Topical Categorization of Japanese Tweets

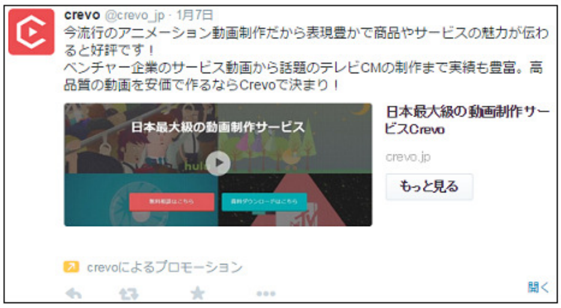
Hikari Shike s1190019 Supervised by Prof. Maxim Mozgovoy

Abstract

Many companies use Twitter for promotion purposes. However, in the current situation, the promotion tweets don't always match with the interests of Twitter users. Our application acquires the most recent 1,000 tweets using Twitter4J and analyzes the interest of the users for the tweets using HatenaKeywordAutoLink API. The developed approach is useful for matching the interest of the users and company marketing.

1 Introduction

Over the past few years, many researchers have shown an interest in data analysis of SNS (Social Networking Service) such as Twitter and Facebook. For example, Twitter's data is effectively used for disaster prevention and company marketing. Regarding company marketing, the company tweets for advertising using Twitter, and the company displays promotion tweets on any timeline. Figure 1 shows an example of a promotion tweet on Twitter user's timeline. However, because this promotion tweet may not match with the interest of all Twitter users, it does not necessarily lead to mutual advantage. On this point, the web service called "Tweet-profiling" [1] exists. It analyzes the interests of the users. In other words, it analyzes tweet topic of the users. However, it analyzes only the most recent 500 tweets. The problem seems to lie in the fact that there are a small number of tweets to analyze and the resulting display is only in pie chart.



crevo @crevo_jp · 1月7日

今流行のアニメーション動画制作だから表現豊かで商品やサービスの魅力が伝わると好評です！
ベンチャー企業のサービス動画から話題のテレビCMの制作まで実績も豊富。高品質の動画を安価で作るならCrevoで決まり！

日本最大級の動画制作サービスCrevo

crevo.jp

crevoによるプロモーション

Figure 1: Promotion tweet

In this research, we tried to solve these problems. Specifically, we double the number of tweets (the most recent 1,000 tweets) to analyze and display the results using a pie chart, a line chart. For the most recent 1,000 tweet, we analyze the category of the user's tweet topic using a pie chart. Especially for the past year's tweets among the most recent 1,000 tweet, we analyze the category of the user's tweet topic using a line chart. Because it is analyzed monthly, time transition is concrete. (see Section 2.4 for the content of the results display) Our application will be helpful for matching the interest of the users and company marketing. In other words, when the advertising personnel of companies use our application, the personnel will be able to perform efficiently behavioral targeting advertising to Twitter users.

1.1 Twitter

Twitter [2] is an online information service. It allows Twitter users to post short message.The message is called a "tweet" and it must be less than 140 characters. The users use a lot of special terms on Twitter. There are several user concepts:

Follow
Follow is used as the setting to display a specific user's tweets on one's timeline.

Follower
Follower is a user who follows a specific user.

Reply
Reply is used to post to other users. By posting a tweet in the format of "@username contents of the tweet", the tweet will be treated as a reply to other users.

Retweet
Retweet is used to re-post tweets of other users.

Timeline
Timeline is used as a screen by which followed tweets and retweets are displayed in chronological order.

Protect
Protect is used as the setting that one's tweets are not open to the public. Our application cannot acquire the user's tweets if they are thus protected.

1.2 Twitter API

API stands for Application Programming Interface. It is the specification of the interface that can be used for programming. In other words, Twitter API is the API to provide the functions of Twitter.

1.3 Twitter4J

Twitter4J [3] is the library that supports all the functions provided by Twitter API. However, it is an unofficial library. By using it, we can easily program to interact with Twitter.

1.4 HatenaKeywordAutoLink API

Hatena keyword [4] is a dictionary service on the Web. You can freely register a lot of keyword with it and classify the keyword into categories. It corresponds to Japanese. English's support is little, however basically, it does not correspond to the English. For this reason, in our study, English tweet is out of the study.

HatenaKeywordAutoLink API [5] is the API to do the following.

1. When we send any text to the API, a keyword which belongs to Hatena keyword and a category name which belongs to the keyword are extracted from the text by the API. The keyword is a keyword that has already been registered with Hatena keyword. If there is no keyword is registered with Hatena keyword in the text, nothing is extracted.

2. Extracted keyword and category name reply to the sender by the API.

Our application uses the API to analyze the user's tweets.

1.5 Apache XML-RPC

Apache XML-RPC [6] is the library of XML-RPC. XML-RPC stands for Extensible Markup Language Remote Procedure Call. This is a protocol to execute a remote procedure call on the Internet. In other words, it is a mechanism to easily execute a service on the Internet. Our application passes the tweets to HatenaKeywordAutoLink API using Apache XML-RPC.

1.6 JFreeChart

JFreeChart [7] is an open source library for generating a chart using Java. We can make an image of the chart of various forms such as a pie chart and a bar chart. Our application uses it in the results display.

2 Implementation

2.1 Registration of application

First of all, it is necessary to register our application with Twitter Developers [8] in order to make a Twitter application. At this point, we can acquire a Consumer key and a Consumer Secret key. These are the public key and the private key. The Consumer key is the information of the application. It is necessary in order to use API.

Next, it is necessary to access Twitter in order to acquire an Access Token. The Access Token is used to encrypt account infomation such as Twitter username and password. It is prepared for the third party such as a web service and is necessary to login to Twitter. In other words, it is like a "pass" for applications to access to the user's data instead of the user. Because Twitter Developers has simplified the following 3 steps to acquire the Access Token, we can easily acquire the Consumer key and the Access Token by using Twitter Developers.

1. We send the Consumer key and the Consumer Secret key to Twitter through Twitter API, we acquire a Request Token.

2. We request an Access Token using the Request Token and we acquire an Access Token.

3. We access to Twitter using the acquisitions.

2.2 Acquisition of tweets

Based on any Twitter username, acquisition of the most recent 1,000 tweets is done through Twitter4J. The tweets are acquired for each page. Each page contains 200 tweets and we can acquire 1,000 tweets when we repeat the process five times. Whenever we acquire 200 tweets, we must wait 10 seconds for the operation. This specification is incorporated in order to reduce the burden on Twitter API so as not to exceed the Twitter API limitations.

Originally, the number of limit of tweets can be acquired from Twitter API is 3,200 tweets. However, in order to acquire 3,200 tweets by our application, it takes about 3 minutes. Because this execution time is too long, the number of limit of tweets can be acquired by our application is 1,000 tweets.

In addition, based on any Twitter username, we acquire the account information such as the following.

- The number of tweets

- The number of follows

- The number of followers

- The number of retweets in the most recent 1,000 tweets

At the same time, we acquire the date and time of tweets for analysis. Figure 2 shows the main operation screen for acquiring and analyzing tweets. (see Figure 2) It takes about 1 minute in order to acquire 1,000 tweets by our application. Although the analysis time's length is depends on the user, we effectively utilize time during the analysis. While my application analyzes the tweets, the user can see the first results. For the reason, we divide the button into two. Figure 3 shows the screen in which we acquired the most recent 1,000 tweets. (see Figure 3) Figure 4 shows the screen in which we acquired the account information. (see Figure 4)

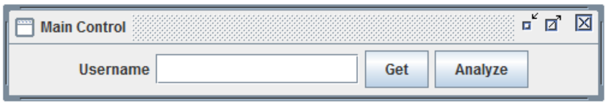
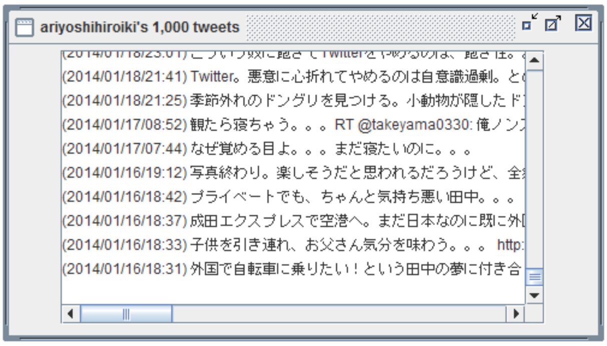


Figure 2: Main control screen


ariyoshihiroiki's 1,000 tweets

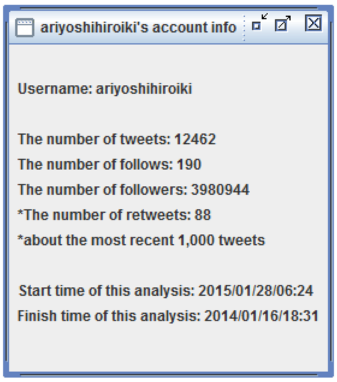
```
(2014/01/18/23:01) こういう奴に飽きてTwitterをやめるのは、飽き性。
(2014/01/18/21:41) Twitter。悪意に心折れてやめるのは自意識過剰。と
(2014/01/18/21:25) 季節外れのドングリを見つける。小動物が隠したド
(2014/01/17/08:52) 観たら寝ちゃう。。。RT @takeyama0330: 俺ノン
(2014/01/17/07:44) なぜ覚める目よ。。。まだ寝たいのに。。。
(2014/01/16/19:12) 写真終わり。楽しそうだと思われるだろうけど、全
(2014/01/16/18:42) プライベートでも、ちゃんと気持ち悪い田中。。。
(2014/01/16/18:37) 成田エクスプレスで空港へ。まだ日本なのに既に外
(2014/01/16/18:33) 子供を引き連れ、お父さん気分を味わう。。。 http:
(2014/01/16/18:31) 外国で自転車に乗りたい！という田中の夢に付き合
```

Figure 3: Acquisition of the most recent 1,000 tweets

2.3 Categorization of tweets

The tweets are passed to HatenaKeywordAutoLink API using Apache XML-RPC, and then we extract a keyword which belongs to Hatena keyword and a category name which belongs to the keyword from the tweets by the API. The keyword is classified into 14 categories.

- Book
- Music
- Movie


ariyoshihiroiki's account info

Username: ariyoshihiroiki

The number of tweets: 12462
The number of follows: 190
The number of followers: 3980944
*The number of retweets: 88
*about the most recent 1,000 tweets

Start time of this analysis: 2015/01/28/06:24
Finish time of this analysis: 2014/01/16/18:31

Figure 4: Account information

- Web
- Animal and Plant
- TV
- Anime
- Food
- Sports
- Game
- Comic
- Idol
- Society
- Science

Because the execution time of HatenaKeywordAutoLink API's two-steps process takes approximately 1 second, if we pass the 1,000 tweets to HatenaKeywordAutoLink API one by one, it takes approximately 1000 seconds. As a result, the burden of HatenaKeywordAutoLink API will increase. Therefore, we pass the 1,000 tweets to HatenaKeywordAutoLink API given month. Based on the date of the acquired tweets, we acquire a keyword which belongs to Hatena keyword and a category name which belongs to the keyword for the following tweets.

- The most recent 1,000 tweets (The data is used in section 2.4.1)
- The past year's tweets (The data is used in section 2.4.2)

2.4 Data visualization system

We will visualize the data as follows using JFreeChart.

2.4.1 Analysis of the most recent 1,000 tweets

We visualize the statistics of the most recent 1,000 tweets of any Twitter user using a pie chart. The following expression is an expression that we used for analysis. The unit is percent.

$$CategoryScore = \frac{100 \times CategoryCount}{TotalCount}$$

- TotalCount is the number of all the keywords which belong to the Hatena keyword extracted from the most recent 1,000 tweets.

- CategoryCount is the number of each category name which belongs to the keyword.

Figure 5 shows the result screen. (see Figure 5)

2.4.2 Analysis of the past year's tweets

We visualize the statistics of the past year's tweets of any Twitter user using a line chart. In the vertical axis of the chart, we display the number of each category name which belongs to the keyword extracted every month from the past year's tweets. In the horizontal axis of the chart, we display the month and the year. Figure 6 shows the result screen. (see Figure 6)

3 Results

In this research we acquired the most recent 1,000 tweets about any Twitter users, we analyzed the tweets by 14 categories, and we provided the information of the field of interest of the users. The results were displayed as shown in Figure 3, Figure 4, and Figure 5. In order to use our application, the user must perform the following steps.

1. Input any Twitter username in the text field next to ”Username” (see Figure 2).

2. Click the ”Get” button. You can acquire the most recent 1,000 tweets and the account information by this operation. The result is displayed as shown in Figure 3 and Figure 4.

3. Click the ”Analyze” button. You can analyze the most recent 1,000 tweets by this operation. The result is displayed as shown in Figure 5 and Figure 6.

4 Discussion

As we explained in Section 1, the web service called ”Tweet-profiling” already exists. For the most recent 500 tweets, it analyzes the interest of users and displays the result using a pie chart. In comparison with this web service, for the most recent 1,000 tweets, our application analyzes the interest of users and displays the result using a pie chart, a line chart.

Moreover, because our application acquires the number of followers and the number of retweets, we can know the level of the user's informational diffusing power. When a user who has a lot of those tweets or retweets, many users can get the opportunity to see the tweet or the retweet. If the company can do the promotion tweet that matches the interest of the user, the user may retweet the promotion tweet. By our application, a lot of users can get the opportunity to see the promotion tweet and the company can do efficient marketing.

5 Conclusion and Future Work

In this research, we developed an application to acquire the most recent 1,000 tweets and analyze the tweets by 14 categories. Because we can know the information of the field of interest of the users by our application, the company can do the promotion tweet that matches the interest of the user efficiently. In other words, our application will be helpful for matching the interest of the users and company marketing.

In future research, there are two objectives. The first is to increase the accuracy of the analysis so that we can do category analysis from the context. In the current situation, because we analyzed by HatenaKeywordAutoLink API, if one word has more than one meaning, all those category names with all of its meaning may return from the API. We then have to guess the category name in which the word belongs from the context, and it is therefore important to identify the necessary category name.

The second is to allow anyone to use this system. As we explained in Section 2.1, Twitter Developers has three simplified steps to acquire the Access Token. However, in this system, because the access token is built into this program, this program is used only by us. In other words, without simplifying the three steps, in order to allow anyone to use this system, there is a need to make it easier for the user to acquire the Access Token, by developing programming to carry it out.

Acknowledgements

I would like to express my gratitude to Prof. Mozgovoy for insightful advice and encouragement. Also, I

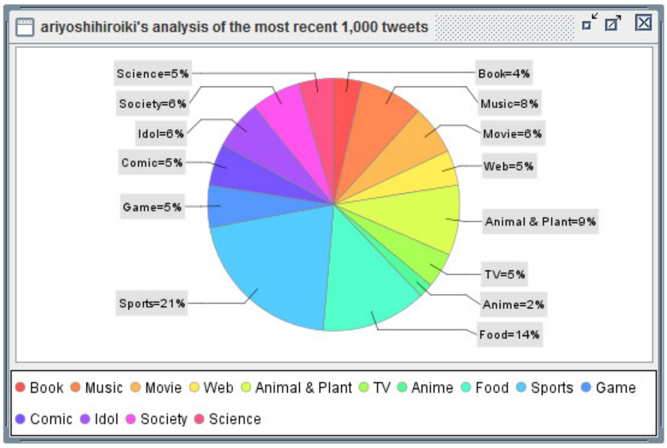


Figure 5: Analysis of the most recent 1,000 tweets

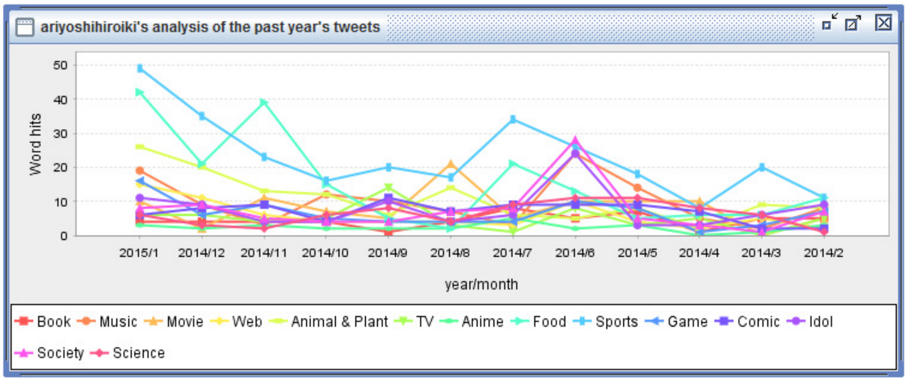


Figure 6: Analysis of the past year's tweets

would like to express my gratitude to Prof. Rozycki for generous support.

References

[1] Tweet-profiling. http://tweet-profiling.ap01.aws.af.cm/.

[2] Twitter. https://twitter.com.

[3] Twitter4J. http://twitter4j.org/en/index.html.

[4] HatenaKeyword. http://d.hatena.ne.jp/keyword/.

[5] HatenaKeywordAutoLink API. http://developer.hatena.ne.jp/ja/documents/keyword/apis/autolink.

[6] Apache XML-RPC. https://ws.apache.org/xmlrpc/.

[7] JFreeChart. http://www.jfree.org/jfreechart/.

[8] Twitter Developers. https://dev.twitter.com/.