

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/308815966>

Data Crawling Otomatis pada Twitter

Conference Paper · September 2016

DOI: 10.21108/INDOSC.2016.111

CITATIONS

18

READS

11,911

3 authors:



Jaka Eka Sembodo

Telkom University

4 PUBLICATIONS 25 CITATIONS

SEE PROFILE



Erwin Budi Setiawan

Telkom University

40 PUBLICATIONS 124 CITATIONS

SEE PROFILE



Abdurahman Baizal

Telkom University

36 PUBLICATIONS 194 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Social Network Analysis [View project](#)

Data Crawling Otomatis pada Twitter

Jaka Eka Sembodo¹, Erwin Budi Setiawan², ZK Abdurahman Baizal³

Computational Science, School of Computing, Telkom University.

Telekomunikasi street 01, Terusan Buah Batu, Bandung, Barat, Indonesia.

¹ jecksart@gmail.com

² setiawanerwinbudi@gmail.com

³ bayzal@gmail.com

Abstract

Twitter is a social networking that first launched on July 2006 and nowadays use by many people in the world. Twitter has functions as a type of micro-blogging (blog small) with tweet (post in twitter) maximum is 140 characters. The problems is hard to get data (data crawling data) from twitter either user data or tweet data automatically. So that, in a research about using tweet especially in collecting data process be more efficeint. In this research writer use Application Programming Interface (API) twitter using the PHP programming language to build crawling data system from twitter automatically. Data crawling from twitter can be use for two search method, by user and by keyword. Search by keyword that is searching using fragment of the sentence otherwise hashtag with the tweet total in a process maximum 100 tweets. And search by user that is searching using name of user twitter with the tweet total is maximum 200 tweets. Feature extraction that get form twitter index for the user data are number of tweets, number of followers, the following amounts, total love, websites, resources, bio profile, id, account, in the name of aan location. The feature extraction for the tweet data are url, mentions, tweets, hashtag, Period Period Like Dan tweet.

Keywords: Twitter, Crawling Data, API Twitter

Abstrak

Twitter merupakan salah satu jejaring sosial yang pertama kali diluncurkan pada Juli 2006 dan saat ini banyak digunakan oleh masyarakat seluruh dunia. Twitter mempunyai fungsi sebagai media sosial bertipe *micro-blogging* (blog berukuran kecil) dengan jumlah karakter dalam *tweet* (post dalam twitter) maksimal 140 karakter. Permasalahan saat ini adalah sulit untuk mengambil data (*crawling data*) dari twitter baik berupa user maupun tweet secara otomatis. Sehingga dalam beberapa penelitian yang menggunakan data tweet menjadi kurang efisien dalam proses pengumpulan data. Pada penelitian ini penulis mengembangkan aplikasi dengan memodifikasi *Application Programming Integration* (API) twitter dengan menggunakan Bahasa pemograman PHP untuk membangun sistem crawling data di twitter secara otomatis. Crawling data di twitter dapat menggunakan dua sistem pencarian, *by user* dan *by keyword*. Pencarian menggunakan *by keyword* yaitu pencarian menggunakan penggalan kata maupunu hashtag dengan total tweet yang diunduh dalam sekali proses maksimum 100 tweet. Sedangkan pencarian dengan *by user* yaitu pencarian berdasarkan nama akun user twitter dengan total tweet yang diunduh dalam sekali proses maksimum 200 tweet. Ekstrasi fitur yang didapat dari index twitter untuk data user berupa total tweet, total follower, total following, total likes, website, source, bio profile, id, akun, nama dan lokasi. Sedangkan ekstraksi fitur yang didapat dari index twitter untuk data tweet berupa url, mention, retweet, hashtag, jumlah likes dan jumlah retweet.

Kata Kunci: Twitter, Crawling Data, API Twitt

I. PENDAHULUAN

Twitter merupakan media sosial bertipe micro-blogging (blog berukuran kecil) yang didirikan oleh Jack Dorsey pada Maret 2016 dan diluncurkan pada Juli 2006. Keunikan dari twitter adalah mempunyai tweet atau post yang ada di twitter dengan ukuran maksimum 140 karakter. Crawling data di twitter adalah suatu proses untuk mengambil atau mengunduh data dari server twitter dengan bantuan Application Programming Integration (API) twitter baik berupa data user maupun data tweet. Beberapa penelitian sebelumnya menggunakan data yang didapat dari twitter sebagai bahan acuan untuk perkembangan penelitian mereka. Misalnya penelitian (1) yang membahas tentang cara menemukan kurator berita di twitter. Penelitian ini memanfaatkan data berupa data user dan data tweet yang didapatkan dari twitter untuk menemukan user dengan peran sebagai kurator berita. Adapun penelitian (2), yaitu penelitian tentang mencari intisari dari keramaian berita yang ada di twitter. Penelitian ini juga menggunakan data tweet dari trending topic yang ada di twitter untuk menemukan topik yang menjadi perbincangan utama. Hanya saja dari beberapa penelitian sebelumnya, proses crawling data dari tweet masih sulit dilakukan secara otomatis. Sehingga menyebabkan tahap pengumpulan data dari suatu penelitian yang menggunakan data dari twitter kurang efisien dari segi waktu. Berangkat dari permasalahan ini penulis membangun aplikasi yang berfungsi untuk crawling data dari twitter secara otomatis.

II. TINJAUAN PUSTAKA

Crawling data. Crawling data merupakan tahap dalam penelitian yang bertujuan untuk mengumpulkan atau mengunduh data dari suatu database (3, 4). Pengumpulan data dari penelitian ini yaitu data yang diunduh dari server twitter berupa user dan tweet beserta atribut-atributnya.

API Twitter atau Application Programming Interface (API) twitter adalah suatu program atau aplikasi yang disediakan oleh twitter untuk mempermudah developer lain dalam mengakses informasi yang ada di website twitter (3, 4). Pendaftaran sebagai developer aplikasi twitter untuk menggunakan API twitter dapat dilakukan di lama <https://dev.twitter.com>. Setelah mendaftar developer akan mendapatkan consumer key, consumer access, access token dan access token secret yang akan digunakan sebagai syarat otentifikasi dari aplikasi yang akan kita bangun. Tujuan dari otentifikasi adalah untuk hak akses developer dalam mengunduh data yang ada di twitter.

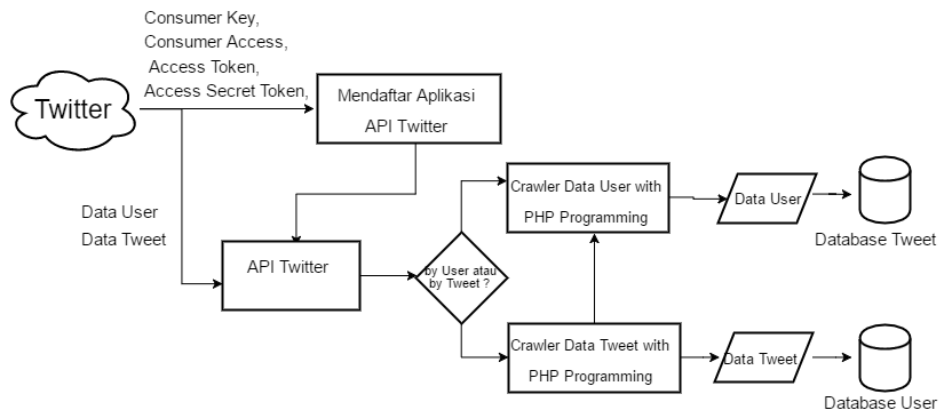
Bahasa Pemrograman PHP dan MySQL. Bahasa pemrograman PHP yang mempunyai singkatan dari Hypertext Preprocessor merupakan Bahasa pemrograman yang digunakan secara luas untuk penanganan pembuatan dan pengembangan sebuah situs web dan bisa digunakan bersamaan dengan HTML (6). MySql merupakan perangkat lunak sistem manajemen basis data SQL yang memiliki antarmuka (*interface*) terhadap berbagai aplikasi dan Bahasa pemrograman dengan menggunakan fungsi API.

Penelitian terkait. Penelitian (3) yang membahas tentang penggunaan twitter sebagai media sosial dan media berita yang didukung dengan API twitter. Penelitian ini bertujuan untuk mengumpulkan data dari twitter berdasarkan user profil, topik yang sedang trending dan tweet yang berhubungan. Penelitian lainnya yang terkait adalah penelitian (4) yang membahas tentang crawling data di twitter dengan API yang terdiri jadi dua, yaitu REST-API dan API untuk streaming. REST-API menggunakan HTTP dalam pengunduhan data yang didukung berbagai macam query di URL dengan berbagai macam informasi. Sedangkan API untuk streaming untuk mengakses data dari aktivitas tweet secara real time.

III. METODE PENELITIAN

Tujuan dari penelitian ini adalah mengimplementasikan *Application Programming Interface* (API) twitter untuk membangun sistem crawling data dari twitter secara otomatis dengan menggunakan Bahasa pemrograman PHP dan basisdata MySQL. Data yang diunduh berupa data user dan data tweet beserta fitur-fitur yang dapat diekstraksi dari indeks yang ada di twitter.

Flowchart dari sistem yang dibangun seperti yang disajikan pada **Picture 1** sebagai berikut:



PICTURE I. FLOWCHART SISTEM

API twitter berfungsi sebagai penghubung antara sistem yang dibangun dengan twitter. API twitter membutuhkan consumer key, consumer access, access token dan access secret token yang didapatkan dengan cara mendaftarkan aplikasi API twitter di <http://dev.twitter.com>. Berikutnya pencarian data dari twitter dilakukan berdasarkan dua metode pencarian, yaitu by user dan by keyword. Sistem yang dibangun menggunakan bahasa pemrograman PHP dengan pencarian data berdasarkan user dan keyword. Terakhir data diunduh dan disimpan ke dalam database tweet dan user.

IV. HASIL DAN DISKUSI

Implementasi dan konfigurasi. Implementasi crawling data di twitter secara otomatis dalam penelitian ini menggunakan Bahasa pemrograman PHP dan basisdata MySQL. Konfigurasi yang disediakan adalah dibutuhkan file cacert.pem yang merupakan library ekstensi untuk menjalankan fungsional dari Application Programming Interface (API) twitter. Modifikasi dari file php.ini dengan menambahkan kode `curl.cainfo = "C:\xampp\php\cacert.pem"` yang bertujuan untuk memanggil fungsi dari file cacert.pem. Persyaratan untuk menggunakan API twitter lainnya adalah diperlukannya consumer key, consumer access, access token dan access secret token yang didapatkan dengan cara mendaftarkan aplikasi API ke dev.twitter.com. Adapun contoh dari keempat data tersebut ditampilkan pada **Table 1**.

TABLE I
FITUR DATA USER

Nama	Nilai
Consumer Key	C5446x1hFCSsXwuozZDm3HYAM
Consumer Access/Consumer Secret	7odd1BjaMDbTqZL2GIOFGjcoNYYn2jLtmpsKY46HKvj85BkliR
Access Token/Oath Token	150547297-qCps5CzlsmcIEbxBqtZh3eJPHoYwzZSu9wN5l6zD
Access Secret Token/Oatch Secret Token	elLivqbIOlbr2XQVEg1t2g26KaOatoxyVqW5y9EvMQIFD

Metode pencarian. Metode pencariann yang disediakan oleh API twitter terdiri dari dua, yaitu pencarian berdasarkan user dan pencarian berdasarkan keyword. Pencarian berdasarkan user adalah pencarian tweet-tweet dari user yang akan dicari. Pencarian ini hanya menampilkan tweet dari satu user. URL yang digunakan untuk pencarian ini adalah `'https://api.twitter.com/1.1/statuses/user_timeline.json'` dengan parameter masukan `'?screen_name'` untuk nama dan `'?count'` untuk jumlah tweet yang akan diunduh. Sedangkan pencarian

bedasarkan keyword adalah pencarian tweet-tweet yang berkaitan dengan kata yang akan dicari. URL yang digunakan untuk pencarian ini adalah `'https://api.twitter.com/1.1/search/tweets.json'` dengan parameter masukannya `'?q'` untuk kata yang akan dicari dan `'?count'` untuk jumlah tweet yang akan diunduh.

Jumlah data. Penulis melakukan pengujian dengan mengunduh data menggunakan aplikasi crawling data yang dibangun. Data yang diunduh terdiri dari data user dan data tweet. Hasilnya adalah jumlah maksimum data tweet berdasarkan user yang dapat diunduh dalam sekali proses crawling maksimum 200 tweet. Sedangkan untuk pencarian berdasarkan keyword, jumlah maksimum tweet yang dapat diunduh adalah 100 tweet. Hal ini kemungkinan disebabkan karena ada batasan konfigurasi runtime dari php atau batasan default dari API twitter. Pada kasus ini, solusi penulis adalah dengan melakukan beberapa proses crawling sehingga jumlah data yang didapatkan baik data user maupun tweet sesuai dengan kebutuhan penelitian.

Ekstraksi fitur. Ekstraksi fitur merupakan pemecahan indeks yang disediakan oleh twitter menjadi beberapa fitur yang akan digunakan dalam penelitian. Ekstraksi fitur disini didapatkan dengan menampilkan *respon code* dari sistem yang dibangun dengan cara menghilangkan komentar pada kode `'//print_r($response);'`. *Respon code* berupa respon dari twitter kepada sistem dengan Bahasa pemrograman json. Sehingga diperlukannya penerjemahan Bahasa json menggunakan laman <http://jsonformatter.curiousconcept.com>. Adapun fitur yang bisa didapatkan untuk data user disajikan pada **Table 2**, sedangkan fitur yang bisa didapatkan untuk data tweet disajikan pada **Table 3**.

TABLE 2
FITUR DATA USER

No	Nama	Keyword Index	Tipe	Keterangan
1	ID User	['user']['id_str']	Integer	Primary dan unique dari user.
2	Nama	['user']['name']	Text	Nama dari user.
3	Akun	['user']['screen_name']	Text	Akun dari user, selalu ada karakter '@' diawal kata dan tanpa menggunakan spasi.
4	Total Tweet	['user']['statuses_count']	Integer	Jumlah tweet yang dibuat oleh user
5	Total Followers	['user']['followers_count']	Integer	Jumlah follower dari user
6	Total Following	['user']['friends_count']	Integer	Jumlah user yang diikuti
7	Total Likes	['user']['favourites_count']	Integer	Jumlah tweet yang menjadi favourite / like
8	Bio Profile	['user']['description']	Text	Biografi singkat dari user
9	Website	['user']['url']	Text	Website yang dicantumkan oleh user
10	Lokasi	['user']['location']	Text	Lokasi yang dicantumkan oleh user
11	Photo Profil	['user']['profil_image_url_https']	Image	Photo profil user

TABLE 3
FITUR DATA TWEET

No	Nama	keyword Index	Tipe	Keterangan
1	ID Tweet	['id_str']	Integer	Primary dan unique dari tweet.
2	Text	['text']	Text	Isi teks dari tweet yang dibuat oleh user.
4	URL	['entities']['urls']	Text	Entitas dari tweet yang mengandung URL
5	Retweet	['retweeted_status']	Boolean	Tanda tweet user termasuk sebagai tweet atau retweet.
6	Hashtag	['entities']['hashtags']	Text	Entitas dari tweet yang mengandung hashtag
7	Mention	['entities']['user_mentions']	Text	Entitas dari tweet yang menandung mention
8	Total Like	['retweet_count']	Integer	Jumlah user yang menyukai tweet
9	Total Retweet	['favourite_count']	Integer	Jumlah user yang membuat retweet
10	Source Tweet	['source']	Text	Sumber dari tweet yang dibuat
11	Waktu Tweet	['created_at']	Text	Waktu saat tweet dibuat

Screenshot aplikasi. Aplikasi yang dikembangkan oleh penulis bertujuan untuk crawling data dari twitter secara otomatis dengan masukan keyword, akun user, jumlah tweet dan radio button untuk pilihan simpan data ke dalam database. Screenshot dari aplikasi yang dibangun oleh penulis untuk crawling data dari twitter secara otomatis ada pada **Picture 2**.

AUTOMATIC CRAWLING DATA IN TWITTER
developed by Jaka Eka Sembodo || lectured by Mr. Erwin Budi Setiawan, S.Si, M.T


Let's Crawl the Tweet !

Search Keyword: Search akun user :

Jumlah tweet (Max 200):

☐ Simpan data hasil crawling ke dalam database

----- HASIL CRAWLING DATA TWEET -----

1	 <p>Erwin Budi Setiawan @erwinbudis ID User: 150547297 Total Tweet: 22 Jumlah Following: 278 Jumlah Follower: 128 Jumlah Likes: 2 Bio Profile: Lecturer and Researcher Lokasi: Telkom University, Bandung ID Tweet: 705980865235431424 RT @maspiyungan: Hamas Kini Miliki 12 Ribu Raket https://t.co/1MDpzNfB9C URL: yes Mention: yes Retweet: yes Tweet biasa: no</p>
---	---

PICTURE 2. FLOWCHART SISTEM

V. KESIMPULAN

Kesimpulan yang didapatkan berdasarkan penelitian ini adalah sistem crawling data dapat diimplementasikan menggunakan bahasa pemrograman PHP dengan modifikasi Application Programming Interface (API) twitter. Diperlukan file cacert.pem dan konfigurasi dalam file php.ini untuk menjalankan fungsional crawling data. Pencarian dapat dilakukan dengan by user dan by keyword. Batasan maksimum data user yang diunduh maksimum 200 tweet dan maksimum data tweet yang diunduh maksimum 100 tweet. Ekstraksi fitur berdasarkan indeks twitter untuk data user meliputi: id user, nama dan akun user, total tweet, total followers, total status, total likes, bio profile, website dan lokasi. Sedangkan untuk ekstraksi fitur untuk data tweet meliputi: Id tweet, teks tweet, retweet, url, hashtag, mention, sumber tweet, total like, total retweet dan waktu tweet.

REFERENCE

- [1] J. Lehman, C. Castillo, M. Lalmas and E. Zuckerman, "Finding News Curators in Twitter," WWW Workshop on *Social News On the Web* (SNOW), May 13-17, 2013, , Rio de Janeiro, Brazil. ACM 978-1-4503-2038-2/13/05.
- [2] J. Lehman, C. Castilo, M. Lalmas and E. Zuckerman, "Transient News Crowds in Social Media," Seventh International AAAI Conference on Weblogs and Social Media (ICWSM), 8-10 July 2013, Cambridge, Massachusetts.
- [3] K. Haewoon, L. Changhyun, P. Hosung and M. Sue, "What is Twitter, a Social etwork or a News Media?," International Conference WWW 2010, April 26-30, 2010, Raleigh, North California, USA. ACM 978-1-60558-799-8/10/04.
- [4] V. George, S. Antonia and G. Dimitros, "A Faceted Crawler for the Twitter Service", WISE 2014, Oc 12-14, 2014, Thessaloniki, Greece.
- [5] N. Diakopoulos, "Finding and Assesing Social Media Information Sources in the Context of Journalism," *CHI'12*, May 5-10, 2012, Austin, Texas, U.S.A. ACM 78-1-4503-1015-4/12/05
- [6] W. Jason Gilmore, "Beginning PHP and MySQL", Apress, 2010. ISBN 978-1-4302-3114-1.

