Tweet Analysis | Place-name Identification Technology | Location Information

Technology Reports

# Analysis of Location-related Tweets and its Applications

*Twitter*[*1] *is an SNS for sharing real-time information, and this information can be analyzed to reveal trends around the world; detecting real-world events as they occur and identifying locations that are attracting attention. This article describes technology that analyzes tweets and associates a location with them. It also describes applications for such location-associated tweets, including visualizing real-time tweets on a map, detecting areas of activity for specific keywords such as "autumn colors", and analyzing the quality of communications coverage by area.*

Service & Solution Development Department† *Keiichi Ochiai*
*Daisuke Torii*
Research Laboratories *Haruka Kikuchi*
*Wataru Yamada*

## 1. Introduction

Twitter is widely used as a social media for sharing user posts in real time, and by analyzing this content, it is possible to discover trends currently occurring around the world, including events and locations that are attracting attention and the latest news. For example, NTT DOCOMO's dmenu has real-time search [1] entries called "Location-based Tweet Search" and the "Trending Spot Ranking" in the area guide [2], which gather real-time tweets[*2] related to sight-seeing spots and provide users with instant information on popular near-by locations [3]. The ability to gather information related to place-names and other locations accurately and in real-time is important for providing such localized information services. To fulfill this requirement, NTT DOCOMO has developed technology to analyze tweets and associate locations with them.

This article describes the technology for associating locations with tweets. It also describes examples of applications for location-associated tweets, including visualizing location-related tweets in real time on a map, detecting burst areas for specific location-related keywords such as "autumn colors", and analyzing mobile phone connectivity and other aspects of communication quality within coverage areas.

## 2. Linking Tweets with Location Information

There are three methods for gathering location-associated tweets.

(1) Geo-tagged Tweets

Using tweets that have been posted with geo-tags, which hold latitude and longitude information, attached.

(2) Place-name Identification Through Textual Analysis

Extracting place-names from tweets using textual analysis and using a place-name dictionary to associate a location with them.

(3) Use of Accounts Having an Associated Location



† Currently Service Innovation Department

*1 **Twitter:** The name Twitter, the logo and the Twitter bird are registered trademarks of Twitter, Inc. in the U.S.A. and other countries.

*2 **Tweet:** A term used to refer to entries on the micro-blogging service provided by Twitter, Inc.

Locations can also be associated with tweets based on the profile of the user posting the tweet. In particular, with sight-seeing spots and facilities that have their own public accounts, a location can be associated with the account.

We discuss details of these methods below. Note that geo-tagged tweets already have a location associated with them, so we discuss only place-name identification using textual analysis and the user-profile based method in this article.

## 2.1 Place-name Identification through Textual Analysis

Of the three techniques mentioned above, textual analysis to associate locations has the potential to extract more tweets with associated locations than either the geo-tagged tweet or the location-associated account methods. As such, method (2) is most effective for extracting information.

When associating locations with tweets based on textual analysis, names within the tweet text could refer to multiple different locations, or to something other than a place such as a person, and it is necessary to eliminate the ambiguity in such cases [4]. We use a place-name identification technology based on co-occurrences*3 of words to eliminate ambiguity regarding different places with the same name, and a place-name identification technology based on Conditional Random Fields (CRF)*4, which identifies place names by examining words before and after the place-name candidate, to eliminate ambiguity due to names that may refer to something (or someone) other than a place.

1) Place-name Identification Using Co-occurrences

Identifying place names using co-occurrences works on the assumption that a place name is more likely to co-occur with names of near-by places or words related to local characteristics, and this can indicate whether the name is being used as the place name or for some other purpose.

For example, "Maruyama Park" can refer to a park in Kyoto or a different park in Hokkaido, but the name "Kyoto" is more likely to co-occur when referring to the one in Kyoto. Similarly, "Matsushima" can be the name of a person, or of a place related to the poet Matsuo Bashou, so it is more likely to co-occur with "Matsuo Bashou" when referring to the place. Such words that are likely to co-occur are called co-occurrences in this article.

A flowchart of place-name identification using co-occurrences is shown in **Figure 1**. Associations between ambiguous place names and co-occurrences are stored in a DB before-hand, tweets are selected by checking if they include names in a place-name DB, the names are checked for ambiguity, and if they are ambiguous, they are correlated with the co-occurrence DB to extract only the tweets containing co-

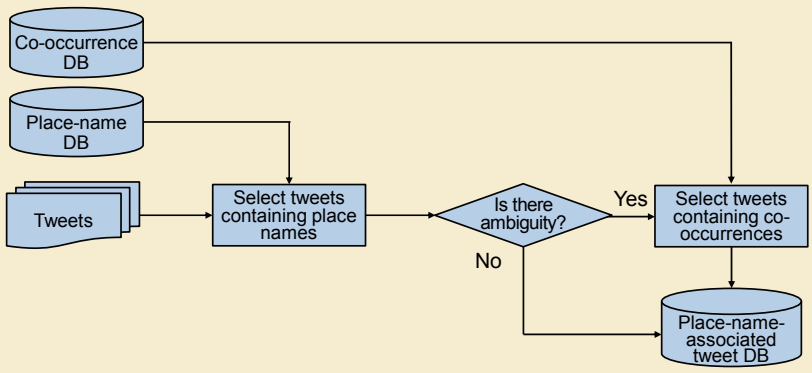


**Figure 1 Co-occurrence place-name identification flowchart**

*3 **Co-occurrence:** When two particular words appear in the same sentence.

*4 **CRF:** Conditional Random Field. A type of method for assigning pre-defined labels to a sequence of input entities based on feature values of the entities. In this document, labels for place-names and non-place-names are assigned to eliminate ambiguity regarding place names within the text, based the sentence syntax and the surrounding words.

occurrence for that place name. This enables ambiguous place names to be associated correctly with tweets.

2) Place-name Identification Using CRF

The co-occurrence place-name identification technology described above only selects tweets that contain both a place-name and corresponding co-occurrences. This is able to associate tweets with place-names very accurately, but it cannot find associations for tweets that do not contain co-occurrences. However, even tweets that do not contain co-occurrences can be identified by looking at how words other than the place names are used within the text.

For example, neither of the following contains co-occurrences:

"I had dinner at Matsushima," and

"Mr. Matsushima had dinner."

However, seeing that they feature "at" or "Mr." near "Matsushima" can be used to determine that they refer to a place and a person, respectively. These types of sentence characteristics are called features in the field of natural language processing.

Place names in the text can be discriminated automatically from non-place names by training a classifier*5 with features for place names and non-place names. For our method, we used a classifier called CRF.

A flowchart for feature extraction and CRF training is shown in **Figure 2**. For classification using CRF, tweets that include place names are selected from the set of tweets, and truth values indicating whether they are place names or person names are applied manually. Then, features are extracted from the tweets marked with truth value, and the CRF is trained to increase its classification performance.

A flowchart for place-name identification using CRF is shown in **Figure 3**. Tweets containing place names extracted from the set of tweets are first judged whether there is ambiguity as a place-name or some other word. If there is ambiguity, the CRF trained using the steps in Fig. 2 is used to discriminate between place names and other words, and then place names are associated

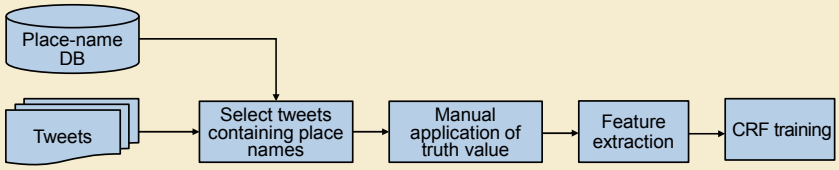


**Figure 2 Feature extraction and CRF training flowchart**

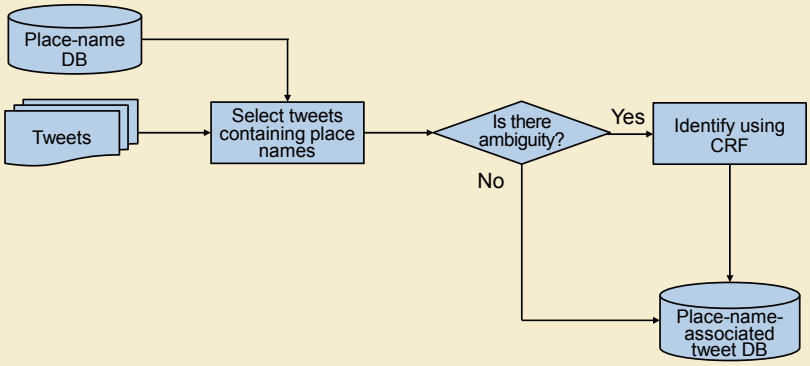


**Figure 3 CRF place-name identification flowchart**

*5 **Classifier:** A device that sorts inputs into predetermined groupings based on their feature values.

with the tweets containing them. This enables ambiguity between place names and non-place names to be eliminated, even if there are no co-occurrences, so accurate associations between place names and tweets can be implemented.

### 2.2 Use of Accounts with Associated Locations

Sightseeing spots and facilities sometimes have their own public accounts, and these public accounts often post tweets with helpful information regarding the location or facility. For example, the public account for a shopping mall might post sale or fair information, or a town hall account might post event information. As such, we investigated public accounts of locations and facilities, and used them to create associations. This enables us to extract more useful tweets using tweets from these public accounts.

## 3. Displaying Tweets on a Map and Example Applications

To display tweets on a map, latitude and longitude values indicating the display position must be assigned to each tweet. Display positions are determined either from the tweets themselves, or from the account that posted the tweet (**Table 1**). There are two issues when using tweets with associated locations, as follows:

- Whether understanding of the real-time situation is needed
- Whether tweets related to a particular topic need to be selected

### 3.1 Real-time Display of Tweets on a Map

Displaying tweets in real time on a map enables viewing of what is actually being said at the time regarding a particular location on the map. When not limiting to a particular topic, one can check the sorts of topics that are being talked about in each location. Tweets are displayed with a background color according to how the location information was determined, with blue indicating textual analysis, orange indicating a public account, and green indicating a geo-tagged tweet (**Figure 4**).

The number of tweets that can be associated with a location varies greatly with the region, and depending on the scale of the map and the screen size, it may not be possible to display all of the tweets, or there may not be a single tweet to display.

For this reason, tweets that are more likely to be useful for the user are displayed with higher priority when there are many location-associated tweets in the display area. Specifically, tweets

**Table 1 Tweet display position on a map**

| Tweet type | Display position on the map |
|---|---|
| Geo-tagged tweet | Coordinates attached to the tweet itself |
| Tweets located by textual analysis | Coordinates of place name extracted from text |
| Tweets from public accounts (e.g.: accounts of cities or regions) | Location of the public account (store, city hall, etc.) |


**Figure 4 Tweet display position on a map**

with a high number of retweets*6 are displayed first, and high-priority tweets are eventually replaced with lower priority tweets after being displayed continuously for a period of time.

On the other hand, if there are no location-associated tweets within the display area, tweets outside of the screen area are displayed with arrows indicating the direction in which they are located (shown with red frames in **Figure 5**). The arrows also work as navigation, allowing users to move to the location of the off-screen tweet by clicking them. This reduces the amount of map scrolling and zooming required to find tweets, and enables users to check tweets efficiently.


**Figure 5 Navigation to tweets off the map**


**Figure 6 Reproducing the advance of autumn colors by detecting areas of increased activity**

## 3.2 Detection of Areas with Increased Activity

Area analysis is a special case of location-related analysis. Characteristics of an entire area can be understood by analyzing with respect to multiple locations within the area instead of only one particular location. For example, it is possible to estimate the best time to see autumn colors in a given area by indexing the number of tweets related to the topic at spots in that area. Or as shown in **Figure 6**, the advancement of changing colors can be reproduced by computing indices from past tweets, and arranging the best times in sequence.

When determining characteristics of an overall area, we index according to multiple locations within the area, and avoid indexing using only a single location within the area.

To illustrate, consider the pattern of tweets produced by three spots within a given area. **Figure 7**(a) shows three tweets from a single spot, and Fig. 7(b) shows one tweet from each of three spots (assuming the example of tweets regarding autumn colors). If the index of activity is taken as simply the total number of tweets in the area, the two areas would appear the same, with three tweets. However, since the former could be due to some event or other phenomenon characteristic of the single spot, the latter, with tweets occurring at

*6 **Retweet:** A tweet from another user that is re-posted, without modifying the content. The number of times an individual tweet is retweeted is called the number of retweets.

all spots, may be a better indication of a phenomenon occurring over the whole area (such as the best time to see autumn colors). There are other possible indices that incorporate a perspective on this sort of diversity, such as the number of unique spots producing tweets (counting each spot as one, even if it produced multiple tweets).

The level of activity of an area can be defined by considering a perspective on diversity as well as quantity, such as the total number of tweets in the whole area. In Fig. 7(b), for the example, if five tweets at each of the spots were produced instead of one each, this could be considered as a higher level of activity. By computing an index that combines both the quantity and the diversity of tweets in an area for successive periods of time, the time periods with the highest levels of activity in an area (e.g., the best time to view autumn colors) can be estimated.

This sort of analysis holds promise for applications beyond autumn colors, such as cherry blossom viewing, or more broadly, for disaster situations such as typhoons, guerrilla rainstorms, tornados and earthquakes.

### 3.3 Analysis of the Communications Quality

In this section, we describe an application of location-associated tweets for analyzing the quality of communication in an area. A data processing flowchart is shown in **Figure 8**. Tweets referring to communication quality in various locations can be extracted by selecting for words related to communication service, such as "connect" or "disconnect", and the names of communication providers. A screen-shot of the system that displays the selected results on a map is shown in **Figure 9**. Fig. 9(a) and (b) show a list of place names extracted from the tweets, and the numbers of tweets selected by geo-tagging and textual analysis respectively.

Fig. 9(c) shows a tweet regarding communication quality displayed on the map at the corresponding location. Fig. 9(d) shows a time-line of tweets related to communication quality for a selected place name. Tweets with positive content, such as the word "connected", are displayed with a pink background, while negative content, such as "disconnected", are shown with a blue background. This allows the user to get a visual impression of the user comments

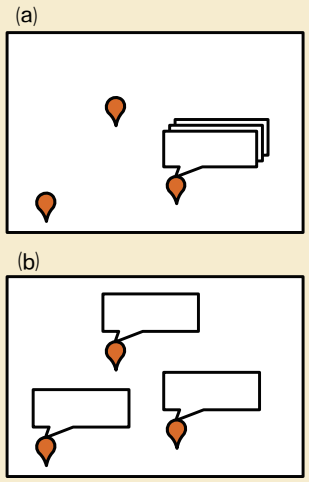


**Figure 7 Tweet patterns in a given area**

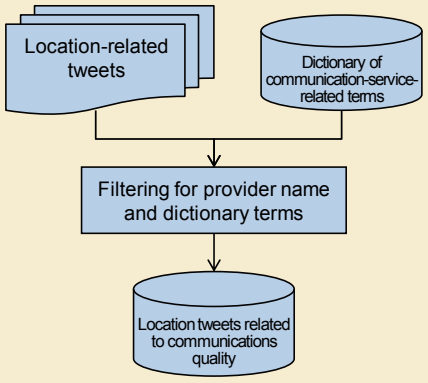


**Figure 8 Communication area quality tweet processing flowchart**


**Figure 9 Screen-shot of checking communication area quality**

for a selected location.

Work to improve the quality of communication in a given area can be supported by providing information in the form of tweets related to communication quality displayed on a map in this way.

## 4. Conclusion

In this article we describe a method for associating locations with tweets and examples of applying such location-associated tweets. Our examples included mapping location-related tweets onto a map in real-time, searching maps for burst areas with respect to specific location-related keywords such as "autumn colors", and analyzing communication quality, such as ease of mobile phone connectivity, in a given area.

In the future, we intend to expand this into a commercial service, and to perform analysis that combines this with other location-related data.

### References

[1] NTT DOCOMO: "dmenu Real-time Search." http://realtime.search.smt.docomo.ne.jp/

[2] NTT DOCOMO: "Trending Spot Ranking – docomo Map Navi." http://s.dmapnavi.jp/kanko/ranking/ranking_top.php?geo=&val=&linkfrom=T009

[3] D. Torii et al.: "Development of Real-time Search Services Offering Daily-life Information," NTT DOCOMO Technical Journal, Vol. 14, No. 4, pp. 10–16, Apr. 2013.

[4] E. Amitay, N. Har'El, R. Sivan, A. Soffer: "Web-a-Where: Geotagging Web Content," SIGIR'04 Proc. of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 273–280, 2004.