Ricardo César Gonçalves Sant'Ana
Moisés Lima Dutra
Guilherme Ataíde Dias
Organizadores

ANAIS
WIDaT 2018

Organização do WIDaT 2018

- **Organização Geral:**

  Guilherme Ataíde Dias (PPGCI-UFPB) - Coordenador geral do evento

  Moisés Lima Dutra (PPGCIN-UFSC) - Vice-coordenador

- **Coordenador da Comissão Científica:**

  Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)

- **Comissão científica**

  Adilson Luiz Pinto (PPGCIN-UFSC)

  Ana Alice Baptista (Universidade do Minho, Portugal)

  Ana Carolina Simionato (PPGCI-UFSCar)

  Angela Maria Grossi de Carvalho (PPGCI-UNESP)

  Bernardina Maria Juvenal Freire de Oliveira (PPGCI-UFPB)

  Cristian Berrío-Zapata (PPGCI-UFPA)

  Dalton Lopes Martins (FCI-UnB)

  Denysson Axel Ribeiro Mota (PPGB-UFCA)

  Douglas Dyllon Jeronimo de Macedo (PPGCIN-UFSC)

  Ed Porto Bezerra (PPGI-UFPB)

  Edgar Bisset Alvarez (PPGCIN-UFSC)

  Edna Gusmão de Goés Brennand (MPGOA-UFPB)

  Edna Gomes Pinheiro (DCI-UFPB)

  Elaine Parra Affonso (FATEC-SP)

  Elvis Fusco (UNIVEM-Marília)

  Enrique Muriel Torrado (PPGCIN-UFSC)

  Evandro de Barros Costa (IC-UFAL)

  Fábio Paraguaçu (IC-UFAL)

  Fernando de Assis Rodrigues (PPGCI-UNESP)

  Gustavo Medeiros de Araújo (PPGCIN-UFSC)

  Henry Pôncio Cruz de Oliveira (PPGCI-UFPB)

  Joana Coeli Ribeiro Garcia (PPGCI-UFPB)

  José Eduardo Santarém Segundo (USP-FFCLRP)

  Leonardo Castro Botega (UNIVEM-Marília)

  Luana Farias Sales Marques (PPGCI-IBICT-UFRJ)

  Marckson Roberto Ferreira de Sousa (PPGCI-UFPB)

  Luís Fernando Sayão (CNEN)

  Marcelo Morandini (EACH-USP)

  Márcio Matias (PPGCIN-UFSC)

  Marcos Mucheroni (CBD-USP)

  Marynice de Medeiros Matos Autran (PPGCI-UFPB)

Maurício Barcellos Almeida (PPGGOC-UFMG)
Moisés Lima Dutra (PPGCIN-UFSC)
Plácida Leopoldina V. da Costa Santos (PPGCI-UNESP)
Pedro Luiz Pizzigatti Corrêa (POLI-USP)
Renata Baracho (PPGGOC-UFMG)
Ricardo César Gonçalves Sant'Ana (PPGCI-UNESP)
Robson Rodrigues Lemos (UFSC-Araranguá)
Rogério Ramalho (PPGCI-UFSCar)
Ryan Ribeiro de Azevedo (UFRPE-UAG)
Sandra de Albuquerque Siebra (PPGCI-UFPE)
Sandro Rautenberg (DECOMP-UNICENTRO)
Silvana Aparecida Borsetti G. Vidotti (PPGCI-UNESP)
Virginia Bentes Pinto (PPGCI-UFC)
Wagner Junqueira de Araújo (PPGCI-UFPB)
Zaira Regina Zafalon (PPGCI-UFSCar)

- **Coordenador do Cerimonial:**
  André Luiz Dias de França (PPGCI-UFPB)

- **Coordenador da Equipe Técnica Local:**
  Laerte Pereira da Silva Júnior (CCHLA-UFPB)

- **Equipe Técnica Local:**
  Adriana Alves Rodrigues (PPGCI-UFPB)
  Antonio Felipe dos Santos (MPGOA-UFPB)
  Débora Gomes de Araújo (PPGCI-UFPB)
  Pedro Augusto de Lima Barroso (PPGCI-UFPB)
  Pollianna Marys de Souza e Silva (PPGCI-UFPB)
  Renata Lemos dos Anjos (PPGCI-UFPB)

METADADOS E TWITTER:
uso do Tweet Object para identificação de local

METADATA AND TWITTER:
use of Tweet Object for location identification

Denysson Axel Ribeiro Mota[1], Gracy Kelli Martins[(2)]
(1) Universidade Federal do Cariri (UFCA), Av. Tenente Raimundo Rocha S/N - Bairro Cidade Universitária - Juazeiro do Norte - CE – Brasil, CEP 63048-080, denysson.mota@ufca.edu.br
(2) Universidade Federal da Paraíba (UFPB), Cidade Universitária - João Pessoa - PB – Brasil, CEP: 58051-900, gracykmg@gmail.com

Resumo:
Esta pesquisa apresenta um estudo, com base nos Estudos Métricos da Ciência da Informação, sobre a disseminação de informações na plataforma Twitter. Este trabalho teve como objetivos identificar elementos para análises métricas nos Tweets, utilizando como recorte mensagens coletadas na ferramenta via API. Para alcançar estes objetivos, utilizou dados obtidos via API do Twitter, usando sistema operacional Debian Linux e uma aplicação Ruby, armazenados em formato JSON em uma base de dados no SGBD PostGreSQL, e manipulados no SublimeText 3 e OpenOffice quando necessário. Apresenta como resultados a distribuição geográfica em cada uma das etiquetas analisadas, mostrando que mesmo que existem diferenças em valores totais, a distribuição em geral se mantém parecida. Conclui-se que as etiquetas disponíveis no JSON são extremamente úteis, mas que é necessário mais estudo sobre as possíveis formas de identificação de metadados úteis para os processos métricos de informação.
Palavras-chave: Disseminação de dados. Geolocalização. Tweet Object. Twitter.

Abstract:
This research proposes to study, based on the Metric Studies of Information Science, the dissemination of information on the Twitter platform. This work aimed to identify elements for metric analysis in Tweets, using as clipping messages collected in the tool via API. To achieve these goals, it used data obtained via the Twitter API, using Debian Linux operating system and a Ruby application, stored in JSON format in a database in PostGreSQL DBMS, and manipulated in SublimeText 3 and OpenOffice when required. It presents as results the geographical distribution in each of the analyzed tags, showing that even if there are differences in total values, the distribution in general remains similar. It is concluded that the tags available in JSON are extremely useful, but that more study is needed on possible ways of identifying metadata useful for metric information processes.
Keywords: Data dissemination. Geolocation. Tweet Object. Twitter.

1 INTRODUÇÃO

Em momentos de grande comoção, como desastres naturais, surtos epidêmicos e eleições, por exemplo, a sociedade busca por informações relevantes para guiar suas decisões no dia a dia, sejam em rádios, jornais e programas televisivos, e em especial pela internet, em portais de notícias, serviços de vídeo, blogs e redes sociais.

As redes sociais em ambiência digital são ferramentas de comunicação que surgiram no final do século XX, iniciando em 2003 com Myspace, que buscava integrar artistas com fãs, e LinkedIn, criada especificamente como uma rede social de negócios. Posteriormente, em 2004, surgiram duas redes: Orkut, baseado na teoria do mundo pequeno, também conhecida como a teoria dos seis graus de separação, de Stanley Milgram (1967), e *Facebook*, atualmente a rede social mais utilizada, com cerca de 2.2 bilhões de usuários mensais (STATISTA, 2018a). Em 2006 surge o *Twitter*, que dentre estes últimos, é um serviço que se caracterizou pela rápida velocidade de disseminação de informações.

O *Twitter*, desde sua criação em 2006, tem passado por diversas mudanças ao longo de sua história. No entanto algo que se mantém constante é o seu poder como ferramenta de comunicação, principalmente devido à sua estrutura de organização em formato de seguidores, em vez de contatos e/ou amigos como as demais. Criado como uma ferramenta para simular um serviço de SMS na Web, passando para a estrutura de *microblog*, e evoluindo para ferramenta de rede social, o *Twitter* conta atualmente com 330 milhões de usuários ativos mensais em todo o mundo (TWITTER, 2018; STATISTA, 2018b), e por ela são enviadas cerca de 500 milhões de novas mensagens por dia (OMNICORE, 2018), entre estas mensagens informativas enviadas por diversos perfis governamentais oficiais.

Todo este fluxo de mensagens não é composto apenas pelos textos enviados. Junto com a mensagem, o *Twitter* armazena alguns metadados que permitem identificar além das mensagens e outras características, por meio de uma estrutura que permite sua consulta posterior, chamada de *Tweet Object* (TWITTER, 2017a). O *Tweet Object*, que será chamado daqui em diante de TO, contém em sua estrutura dados como a identificação do usuário que realizou o *tweet*, dia e hora do *tweet*, se é uma mensagem original ou uma replicação (chamado de *retweet*), se outros usuários foram citados, entre outros.

Esse tráfego de dados pode ser acessado via APIdo *Twitter*, que fornece acesso à base de dados retornando a solicitação via documento texto em formato *JavaScript Object Notation* (TWITTER, 2017b). A *JavaScript Object Notation*, mas mais conhecido pelo seu acrônimo: JSON, é um padrão de notação para intercâmbio de dados estruturados na internet via texto, criado em 2001 e adotado em diversas plataformas devido a ser independente de linguagem, suas definições podem ser encontradas na RFC8259 e ECMA-404 (BRAY, 2017; ECMA, 2017).

2 OBJETIVOS

Este trabalho tem como objetivo principal identificar elementos para análises métricas nos *tweets*, utilizando como recorte mensagens coletadas na ferramenta via API. Como objetivos específicos, temos: a) identificação dos elementos no *Tweet Object*; b) manipulação dos objetos JSONB na base de dados; c) exportação dos dados para ferramentas de mapas *online*.

3 PROCEDIMENTOS METODOLÓGICOS

A captura dos *tweets* foi realizada via acesso à API usando um *script* na linguagem Ruby, que foi executado em um servidor Debian no período compreendido entre outubro de 2017 até março de 2018, e armazenados em banco de dados PostgreSQL em formato JSON.

Após os estudos da API e da estrutura do *Tweet Object*, o elemento que despertou a curiosidade no primeiro momento a possibilidade de descobrir a distribuição geográfica desses *tweets*.

Para isto inicialmente foram selecionados os *tweets* onde a tag [coordinates] não estava vazia ou nula. Esta etiqueta retorna a localização onde o *tweet* foi realizado, em coordenadas com longitude e latitude, nesta ordem. O Comando SQL, que busca todos os registros de coordenadas que não estiverem vários ou nulos, está exposto na Figura 1.

Figura 1 – Distribuição de Tweets via Etiqueta [coordinates]

```
select
  dados->'coordinates' as coord
from tweets
where
  dados->'coordinates' != '"""' and
  dados->'coordinates' != 'null';
```

Fonte: Dados da Pesquisa, 2018.

Após a extração destes dados, foi realizado mapeamento das coordenadas na ferramenta *MapMakerApp*, retornando a o mapa exibido no Apêndice A. Esta plataforma mapa permite a navegação por área, detalhando as quantidades de *tweets* por local, com zoom e navegação interativa.

No entanto apenas 2.125, do total de 2.907.720 de *tweets*, contém a etiqueta [coordinates] preenchida, pois é um campo opcional e recomendado o não preenchimento por organizações de segurança. Foi realizada então uma extração usando a subetiqueta [location] dentro da etiqueta [user]. Esta etiqueta também é opcional, mas está preenchida em 2.003.338 dos *tweets* coletados.

Utilizando o comando exposto na Figura 2, foram obtidos 2.003.338 *tweets* que têm esta etiqueta, que posteriormente foram importados na plataforma Google Maps, que permite a inserção de locais em texto, e mediante a API de Geocoding disponível marca a localização em latitude e longitude no mapa.

Figura 2– Distribuição de Tweets via Etiqueta [user]->[location]

```
select
  dados->'user'->'location' as loc
from tweets
where
  dados->'user'->'location' != '"""'
  and
  dados->'user'->'location' != 'null';
```

Fonte: Dados da Pesquisa, 2018.

Os primeiros 12 mil registros foram inseridos no site *Google Maps*, pois ele permite a inserção de locais por nome e a própria plataforma faz a conversão para coordenadas, resultando no mapa exibido aqui no Apêndice B.

No entanto, devido ao grande número de registros, e ao limite de upload da ferramenta, que apenas permite que 2.000 registros sejam lançados por vez, foi decidido seguir por um caminho de marcar apenas uma vez por local.

Considerando que o preenchimento é livre, podendo inclusive conter caracteres especiais, os registros foram então tratados e filtrados usando funcionalidades do *Sublime Text* 3 de busca usando expressões regulares, em conjunto com o pacote *Remove Non Ascii Chars* para remover todos os caracteres especiais que o Twitter permite incluir na localização, como *emoticons*, por exemplo. É importante ressaltar que isto também elimina os registros de cidades que tenham sido redigidos usando ideogramas, como de países da Ásia, mas que devido aos pesquisadores não conhecerem os caracteres não seria possível filtrar o que de fato é cidade e o que não é.

Utilizando a ferramenta de substituição de texto do *Sublime Text* 3, foram utilizadas expressões regulares para excluir, do início e do final da linha, usando os caracteres ^ e $, respectivamente, todos os caracteres não alfanuméricos e espaços em branco ou tabulação. Esta substituição foi realizada de forma repetitiva até que a busca não encontrasse mais registros que atendessem os requisitos.

Posteriormente a esta busca foram excluídas todas as linhas vazias, também utilizando a ferramenta de busca textual e informando a expressão regular ^\n e deixando o campo {Replace:} vazio. Depois disto, foram excluídas as linhas duplicadas no menu {Edit}>{Permute Lines}>{Unique}. Com estes mecanismos, foi possível reduzir o número de registros de locais para 129.875, dos 2.003.338 registros recuperados anteriormente.

Mesmo reduzindo o número total de registros, a limitação da plataforma *Google Maps* ainda é uma barreira para o adequado uso dela, sendo necessária a realização de 65 processos de importação, além das correções e validações necessárias nos casos em que a plataforma não compreenda o local.

Por fim, foram realizadas extrações usando a etiqueta [place]. Esta etiqueta indica a qual local o *tweet* está relacionado, mas não indica, necessariamente, onde ele foi realizado. Dentro desta etiqueta existem duas outras que são interessantes: [full_name] e [bounding_box]. A primeira tem o local em formato texto, passível de importação no *Google Maps*, enquanto a segunda tem uma outra etiqueta aninhada chamada de [coordinates], que contém as coordenadas para formar uma caixa de exibição do local. O comando SQL utilizado pode ser visto na Figura 3.

Figura 3– Distribuição de Tweets via Etiqueta [place]->[bounding_box]->[coordinates]

```
select
  dados->'place'->'bounding_box'->'coordinates' as lugar
from tweets
where
  dados->'place' != '""' and
  dados->'place' != 'null';
```

Fonte: Dados da Pesquisa, 2018.

Devido à limitação de importação do *Google Maps*, decidiu-se usar esta segunda opção, tratando as coordenadas no *OpenOffice* para obter o ponto central indicado. Após a extração destas

coordenadas, as mesmas foram importadas na plataforma *MapMakerApp*, resultando no mapa exibido no Apêndice C.

4 RESULTADOS

É possível afirmar, com a análise apresentada, que os três objetivos foram alcançados de forma total, mas que ainda permitem a expansão da pesquisa com novas etiquetas e análises.

As diferentes possibilidades de coleta de localização dos tweets mostram a flexibilidade da ferramenta. Mesmo com as diferenças entre os valores obtidos, como pode ser visto nos Apêndices A,B e C, os grandes aglomerados de tweet permanecem (Brasil, EUA, India e Tailândia).

Isto pode demonstrar alguma consistência dos metadados coletados automaticamente pela plataforma com os informados livremente pelos usuários, mas mais testes são necessários considerando as possibilidades de representação na etiqueta [user]->[location].

5 CONSIDERAÇÕES FINAIS

A plataforma disponibilizada pelo Twitter nos permite avaliar diversos metadados para estudos métricos da informação. No entanto, é necessária a identificação das etiquetas disponíveis, para saber onde ir para buscar a informação necessária, e quais as possibilidades de combinação entre elas para enriquecimento da análise.

Nesta perspectiva, buscou-se aqui estudar as etiquetas e geolocalização, de forma relativamente simples, mas é possível realizar posteriores pesquisas que envolvam cruzamentos destes dados com outras etiquetas, e com isso obter novas possibilidades de análise.

De forma semelhante, podem ser feitas comparações entre os metadados das etiquetas [coordinates] da raiz do tweet com os da etiqueta [place]->[bounding_box]->[coordinates] dos tweets que tenham ambas.

A possibilidade de importação destes metadados em plataformas de mapas que permitem navegação é um caminho interessante de disponibilização de dados de pesquisa para uso posterior.

REFERÊNCIAS

BRAY, T. (Ed.). The JavaScript Object Notation (JSON) Data Interchange Format. 2017. Disponível em <https://www.rfc-editor.org/info/rfc8259> Acesso em: < 25 mar. 2018 >.

ECMA INTERNATIONAL (ECMA). The JSON Data Interchange Syntax. 2017. Disponível em <https://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf > Acesso em: < 25 mar. 2018>.

MILGRAM, S. The small world problem. **Psychology Today**, v. 2, p. 60-67, 1967.

STATISTA. Most popular social networks worldwide as of April 2018, ranked by number of active users (in millions). 2018a. Disponível em < https://www.statista.com/statistics/272014/global-social--networks-ranked-by-number-of-users/ > Acesso em: < 6 abr. 2018 >.

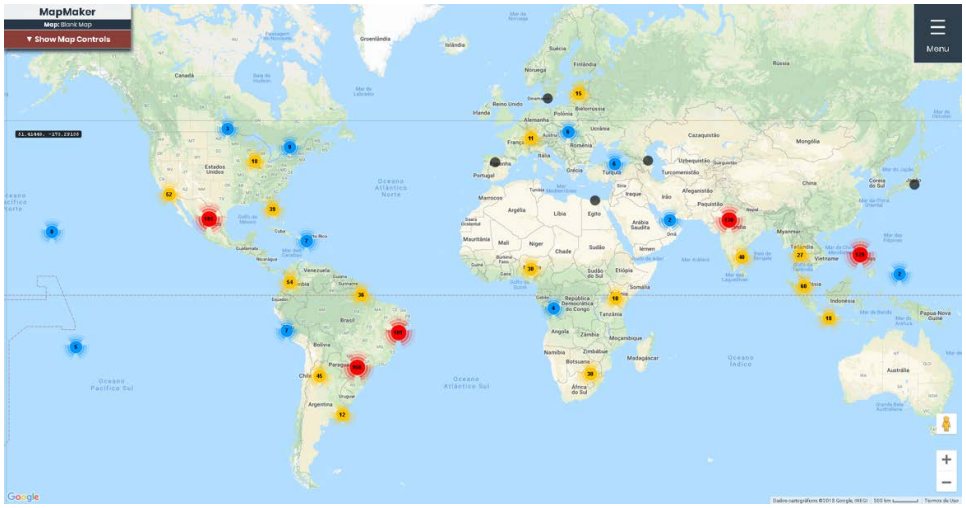
STATISTA. Twitter: number of monthly active users 2010-2017. 2018b. Disponível em <https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ > Acesso em: < 6 abr. 2018 >.

TWITTER. Q1 2018 - Letter to Shareholders. 2018. Disponível em <http://files.shareholder.com/downloads/AMDA-2F526X/6236645277x0x978181/2FD6D58F-A930-4EB2-90B0-9C3A120648DE/Q1_2018_Shareholder_Letter.pdf> Acesso em: < 03 maio 2018 >.

TWITTER. Tweet Object – Twitter Developers. 2017a. Disponível em <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object> Acesso em: < 25 mar. 2018 >.

TWITTER. Introduction to Tweet JSON – Twitter Developers. 2017b. Disponível em <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/intro-to-tweet-json> Acesso em: < 25 mar. 2018 >.

Apêndice A – Distribuição de Tweets via Etiqueta [coordinates]

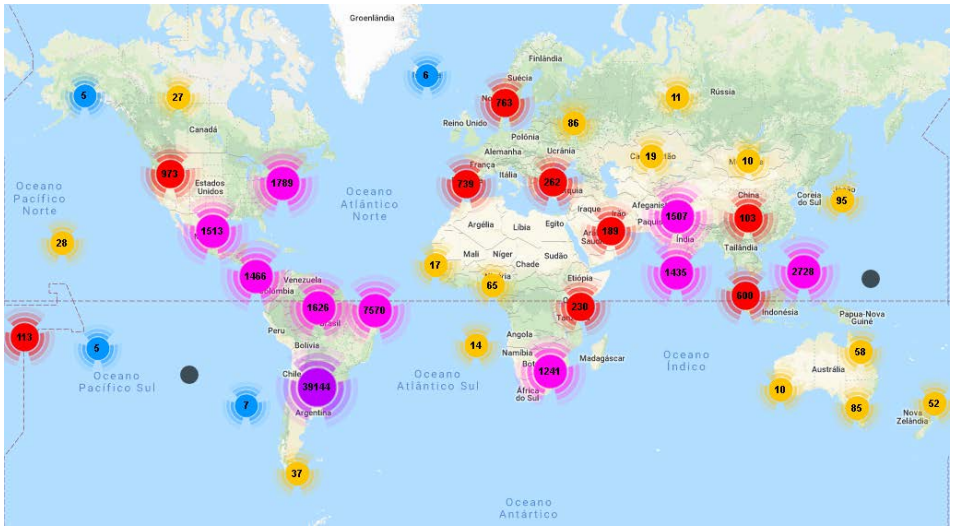


Fonte: Dados da Pesquisa, 2018.

Apêndice B – Distribuição de Tweets via Etiqueta [user]->[location]

Fonte: Dados da Pesquisa, 2018.

Apêndice C – Distribuição de Tweets via Etiqueta [place]->[bounding_box]->[coordinates]



Fonte: Dados da Pesquisa, 2018.