Prof. Luis Carlos Erpen de Bona

Curitiba
2019

Especialização em Data Science & Big Data
dsbd.leg.ufpr.br

Universidade Federal do Paraná
Trabalho de Conclusão de Curso, nº 09, 01 (2019)

Explorando análises descritivas do Twitter

Uma ferramenta para as Ciências Sociais

***Jairo Ataide*[1]**

Departamento de Estatística - Universidade Federal do Paraná Campus III - Centro Politécnico Rua Evaristo F. F. da Costa, 418 Jardim das Americas Curitiba - PR

Resumo

Este trabalho parte da questão da Teoria das Redes e dos Grafos em uma perspectiva da análise estrutural de relações sociais no Twitter, sendo este fonte massiva de dados para a exploração nas ciências sociais, visto que tamanho e a complexidade destes dados podem ser apresentados de forma a melhorar a habilidade de analisar tais volumes. Este trabalho apresenta uma estrutura projetada para facilitar a pesquisa neste campo, fornecendo um conjunto de recursos necessários para extração e análise pelos pesquisadores.

Palavras-Chaves: Big Data, Twitter, Redes Sociais, Grafos, Sociograma, sociometria.

Abstract

This paper is based on Networks and Graphs Theory in a structural analysis perspective regarding the social relations on Twitter as it is a massive data source to be explored by Social Sciences. The size and complexity of these data can be presented in order to improve the ability to analyze such volumes. Here it is being presented a designed structure to help and easywell/facilitate this field research providing a resources set to extract and good analysis by the researchers..

Keywords: Bid Data, Twitter, Social Networks, Graphos, Sociogram, Sociometry.

1. Introdução

O Twitter é uma das redes sociais mais populares do mundo. Parte do recurso é a capacidade dos usuários de seguir qualquer outro usuário com um perfil público, permitindo que os usuários interajam entre si e com empresas. O Twitter também se tornou um importante canal de comunicação do mercado e governos. Uma fonte inesgotável de dados para pesquisas nas ciências sociais.

A aplicação de infraestrutura para ciência de dados e Big Data no contexto das Redes Sociais geralmente é restrita a aplicações comerciais ou soluções internas corporativas, restringindo assim o acesso de pesquisadores a base de dados representações de fácil entendimento.

A coleta automatizada de dados de sites de redes sociais desempenha um papel importante na pesquisa científica. Em vários problema de pesquisa das ciências sociais, o impacto em massa sobre vários aspectos pode ser analisado a partir dos dados gerados destas redes. a partir desses sites. No Twitter, foco deste trabalho, é uma plataforma sociais com informações abertas e amplamente usadas para compartilhamento de visualizações.

Este trabalho apresenta uma estrutura projetada para facilitar a pesquisa neste campo, fornecendo um conjunto de recursos necessários para extrair e analisar informações no Twitter a

1 Jairo.ataide@ufpr.br

partir da perspectiva teórica das Redes e dos grafos como representações visuais de processos sociais envolvidos e o material apresentado não pressupõe habilidades estatísticas do leitor,

Este trabalho está estruturado em 5 seções. Na seção 2 são definidos conceitos básicos de redes sociais, grafos e a sustentação teórica. Na seção 3 são apresentadas a motivação, problematização e objetivo do trabalho. Na seção 4 é apresentada a metodologia e a estrutura utilizada. Na seção 5 são apresentados os principais resultados computacionais. Finalmente na seção 6 são elaboradas as conclusões.

2. Revisão da literatura

Popularmente quando se ouve o termo "Redes Sociais", automaticamente se referência as mídias sócias, tais como Facebook, Twitter, Instagram, etc. Contudo bem anterior a estas já havia incontáveis redes sociais [1].

Sendo que se pode afirmar, de modo abrangente a análise de redes sociais constitui um conjunto de métodos quantitativos que se aplicam a dados relacionais. A análise de redes serve a diversos domínios científicos tendo a matemática como fornecedora da teoria de grafos, da álgebra de matrizes, e da teoria de probabilidades e atendendo a Psicologia, Sociologia, Ciências da Saúde, Ciências da Administração etc. [1].

O comportamento humano sempre foi estruturado através da interdependência das interações e de tudo aquilo que pressupõe a sua existência. Isso se dá pois não se consegue mais imaginar um mundo sem o uso massivo de dados da internet e isso faz com que o dado científico relevante confunde-se com o que parece ser a nova natureza digital das relações sociais construídas pelo uso cotidiano dos ordenadores de dados [1].

Ao que se refere as redes sócias digitais, estas potencializaram a expansão das conexões sociais de cada indivíduo e propiciando acesso imediato à informação, sensação de pertencimento e tempo escasso têm levado estes a expandirem sua "coleção de amigos" e fazendo com que os processos de decisão, individuais e coletivos, estejam mais estreitamente relacionados [2].

As redes sociais digitais foram concebidas como um meio de comunicação e socialização, não apenas entre parentes, amigos e pequenas comunidades, mas também como uma "sala de reunião" virtual em tempo real para estranhos se encontrarem e discutirem. Seu objetivo é apoiar uma estrutura social através da Internet, a fim de satisfazer a necessidade de comunicação entre indivíduos e/ou organizações [3].

Rheingold defende enfaticamente o nascimento de um novo conceito de comunidade, que reúne os indivíduos em torno de interesses e valores em comum. Para ele, a forte afinidade entre seus membros é fundamental para dar sustentabilidade à própria comunidade, além de uma massa crítica de usuários debatendo temas tão interessantes que sejam capazes de atrair outros usuários garantindo renovação e perpetuação. Rheingold define comunidade virtual a partir de seu próprio entendimento de comunidade, como sendo uma teia de relacionamentos mantida ao longo do tempo entre pessoas que se preocupam umas com as outras. Os recursos da tecnologia digital capacitam, melhoram ou amplificam essa rede de relacionamentos. [4] (apud [2])

Granovetter, autor seminal da teoria dos Laços Fortes e Laços fracos [5], [6] em época muito anterior as mídias digitais com a vimos atualmente, descreve as redes de Lações Fortes como formadas por indivíduos com uma identidade comum, cujas dinâmicas geradas nessas interações não se estendem além dos agrupamentos por estes formados, são redes sustentadas pelas referências para a tomada de decisão e são relações com alto nível de credibilidade e influência. E indivíduos que compartilham Laços Fortes comumente participam de um mesmo círculo social. ao passo que os indivíduos com os quais se tem relações de Laços Fracos são importantes porque conectam com vários outros grupos, rompendo a configuração de agrupamentos isolados e assumindo assim a configuração de rede social. Nesse sentido, as relações baseadas em Laços Fortes levam a uma topologia da rede, isto é, definem a configuração dos nós da rede de conexões entre os indivíduos , no qual as relações de Laços Fracos funcionam como pontes entre grupamentos isolados. Granovetter sustenta que por meio das relações de Laços Fracos os indivíduos são expostos à inovação, mas para abraçá-la dependem do aval de suas relações de Laços Fortes;

ou pensar os Laços Fracos como redes eficientes no transporte de informação, mas não tão eficientes para provocar uma tomada de decisão.

Embora as teorias de Granovetter [5], [6] sejam anteriores ao advento das mídias digitais na escala que são utilizadas atualmente, como já citado anteriormente, Kaufman [2] descreve que as conexões das mídias digitais guardam semelhanças com a definição de Laços Fracos de Granovetter, ao desempenharem o papel de propagadores de inovações, difundindo referências e experiências, facilitadas pela tecnologia que amplia o acesso e acelera as interações com um número maior de pontos de contato.

Novos tempos trazem novos desafios e com eles novas metodologias de investigação são criadas, reestruturadas ou simplesmente retomadas na tentativa de explicar determinada realidade. A metodologia de análise de rede é apenas mais um desses [7]. Considerando o campo das Ciências Sociais, foco deste trabalho, rede seria o conjunto de relações sociais entre um conjunto de atores e também entre os próprios atores. Designa ainda os movimentos pouco institucionalizados, reunindo indivíduos ou grupos numa associação cujos limites são variáveis e sujeitos a reinterpretações [8] (apud [9]). e para a Antropologia Social a noção de redes sociais busca apoiar "a análise e descrição daqueles processos sociais que envolvem conexões que transpassam os limites de grupos e categorias" [10, p. 163] (apud [9]).

Estudos visando explorar a descoberta de conhecimento em redes sociais envolvem áreas diversas como análise de links, mineração de dados, mineração de grafos, aprendizagem de máquina e técnicas de visualização. Especificamente, o princípio fundamental na visualização de redes sociais é facilitar a compreensão dos dados. As técnicas existentes direcionam-se a selecionar subgrupos de informações com o objetivo de simplificar a visualização [11].

A linguagem dos grafos é a ferramenta fundamental para descrever a morfologia das redes sociais, ou seja, a representação da sua forma e configuração. Teoria de Grafos é um tipo de geometria, conhecida de forma específica como topologia, de grande expansão e com diversos usos aplicados, especialmente em Ciências da Computação e nas Ciências Sociais. A geração de um grafo pressupõe um procedimento matemático formal que consiste numa matriz de adjacência na qual se registram as relações entre os agentes do sistema. [1].

A Teoria de Redes ou Análise de Redes Sociais, teve a sua fundamentação matemática na Teoria dos Grafos de forma que uma rede pode ser expressa matematicamente por um grafo que é constituído por um conjunto de pontos, os nós ou vértices, conectados por linhas que expressam uma relação entre eles, as arestas. [12], [13]

Neste trabalho um dos dos conceitos fundamentais a serem utilizados são as medidas de centralidade relacionada ao conceito de poder e de influência dos elementos adjacentes numa rede, os vértices mais centrais são aqueles a partir dos quais podemos atingir qualquer outro com mais facilidade ou rapidez. As medidas de centralidade identificam a posição de um individuo relativamente a outros na sua rede. As medidas de centralidade são centralidade de Grau, Proximidade, Intermediação, Vetor Próprio e Katz [13].

Para efeito da formação do grafo representado na 1 a medida de centralidade a considerar é por Grau, que diz respeito à centralidade de um vértice é o número de contatos diretos que ele possui. Um elemento se encontra numa posição que permite o contato direto com muitos outros é vista pelos demais como um canal maior de informação, razão pela qual dizemos ser mais central. Assim, a centralidade de grau nada mais é que a contagem do número de adjacências de um vértice [13].

Outro fator a se considerar na formação do Grafo deste trabalho são os modelos de abordagem macro, micro e bloco. Sendo o que foi considerado para efeito da formação do grafo representado na 1 foi o modelo em bloco.

O modelo de abordagem macro (sociocentrada) põe-se enfase no todo e não apenas nos elementos que compõem a rede. A abordagem micro (egocêntricas) Quando o interesse de investigação recai sobre o indivíduo (ego) e as relações que este desenvolve dentro de uma determinada rede e modelação em bloco (block-modeling) são aplicadas quando se pretende colapsar uma rede complexa em busca de um denominador comum [7].

3. Motivação, problematização e objetivo

A disponibilidade pública de uma quantidade imensa de dados disponíveis nas redes sociais abre um vasto campo de pesquisa.

Muitos pesquisadores das áreas das ciências sociais necessitam utilizar dados de redes sociais digitais em suas pesquisas e apresentam dificuldades da extração e disponibilização de dados estatísticos destes.

Este trabalho propõe o desenvolvimento de uma ferramenta digital, em código aberto, para extração, processamento e análise de comentários do Twitter de forma a permitir que pesquisadores o utilizem de forma simples e personalizável de modo a obter grafos e estatísticas descritivas dos dados que facilitem e automatizem de forma eficiente análises com base nos dados.

4. Metodologia

Este trabalho foi formulado por meio do estudo, conceituação e extração de informação de muitos trabalhos de pesquisa anteriores. A literatura publicada relacionada à Análise de Redes Sociais.

Neste trabalho é guiado a abordagem de Pesquisa-Ação.

> "A Pesquisa-Ação em contexto profissional "é uma proposta de pesquisa mais aberta (com características de diagnóstico e de consultoria), para tentar clarear uma situação complexa e encaminhar possíveis ações" [14, p. 20]

Como metodologia principal foi adotado um *framework* proposto por Cuesta, Barrero, e Moreno [3] os quais desenvolveram uma estrutura para extração massiva e análise de dados do Twitter. Embora a metodologia proposta tenha sido adaptada para os objetivos deste trabalho.

Os tweets estão disponíveis on-line, tanto em formato estático quanto dinâmico. Ambas as formas são disponibilizados por meio de uma API pública fornecida pelo Twitter. Em sua forma estática, é possível consultar os tweets recentes com base em determinados critérios de pesquisa. O resultado para essas consultas é sempre uma visualização estática em um determinado instante no tempo. Na forma dinâmica apresenta um fluxo de tweets, que devem ser filtrados por pelo menos um critério acessível a partir de um "endpoint REST". As requisições podem se conectar a esse endpoint e receber dados não em um único instante no tempo, mas como um fluxo contínuo de dados segundo os critérios de filtragem.

O código foi inteiramente desenvolvido na linguagem R.

Mineração: O núcleo da ferramenta. Extrai os dados de acordo com a aplicação de filtros. Critérios de filtragem disponíveis:

- Termo – Palavra ou expressão a ser pesquisada, com ou sem operadores Boleanos ou de forma exata.
- Filtra por quantidade de termos associados ao termo principal.
- Filtra por período de tempo.
- Filtra a quantidade de Tweets a serem pesquisados.
- Filtra os tipos de Tweets: Recentes, populares ou ambos.
- Filtra a inclusão ou não Retweets.
- Filtra por limites geográficos.
- Resultados retornam ou não como *Data Frames*.
- Filtra por usuário(s) do Twitter.
- Filtra por quantidade de seguidores de usuário(s).
- Filtra por quantidade de seguidores de usuário(s).

O módulo de mineração suporta três modos de operação: a) modo único, que é o padrão de uso padrão da API d. Apenas um filtro está disponível por endereço IP durante o uso deste modo. b) No modo serial, vários filtros são definidos junto com um limite; Os filtros são executados sequencialmente e, quando o limite é atingido, o próximo filtro na lista é aplicado. Isso permite que vários filtros sejam executados no mesmo IP, embora a captura apresente lacunas. O filtro também pode ser configurado para dormir por um número predefinido de segundos. c) O modo paralelo permite vários filtros ao mesmo tempo sem lacunas na captura.

Processamento: A partir do daos da mineração é executado o seguinte *script*:

- Remoção das *Stop Words*.
- Remoção de pontuação.
- Conversão para minusculas.
- Identificação de cada Tweet pelos metadados.
- Contagem de palavras.
- Contagem de seguidores.
- Criação do Data Frame.

Visualização: Criação dos relatórios exemplificados na seção 5.

5. Resultados

As telas abaixo representam algumas das saídas das ferramentes. Todas tem suas fontes os dados coletados pelo REST API do Twitter e processados pelo autor.

A 1 representa a rede de palavras que aparecem conforme o critério de filtro do número de Twiters a serem pesquisados, que neste exemplo foram de 100.000, e a quantidade mínima que essas palavras devem aparecer quando ligadas ao termo principal da busca, que neste exemplo foram de 24 repetições mínimas e filtrados pelos últimos nove dias postagens. A intensidade por centralidade de grau são representadas pelo espectro de cores.

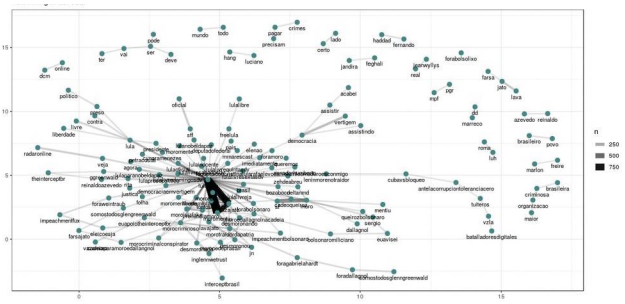

Figura 1: Rede de palavras – Twetts usando a hastag #LulaLivre

A 2 apresenta o detalhe de maior destaque da 1.

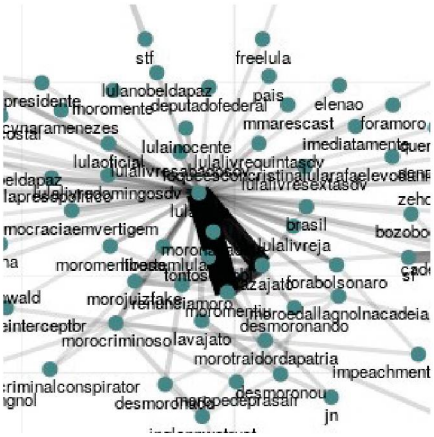

Figura 2: Detalhe 1

A 3 apresenta as localidades dos disparos das postagens conforme os já citados critérios de filtragem.

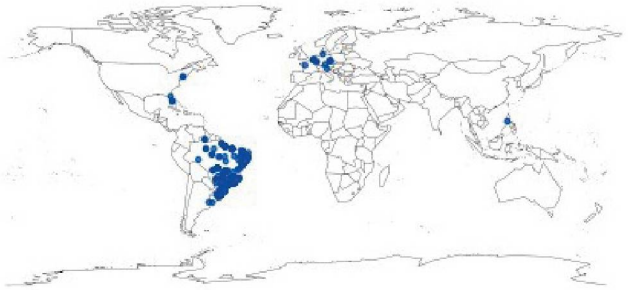

Figura 3: Posição geográfica dos Tweets

A 4 apresenta as palavras mais comuns retornadas conforme os já citados critérios de filtragem e com o limite escolhido das 15 mais relacionadas ao termo chave.

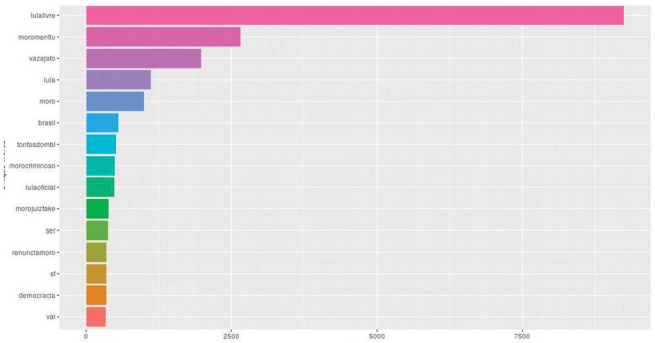

Figura 4: Palavras mais comuns retornadas

A 5 apresenta as hardwares utilizados pelos usuários conforme os já citados critérios de filtragem e com o limite dos 10 mais utilizados.

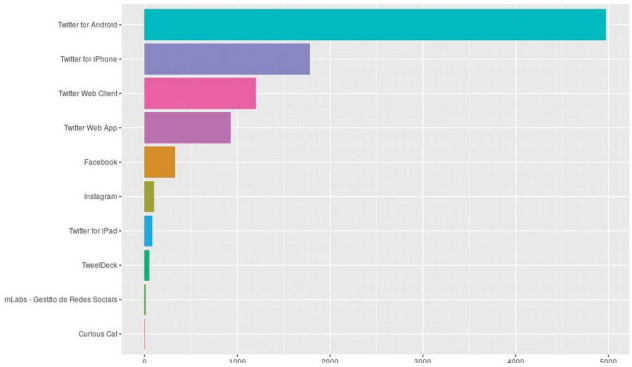

Figura 5: Twetts por fonte

A 6 apresenta a frequência de Tweets retornados conforme os já citados critérios de filtragem e com o limite escolhido dos últimos 9 dias com intervalos de três horas.

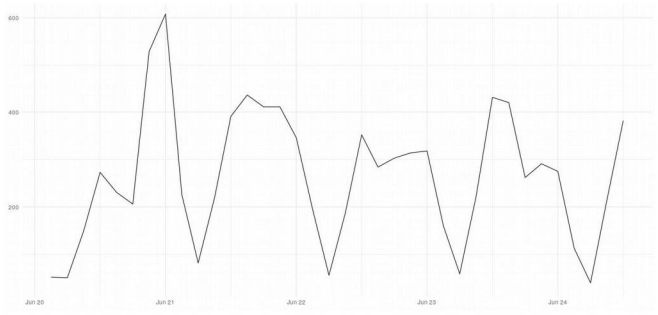

Figura 6: Frequência de Tweets

6. Conclusões

A ferramenta apresentada neste trabalho fornece aos pesquisadores das ciências sociais uma ferramenta para a coleta e análise dos dados Tweets, na qual outras pesquisas podem ser conduzidas. A ferramenta serve ao seu propósito.

A partir das saídas das análises podem ser usadas informação para efeitos no processo de tomada de decisão e ainda ser explorado para estabelecer as relações entre várias situações.

Um trabalho futuro pode ser conduzido no sentido de desenvolver um fronteend em Shyne flexibilizando a interação dos pesquisadores na aplicação dos critérios e filtros.

7. Referências

[1] S. S. Higgins and A. C. A. Ribeiro, "Análise de redes em Ciências Sociais," ENAP, Brasilia, 2018.

[2] D. Kaufman, "A Força dos" Laços Fracos de Mark Granovetter no Ambiente do Ciberespaço," *Galáxia. Rev. do Programa Pós-Graduação em Comun. e Semiótica. ISSN 1982-2553*, no. 23, 2012.

[3] Á. Cuesta, D. F.Barrero, and M. D. R-Moreno, "A Framework for Massive Twitter Data Extraction and Analysis," *Malaysian J. Comput. Sci. Vol 27 No 1 Malaysian J. Comput. Sci.*, Mar. 2014.

[4] H. Rheingold, *Smart mobs: The next social revolution*. Basic books, 2007.

[5] M. S. Granovetter, "The strength of weak ties," *Am. J. Sociol.*, vol. 78, no. 6, pp. 1360–1380, 1973.

[6] M. Granovetter, "The strength of weak ties: A network theory revisited," 1983.

[7] L. A. Santos, "Potencialidades e limitações da metodologia de análise de rede: um modelo teórico voltado para as Ciências Sociais," *Comun. e Soc.*, vol. 33, pp. 183–198, 2018.

[8] A. Colonomos, "Emergence d'un objet et perspectives internacionalistes. In.: CHARILLON, F. et al," *Sociol. des réseaux transnationaux. Paris Ed. L'Harmattan*, 1995.

[9] S. Acioli, "Redes sociais e teoria social: revendo os fundamentos do conceito.," *Informação & Informação*, vol. 12, no. 1esp, pp. 8–19, 2007.

[10] J. A. Barnes, "Social Networks," *Cambridge*, vol. 26, pp. 1–29, 1972.

[11] C. Freitas *et al.*, "Extração de conhecimento e análise visual de redes sociais," *SEMISH-Seminário Integr. Softw. e Hardware, Belém do Pará, Bras. SBC*, pp. 106–120, 2008.

[12] A. S. de Carvalho and E. de Rezende Francisco, "Análise de Redes Sociais e Teoria de Grafos como Suporte para Consumer Insights," in *CLAV 2017*, 2017.

[13] P. A. Laranjeira and L. Cavique, "Métricas de centralidade em redes sociais," *Rev. Ciências da Comput.*, vol. 9, pp. 1–20, 2014.

[14] M. Thiollent, "Pesquisa-Ação nas organizações, Ed," *Atlas, São Paulo, Atlas*, 1997.