Avaliação de Desempenho da Streaming do Twitter

Veruska Ayora
Universidade Federal do ABC, UFABC - Santo André, Brasil
Veruska.ayora@ufabc.edu.br

Resumo: O Twitter, um popular serviço de microblog, permite aos seus usuários enviar mensagens curtas de até 140 caracteres chamadas *tweets*. Por meio de sua API Streaming compartilha uma amostra de no máximo 1% de todos os *tweets* publicados a partir de alguns parâmetros predefinidos pelo usuário da API. Essas amostras de dados são usadas frequentemente pela comunidade cientifica em busca de detecção de eventos da vida real, como por exemplo, desastres naturais. No entanto, devido às restrições na política de serviço do Twitter surge a necessidade de entender o quanto essas amostras coletadas são representativas em relação ao cenário real. Para isso, foi proposta uma solução para avaliar simultaneamente a publicação de novos *tweets* e a captura dos mesmos via API Streaming. Finalmente, foi avaliado a representatividade desses dados, bem como o tempo de atraso dos *tweets* recuperados. Os resultados revelam, com 95% de confiança, que a taxa de captura dos experimentos realizados com 12.000 *tweets* está entre 99,47% e 99,93%. Enquanto que o intervalo de tempo entre a publicação e a captura de cada *tweet* pode variar entre 1,729 e 2,159 segundos.

1. Introdução

O Twitter, um popular serviço de microblog, permite aos usuários enviar mensagens curtas de 140 caracteres chamadas *tweets*. É uma das principais redes sociais do mundo, possui atualmente 328 milhões[1] de usuários ativos, os quais publicam em média 500 milhões[2] de mensagens por dia, por meio de seus navegadores e dispositivos móveis. Devido à grande quantidade de mensagens trocadas e a facilidade de publicação das mensagens pelo celular, o Twitter tem chamado atenção de diversos pesquisadores que visam correlacionar seu conteúdo com os acontecimentos da vida real [1]. Earle et al e Sakaki et al estudaram o comportamento dos *tweets* em eventos de terremotos [2], [3]. Enquanto que Achrekar et al e Aramaki et al estudaram a detecção de epidemias do vírus Influenza [4], [5]. Outros eventos, como crises, incidentes [6], [7] e protestos [8] também foram encontrados na literatura. Essa rede social disponibiliza a API *Streaming* (*Application Programming Interface*) para a coleta de dados em tempo real, de acordo com os parâmetros de pesquisa de interesse do usuário. Esse serviço retorna uma amostra de até 1% de todos os *tweets*. No entanto, dependendo da magnitude do evento que se almeja detectar, a amostra recuperada pode não ser significativa. A solução proposta por esse artigo é produzir um conjunto de *tweets* identificados por uma palavra-chave, e utilizar a API Streaming para recuperar esses mesmos *tweets*, de modo a avaliar a integridade e representatividade da amostra coletada.

2. Análise dos Dados

A Teoria dos Conjuntos[3] ajuda a compreender as análises realizadas nesse experimento. Os cenários avaliados incluíram as seguintes possibilidade: Fig.1 (A), temos $P \cap C = \{x \in U \mid x \in P \text{ e } x \in C\}$, ou seja, quantos foram capturados do conjunto de *tweets* publicados, e quantos são aqueles que pertencem apenas ao conjunto U, ou seja, aquele publicado por qualquer outro usuário do Twitter. Em Fig.1 (B), temos $P \cap C = \{\}$, quais as amostras não conseguiram recuperar nenhum tweet publicado pelo experimento. Em Fig.1 (C), temos: $C \subset P \Leftrightarrow C \subset P$ e $C <> P$, ou seja, são amostras que retornaram apenas *tweets* publicados pelo experimento, mas que, no entanto, em menor quantidade. Em Fig.1 (D) temos: $P \subset C \Leftrightarrow P \subset C$ e $P <> C$, situação que retorna, além de todos os *tweets* que foram publicados pelo experimento, também *tweets* que pertencem somente ao conjunto U. E por fim, em Fig.1 (E) temos: $P = C \Leftrightarrow P \subseteq C$ e $C \subseteq P$, que são as amostras que retornam exatamente os *tweets* que foram publicados pelo experimento.

[1] https://about.twitter.com/pt/company
[2] https://www.statista.com/topics/737/twitter/
[3] A Teoria dos Conjuntos foi criada e desenvolvida pelo Matemático russo George Cantor (1845-1918), trata-se do estudo das propriedades dos conjuntos, relações entre conjuntos e relações entre os elementos e o próprio conjunto.

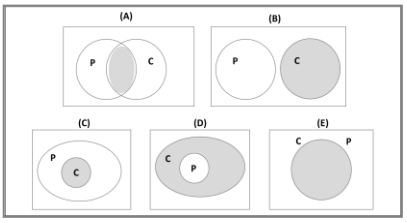

Figura 1- Diagrama de Venn (Fonte: Autor)

Tabela: Conjuntos de Tweets	
Conjunto	***Descrição***
U	O conjunto universo de todos os *tweets* que estão sendo publicados num dado momento;
P	O conjunto dos *tweets* publicados com hashtag específica;
C	O conjunto de amostras de P obtido através da conexão via API Streaming
T	O conjunto de amostras de C ∉ P, obtidas através da conexão via API Streaming

Tabela 1: Conjunto de Tweets

3. Resultados e Conclusões

Os resultados mostram que quando o tamanho dos *tweets* foram mantidos em 70 caracteres, tivemos um retorno médio de 99,63%, enquanto que quando os tamanhos foram mantidos em 130 caracteres, o retorno médio foi de 99,77%. Ao avaliar as características desses *tweets* retornados, foi observado que sete experimentos retornaram *tweets* de acordo com as características do cenário (C) da Fig.1 em que $C \subset P \Leftrightarrow C \subset P$ e $C <> P$. Apenas o experimento com publicação de 1200 *tweets* de 130 caracteres se encaixou no cenário (E) da Fig.3 em que $P = C \Leftrightarrow P \subseteq C$ e $C \subseteq P$, retornando 100% dos *tweets* publicados. Não houve situações em que o conjunto retornado foi vazio (B), e nem situações em que o conjunto foi maior que a quantidade publicada (D). No entanto, se considerarmos os *tweets* em branco, que são aqueles que a streaming envia para manter a conexão aberta, ao invés do cenário (C), teríamos a preponderância do cenário (A) para os sete experimentos que tiveram perda de *tweets*.

Os resultados mostram que mesmo aumentando a quantidade de *tweets* e seus respectivos tamanhos (Fig.2), a *API Streaming* continua retornando um porcentual acima de 98,92%. Dessa forma, não se pode concluir que existe perda para àqueles conjuntos maiores de *tweets* passiveis de serem capturados. O mesmo raciocínio aplica-se para o tamanho dos *tweets*. A Fig.3 mostra que os *tweets* podem sofrer atrasos que fariam entre 1,4 a 2,5 segundos. Para os *tweets* de 130 caracteres houve aumento nos atrasos, na medida em que aumentava o conjunto C. Essa relação não pôde ser observada nos *tweets* de 70 caracteres, o qual obteve seu maior pico de atraso no experimento com 2400 *tweets*. Em relação a taxa de perda dos *tweets*, é possível verificar (Fig. 4), como foi o comportamento da perda de *tweets*. As perdas não foram esparsas durante todo o período de 24 horas, elas aconteceram em determinado hora do dia. Observa-se que houve maior concentração de perdas no meio do dia entre as 10:00h e 14:00h. Além disso, as perdas aconteceram em 100% dos casos ou no início da coleta da amostra ou no fim da coleta.

Por fim, foi possível concluir que existe a possibilidade da API Streaming retornar 100% dos *tweets* alvo. Nessa pesquisa, isso aconteceu em 12% dos casos. A pesquisa revelou com um grau de 95% de confiança que a taxa de retorno dos *tweets* está entre 99,47% e 99,93%. Enquanto que o intervalo de tempo entre a publicação e a captura de cada tweet pode variar entre 1,729 e 2,159 segundos. Esses valores revelam que existe a possibilidade de obter uma representatividade alta de um conjunto de 12.000 *tweets* capturados pela API Streaming. A baixa latência do tempo de atraso, revelam que é concebível detectar os *tweets* alvo quase que em tempo real.

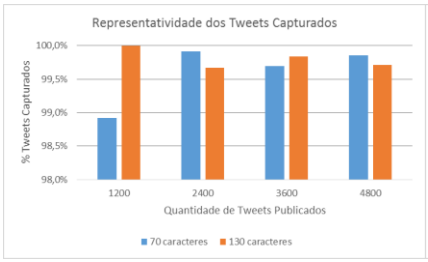

Figura 2 – Gráfico de Representatividade dos Tweets Capturados

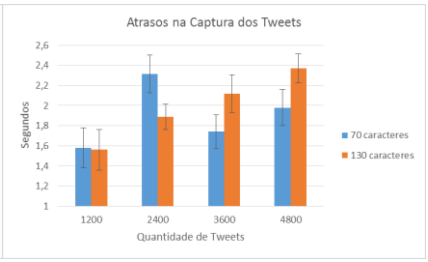

Figura 3 – Atrasos na Captura dos Tweets

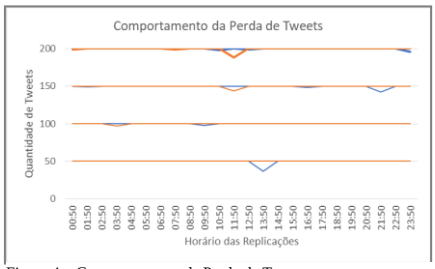

Figura 4 – Comportamento da Perda de Tweets

4. Referências

[1] Ravindran, Sharan Kumar, and Vikram Garg. Mastering social media mining with R. Packt Publishing Ltd, 2015.

[2] Earle, Paul S., Daniel C. Bowden, and Michelle Guy. "Twitter earthquake detection: earthquake monitoring in a social world." Annals of Geophysics 54.6 (2012).

[3] Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Tweet analysis for real-time event detection and earthquake reporting system development." IEEE Transactions on Knowledge and Data Engineering 25.4 (2013): 919-931.

[4] Achrekar, Harshavardhan, et al. "Predicting flu trends using twitter data." Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. IEEE, 2011.

[5] Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. "Twitter catches the flu: detecting influenza epidemics using Twitter." Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, 2011.

[6] Kryvasheyeu, Yury, et al. "Rapid assessment of disaster damage using social media activity." Science advances 2.3 (2016): e1500779.

[7] Singh, Jyoti Prakash, et al. "Event classification and location prediction from tweets during disasters." Annals of Operations Research (2017): 1-21.

[8] Starbird, Kate, Grace Muzny, and Leysia Palen. "Learning from the crowd: collaborative filtering techniques for identifying on-the-ground Twitterers during mass disruptions." Proceedings of 9th International Conference on Information Systems for Crisis Response and Management, ISCRAM. 2012.

[9] Morstatter, Fred, et al. "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose." ICWSM. 2013.

[10] Llewellyn, Clare, and Laura Cram. "Distinguishing the Wood from the Trees: Contrasting Collection Methods to Understand Bias in a Longitudinal Brexit Twitter Dataset." ICWSM. 2017.

[11] Ghosh, Saptarshi, et al. "On sampling the wisdom of crowds: Random vs. expert sampling of the twitter stream." Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.

[12] Li, R., S. J. Wang, and C. C. K. Chang. "Automatic topic-focused monitor for twitter stream." PVLDB (2014): 1966-1977.

[13] Joseph, Kenneth, Peter M. Landwehr, and Kathleen M. Carley. "Two 1% s Don't Make a Whole: Comparing Simultaneous Samples from Twitter's

[14] Perera, Rohan DW, et al. "Twitter analytics: Architecture, tools and analysis." MILITARY COMMUNICATIONS CONFERENCE, 2010-MILCOM 2010. IEEE, 2010.