Um Estudo de Modelos e Sistemas de Bancos de Dados para Redes Sociais

Mariana Duarte de Araujo Salgueiro[1]

[1]Departamento de Informática
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio)

{msalgueiro, sergio}@inf.puc-rio.br

Orientador: Sérgio Lifschitz (DI PUC-Rio)

Nível: Mestrado

Ingresso: Março de 2021

Previsão de Término: Fevereiro de 2023

Etapas Concluídas: Disciplinas concluídas

Defesa de Proposta: Setembro de 2022 (previsto)

Abstract. *Social networking platforms and applications have become a critical part of the information ecosystem, sparking interest from both an application and research perspective. Several works in the literature treat raw social network data as offered by APIs. In the presence of massive datasets, we expect to work with database management systems (DBMS), which are suitable for dealing with large volumes of data. However, database modeling is necessary to have a logical and coherent collection of data with some inherent meaning. The present research work deals with the actual representation and analysis of social networks information.*

Resumo. *As plataformas e aplicativos de redes sociais se tornaram parte crítica do ecossistema de informações disparando o interesse tanto do ponto de vista de aplicação quanto de pesquisa. Diversos trabalhos na literatura tratam dados de redes sociais brutos da forma como eles vêm oferecidos pelas APIs. Na presença de grandes conjuntos de dados, esperamos trabalhar com sistemas de gerenciamento de banco de dados (SGBD), que são adequados para lidar com grandes volumes de dados. A modelagem de um banco de dados se faz necessária para que exista uma coleção lógica e coerente de dados com algum significado inerente. O presente trabalho de pesquisa se propõe discutir a representação e a análise de informações oriundas de redes sociais.*

1. Introdução

Este trabalho diz respeito a uma pesquisa de mestrado em andamento que tem por objetivo principal estudar como são feitas as análises nas informações oriundas de redes sociais envolvendo grandes volumes de dados.

Estamos supondo que trabalhar com dados obtidos de *APIs* ou *EndPoints* das redes sociais, sem considerar um Sistema Gerenciador de Banco de Dados (SGBD), pode ser pouco eficiente ou mesmo inviável. Uma avaliação de trabalhos na literatura revela que ainda há uso intensivo de arquivos no formato de planilhas (e.g XLS ou CSV) ou JSON. Sabe-se que os SGBDs (em qualquer modelo!) são fortemente recomendados quando os dados crescem em tamanho e complexidade (e.g. [Angles et al. 2013]).

As redes sociais são exemplos de sistemas Big Data e grandes quantidades de dados são geradas em pequenas unidades de tempo. Para realizar análises em um tempo aceitável, os dados precisam ser devidamente persistidos para que, posteriormente, seja disponibilizado um acesso rápido aos mesmos. Esta questão de persistência traz alguns questionamentos relevantes, a saber (i) se podemos considerar os bancos de dados em modelos relacionais (tradicionais) ou se é necessário considerar algum outro modelo, como é o caso de sistemas NoSQL em grafos e (ii) se o armazenamento oferecido pelos SGBDs relacionais atende aos requisitos de dados e operações ou teríamos de considerar outros tipos nativos dos sistemas de bancos de dados não convencionais [Wycislik and Warchal 2014].

Um dos objetivos deste trabalho é estudar como construir um *Servidor de Bancos de Dados de Redes Sociais* que permita lidar bem com grandes volumes, integrando informações de mais de uma rede social garantindo a semântica original, permitindo dar acesso a usuários não especialistas de forma que possam fazer consultas e análises diversas sobre os dados gerados nas redes.

2. Trabalhos Relacionados

Ao realizar buscas na literatura especializada, tanto em revistas como em conferências, notamos que não há a preocupação em relatar como os dados foram modelados, armazenados e disponibilizados as análises posteriores. Geralmente a definição dos *datasets* é até bem detalhada, contemplando a forma como os dados foram coletados e limpos, porém, nada ou pouco se diz quanto à maneira como os dados são utilizados na prática. No caso de *datasets* relativamente pequenos, de fato não há necessidade de muitas explicações. Entretanto, em muitos casos práticos temos que lidar com planilhas com algumas dezenas ou centenas de milhões de linhas. Para análises complexas sobre redes sociais, contextualizadas no tempo e no espaço, é necessário definir um modelo de dados adequado em função das características dos dados e as consultas que serão depois executadas.

Em [Paul et al. 2019], os autores estudam 231.246 perfis de usuários do Twitter verificados e as 79.213.811 conexões sociais entre eles, porém, não relatam onde os dados foram armazenados ou manipulados. Em [Alashri and Alalola 2020], o *dataset* é composto por quase 4.825 postagens no Facebook e tweets no Twitter e 6.074.537 de comentários sobre essas postagens mas, novamente, não há relato sobre a forma de guardar e gerenciar os dados. Em [Barbosa et al. 2022], os autores até mencionam a utilização de um sistema de banco de dados - sem citar qual! - para armazenamento de 1.779.024 tweets/retweets. Mencionam a construção de uma rede de de retweets com 502.690 vértices

e 1.067.358 arestas, sem explicar, também, onde e como tal rede foi armazenada. Em [Costa et al. 2022], os autores modelam dados disponibilizados em um SGBD (novamente não citado) em um modelo de grafo com mais de 144 mil nós e mais de 110 mil arestas, novamente sem indicação de como estes dados são guardados e processados. Em [Santos and Goya 2022] foram coletados por volta de 14,7 milhões de tweets e 662.793 usuários para detecção de posicionamento, e em [Paes et al. 2022], os autores coletam aproximadamente 252 mil tweets. Nos casos não também se detalha nem o modelo, nem o sistema de banco de dados utilizado, se é que algum foi.

3. Proposta de Pesquisa

A partir dos exemplos de análises realizadas em dados de redes sociais na literatura, temos estudado e avaliado os modelos e sistemas mais adequados para gerenciar e analisar dados de redes sociais. Nossa pesquisa agrupou alguns dos tipos de análises feitas sobre dados de redes sociais e utilizamos algumas classes para entender questões relacionadas ao melhor desempenho, menor complexidade de representação e categorias de análises.

Muitos autores partem do pressuposto que um sistema baseado no modelo de dados relacional não seria eficaz nem eficiente para atender aos requisitos das análises das informações de redes sociais. Em [Wycislik and Warchal 2014] é feita uma comparação entre o SGBD Oracle 11g (modelo relacional) e o Neo4J (modelo de grafo) executando o algoritmo *Local Clustering Coefficient* (LCC). Como em outras aplicações recentes, o artigo comenta sobre os modelos do tipo NoSQL que se tornaram cada vez mais populares por preencherem lacunas importantes deixadas por sistemas relacionais. No caso particular de análises de redes sociais um sistema de banco de dados baseados em grafos parece ser uma escolha natural por conta da representação das conexões existentes em redes sociais mais diretamente. No entanto, a tecnologia dos sistemas relacionais ainda atendem alguns tipos de análises adequadamente e contemplam soluções de algoritmos avançados para indexação, otimização de consultas e processamento de transações.

A terminologia e grupos propostos em [Tang et al. 2014] sugerem um ponto de partida interessante para as discussões da nossa pesquisa. Os autores atestam que existem 3 tipos de objetos em dados de redes sociais: (1) usuário, (2) relacionamentos e (3) conteúdos gerados por usuários. Consequentemente, as possíveis análises a serem realizadas em dados de redes socias também devem ser divididas em 3 grupos.

As **análises relacionadas ao usuário** têm o intuito de entender os usuários das redes sociais. Alguns exemplos de análise são: *community detection*, *bots detection* e *influenciadores e centralidade*. *Community detection* consiste em encontrar grupos coesos (*clusters*) em estruturas de rede complexas, permitindo descobrir grupos de usuários que têm interesses comuns. Na rede social Twitter, por exemplo, os interesses em comum de dois usuários u_i e u_j podem ser descobertos através de: (i) tópicos comuns encontrados no corpo do texto dos tweets; (ii) URLs em comum; (iii) hashtags em comum; (iv) número de vezes que u_i e u_j retweetaram a mesma pessoa; (v) número de vezes que u_i e u_j retweetaram um ao outro, entre outros [Silva et al. 2017]. Já na rede social Facebook, usuários podem ser atribuídos a determinado grupo por (a) seguirem uma determinada página ou (b) por compartilharem uma determinada postagem. O próprio conceito de grupos que o Facebook possui já é um exemplo natural de comunidade.

A detecção de *bots* identifica contas automatizadas que podem representar

ameaças potenciais à interpretação da opinião pública, democracia, saúde pública, mercado de ações e outras disciplinas. Os robôs tentam emular e produzir contéudo automaticamente como se fossem humanos [Orabi et al. 2020]. Em [Zhang et al. 2016], os autores identificam usuários maliciosos através da métrica de que usuários comuns podem até seguir, no Twitter, usuários ”estranhos”(que eles não conhecem) sem cautela, mas tendem a ser mais cuidadosos e seletivos ao retweetar, responder e mencionar outros usuários, o que não acontece com usuários maliciosos. No artigo [Santia et al. 2019] referente ao Facebook os autores definiram algumas métricas para detectar usuários maliciosos e uma delas tem relação com comentários em posts. Eles descobriram que o número médio de links postados por comentário, por exemplo, por humanos foi de 0,0181, enquanto o mesmo valor para os bots foi de 0,1467. Eles tomaram, então, a postagem frequente de links como um indicador para *bots*.

No que diz respeito aos influenciadores e centralidade, utilizaremos a definição de [Himelboim 2017]. Usuários que são mais importantes em uma rede social são considerados ”centrais”. A centralidade refere-se a quão proeminentemente conectado um usuário está em uma rede. No Facebook, onde as redes são formadas por usuários e amizades, o grau de centralidade de um usuário mede o número de amigos que o usuário tem. Em redes direcionadas, onde cada link tem uma direção, surgem dois tipos de centralidade: *in-degree* e *out-degree*. A centralidade *in-degree* é baseada em links que outros iniciaram com um usuário, e a centralidade *out-degree* é baseada nos links iniciados por um usuário com outros. Muitas redes sociais são direcionadas. No Twitter, por exemplo, os links têm uma direção, pois um usuário pode seguir, mencionar ou responder a outro. O *in-degree* seria o número de menções, respostas ou seguidores que um usuário recebe. Tweets postados por um usuário influenciador, ou seja, de alto nível de *in-degree* desfrutam de um público maior de seguidores ou são difundidos por meio de retuítes, por exemplo. O *out-degree* é o número de outros usuários que um usuário segue, menciona ou responde. Podemos citar o algoritmo *PageRank* como uma das formas de identificar os influenciadores em uma rede. Diversos trabalhos na literatura se utilizam deste método com este intuito: [Hagen et al. 2022], [Elbaghazaoui et al. 2022], [Bodrunova et al. 2017], entre outros.

As **análises baseadas nos relacionamentos** focam na mineração das interações entre os usuários e visam revelar uma visão detalhada e abrangente destes relacionamentos sociais. Alguns exemplos de análise são: *link types* e *densidade*.

Link types são o que basicamente definem com qual rede estamos lidando. Quando olhamos a nível de relacionamento, no contexto de redes sociais, podemos ter diversos tipos de links. No Twitter, seguir, mencionar, retuitar, responder, dar like, criar listas e hashtags são alguns exemplos de links; cada link criando uma rede diferente. No Facebook, amizade, curtir uma postagem, comentar uma postagem, postar em um mural e marcar um usuário em uma foto constituem tipos de links. No Instagram, seguir outro usuário é um tipo de link, enquanto curtir uma foto é outro. Cada tipo de link citado constitui redes diferentes que possuem resultados de centralidade, comunidades, informações diferentes. Em [Ituassu et al. 2018], o professor Sergio e outros autores fizeram a chamada ”análise de mídia”, a fim de identificar qual a mídia compartilhada em cada uma das publicações do Twitter. Nesse caso, o link entre usuários pode ser visto como quando dois usuários compartilham uma mesma mídia (G1, por exemplo).

De acordo com [Himelboim 2017], com relação a densidade, as redes variam em

termos de sua interconexão. Algumas são mais fortemente interconectadas, enquanto em outras os nós são conectados apenas esparsamente. A densidade da rede é medida pelo número de conexões possíveis ou potenciais, sobre o número de conexões reais. A extensão em que uma rede está densamente interconectada afeta a taxa de fluxo de informações dentro dela. [Carley 1991] mostra que a interação entre os indivíduos leva ao conhecimento compartilhado, e o conhecimento compartilhado leva a ainda mais interação.

Por fim, as **análises relacionadas aos conteúdos** gerados trazem a ideia de "Análise do Discurso"ou de narrativas [BARDlN 1977]. A ideia é construir um conhecimento analisando o discurso, a disposição e os termos utilizados por usuários. Alguns exemplos de análise neste contexto são as análises de sentimentos e análises de tópicos.

Análise de sentimentos se caracteriza por detectar a polarização de sentimentos em um texto como sendo positivo ou negativo, as vezes até neutro. No caso de dados do Twitter, por exemplo, consiste em atribuir uma polaridade a cada tweet, que pode ser positiva, neutra ou negativa em relação ao tópico principal sendo discutido. No caso do Facebook, o mesmo se aplica em polarizar as postagens (posts, comentários, etc) dos usuários com relação a sentimentos.

Já a modelagem de tópicos permite organizar, entender e resumir grandes coleções de informações textuais. Com esta técnica é possível descobrir tópicos latentes presentes em uma coleção de documentos (corpus), anotar cada documento com os tópicos para filtrar e agrupar os mesmos. Assim como na análise de sentimentos, tanto para o Twitter quanto para o Facebook, o topic modeling pode ser aplicados nas postagens dos usuários para identificar os tais tópicos latentes. Os *Trending Topics* do Twitter são, nada mais, nada menos, que a agregação de um conjunto de tweets que tratam sobre um mesmo tópico, porém com a característica de estarem acontecendo naquele momento específico.

4. Resultados Esperados

Como objeto de pesquisa que vai compor este trabalho, para todos estes tipos de análise citados, pretendemos realizar uma modelagem tal que as análises pretendidas sejam feitas com melhor eficiência e menos complexidade de expressão e representação.

Pretendemos tratar, também, do problema de semântica ao integrar dados de redes sociais diferentes, para garantir que não estamos comparando "laranjas com bananas". Apenas para citar um exemplo, na rede social Twitter temos o conceito de *like* e no Facebook as reações de usuário. É necessário tomar cuidado para considerá-las conjuntamente de forma a não mudar a intenção original da manifestação.

Com relação a coleta dos dados para a criação do *servidor* de banco de dados de redes sociais, nos utilizaremos das APIs (quando existirem) disponibilizadas pelas redes. O Twitter, hoje em dia, disponibiliza seus dados através de sua API v2. Entretanto, notamos que existem informações acerca dos dados nas redes sociais que não são disponíveis através de suas APIs. Como exemplo de uma lacuna entre o que as redes oferecem para análise e os dados em si, consideremos o caso da *repostagem direta* e da *repostagem indireta*. Uma repostagem direta é quando um usuário reposta (retuíta) diretamente um tweet que aparece em sua linha do tempo. Uma repostagem indireta ocorre quando um usuário reposta uma repostagem; o usuário reposta um tweet porque outra pessoa o repostou e é

por isso que esse tweet está aparecendo em sua linha do tempo. Na resposta da API do Twitter, as repostagem indiretas contam como repostagem diretas. Se tivermos usuário 1, usuário 2 e usuário 3: o usuário 1 postou um tweet, o usuário 2 repostou o tweet do usuário 1 e o usuário 3 repostou a repostagem feita pelo usuário 2 (Figura 1), a resposta vinda da API mostra que o usuário 3 repostou diretamente o usuário 1 (Tabela 1), o que pela interface do Twitter mostra que internamente eles sabem que o usuário 3 repostou o usuário 2. Esta lacuna impede diretamente a criação de um caminho de repostagens, por exemplo. Não somos mais capazes de montar um caminho indicando por onde um tweet passou, ou seja, quem repostou alguém que repostou alguém.

Figura 1. Notificação de repostagem em repostagem na interface do Twitter

Tabela 1. Repostagens extraídas da API do Twitter

Users	1	2	3
text	test	RT @User1: test	RT @User2: test
tweet_id	**1541580**	154824577	15318912
referenced_tweets_type	null	retweeted	retweeted
referenced_tweets_id	null	**1541580**	**1541580**

Assim, alguns resultados possíveis deste trabalho são:

1. agrupar os tipos de análises feitas em dados de redes sociais em modelagens específicas que possuam maior desempenho e menor complexidade de representação;
2. avaliar a ideia de um sistema de banco de dados integrando múltiplas redes sociais;
3. ajudar pesquisadores que não têm conhecimento na área a interpretarem dados de redes sociais;

A extensão de um sistema relacional com funções armazenadas para análises de dados de redes sociais, tal como o Neo4J faz para buscas e manipulações em redes, é um dos trabalhos que está em andamento.

Referências

Alashri, S. and Alalola, T. (2020). Functional analysis of the 2020 U.S. elections on twitter and facebook using machine learning. *Procs Intl Conf ASONAM*, pages 586–589.

Angles, R., Prat-Pérez, A., Dominguez-Sal, D., and Larriba-Pey, J.-L. (2013). *Benchmarking database systems for social network applications*. ACM.

Barbosa, C., Félix, L., Alves, A., Xavier, C., and Vieira, V. (2022). Uso de URLs para caracterização de comunidades em redes sociais online. In *BRASNAM*, pages 25–36.

BARDlN, L. (1977). Análise de conteúdo. *Lisboa: edições*, 70:225.

Bodrunova, S. S., Litvinenko, A. A., and Blekanov, I. S. (2017). Comparing influencers: Activity vs. connectivity measures in defining key actors in twitter ad hoc discussions on migrants in germany and russia. In *Intl Conf on Social Informatics*, pages 360–376.

Carley, K. (1991). A theory of group stability. *American Soc. Review*, 56(3):331–354.

Costa, L., Reis, A., Bacha, C., Oliveira, G., Silva, M., Teixeira, M., Brandão, M., Lacerda, A., and Pappa, G. (2022). Alertas de fraude em licitações: Uma abordagem baseada em redes sociais. In *BRASNAM*, pages 37–48.

Elbaghazaoui, B. E., Amnai, M., and Fakhri, Y. (2022). Data profiling and machine learning to identify influencers from social media platforms. *ICT Stds*, pages 201–218.

Hagen, L., Fox, A., O'Leary, H., Dyson, D., Walker, K., Lengacher, C. A., and Hernandez, R. (2022). The role of influential actors in fostering the polarized covid-19 vaccine discourse on twitter: Mixed methods of machine learning and inductive coding. *JMIR Infodemiology*, 2(1):e34231.

Himelboim, I. (2017). *Social Network Analysis (Social Media)*. Wiley.

Ituassu, A., Lifschitz, S., Capone, L., Vaz, M. B., and Mannheimer, V. (2018). Publicación de medios y preferencia electoral en twitter: análisis de opinión pública durante las elecciones del año 2014 en brasil. *Palabra Clave*, 21(3):860–884.

Orabi, M., Mouheb, D., Al Aghbari, Z., and Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing & Management*, 57(4):102250.

Paes, V., Araújo, D., Brito, K., and Andrade, E. (2022). Análise de sentimento em tweets relacionados ao desmatamento da floresta amazônica. In *BRASNAM*, pages 61–72.

Paul, I., Khattar, A., Kumaraguru, P., Gupta, M., and Chopra, S. (2019). Elites tweet? characterizing the twitter verified user network. In *2019 IEEE 35th International Conference on Data Engineering Workshops (ICDEW)*, pages 278–285.

Santia, G. C., Mujib, M. I., and Williams, J. R. (2019). Detecting social bots on facebook in an information veracity context. *AAAI Conf Web and Social Media*, 13(01):463–472.

Santos, P. and Goya, D. (2022). Detecção de posicionamento e rotulação automática de usuários do twitter: estudo sobre o embate científico-político no contexto da CPI da covid-19. In *BRASNAM*, pages 49–60.

Silva, W., Ádamo Santana, Lobato, F., and Pinheiro, M. (2017). *A Methodology for Community Detection in Twitter*. Association for Computing Machinery.

Tang, J., Chang, Y., and Liu, H. (2014). Mining social media with social theories. *ACM SIGKDD Explorations Newsletter*, 15:20–29.

Wycislik, L. and Warchal, L. (2014). *A Performance Comparison of Several Common Computation Tasks Used in Social Network Analysis Performed on Graph and Relational Databases*, volume 242. Springer Verlag.

Zhang, J., Zhang, R., Sun, J., Zhang, Y., Zhang, C., Zhang, J., Zhang, R., Sun, J., Zhang, Y., and Zhang, C. (2016). Truetop: A sybil-resilient system for user influence measurement on twitter. *IEEE/ACM Trans. Netw.*, 24(5):2834–2846.