Dynamic Mapping of Dense Geo-Tweets and Web Pages based on Spatio-Temporal Analysis

Yuanyuan Wang
Nagoya University
464-8601 Japan
yuanw@db.ss.is.nagoya-u.ac.jp

Goki Yasui
Kyoto Sangyo University
603-8555 Japan
i1458085@cse.kyoto-su.ac.jp

Yukiko Kawai
Kyoto Sangyo University
603-8555 Japan
kawai@cc.kyoto-su.ac.jp

Toyokazu Akiyama
Kyoto Sangyo University
603-8555 Japan
akiyama@cse.kyoto-su.ac.jp

Kazutoshi Sumiya
Kwansei Gakunin University
669-1337 Japan
sumiya@kwansei.ac.jp

Yoshiharu Ishikawa
Nagoya University
464-8601 Japan
ishikawa@is.nagoya-u.ac.jp

ABSTRACT

Twitter evidently stirred a popular trend of personal update sharing. Twitter users can be kept up to date with current information from Twitter; however, users cannot obtain the most recent information, while they browse web pages since these are not updated in real time. Meanwhile, Twitter users are difficult to gain useful information about their current locations since these are often posted on web pages. To solve them, it is important to enrich traditional web pages with real time tweets. Therefore, we developed a novel tweet mapping system to support web and Twitter user communication through both the contents of tweets and web pages based on spatio-temporal analysis. Our system maps geo-tagged tweets to web pages by matching their location names, and categorizes tweets based on category names of floors from web pages according to different time frames. Thus, our system can effectively present the most related tweets and their summary to help users easily gain more detailed current situation in different time periods, and it also can effectively present messages from web users to help Twitter users immediately obtain useful information. In this paper, we discuss our proposed mapping method's effectiveness with our prototype system using dense tweets in urban areas.

CCS Concepts

•**Information systems** → **Location based services;** *Content analysis and feature selection;* Chat; •**Human-centered computing** → *Social media;*

Keywords

spatio-temporal analysis; geo-tagged tweets; web pages; mapping

SAC 2016, April 04-08, 2016, Pisa, Italy

DOI: http://dx.doi.org/10.1145/2851613.2851985

1. INTRODUCTION

The advent of Twitter[1] and Tumblr[2] have recently received attention and gained popularity of personal status update sharing. Twitter users can broadcast and share information about their activities, opinions and statuses in short posts, using smartphones at anytime and anywhere. Despite the useful information on Twitter, there still exists a lack of Twitter users' requirements. That is, Twitter users are difficult to obtain useful information about their current locations, e.g., bus timetables or sightseeing maps, because they are often posted on web pages, Twitter users will have to access each web page using smartphones. Meanwhile, previous works usually focus on detecting a wide range of events based on geographical areas or location mentions. For example, a probabilistic framework for estimating a Twitter user's city-level location based purely on the content of the user's tweets [1]. However, this work did not detect dense tweets by considering locations with floors or height information of landmarks (e.g., composite facilities), because locations include only latitude and longitude and data is sparse during a short period of time. Actually, there are many small events depend on the time of day such as shop sells and seasonal events based on floors of composite facilities at urban areas; users are difficult to obtain the most recent information, whilst they browse web pages since they are not updated in real time. Therefore, it is important to enrich traditional web pages with real time tweets support for web and Twitter user communication.

Although several techniques of cross-media user communication have been studied [5, 6], they have focused on user communication through the contents of tweets and web pages based on locations only, they do not solve the mentioned issues about height and temporal information of tweets with web pages. We have developed a tweet viewing system to associate web pages with the most related tweets support for web and Twitter user communication. To achieve this goal, we first acquire geo-tagged tweets based on content analysis and region selection. Therefore, our method can detect and filter tweets if they are related to, or nearby target locations, even though they do not include location names. The system then dynamically maps acquired tweets to web pages by matching lo-

[1]https://twitter.com
[2]https://www.tumblr.com/

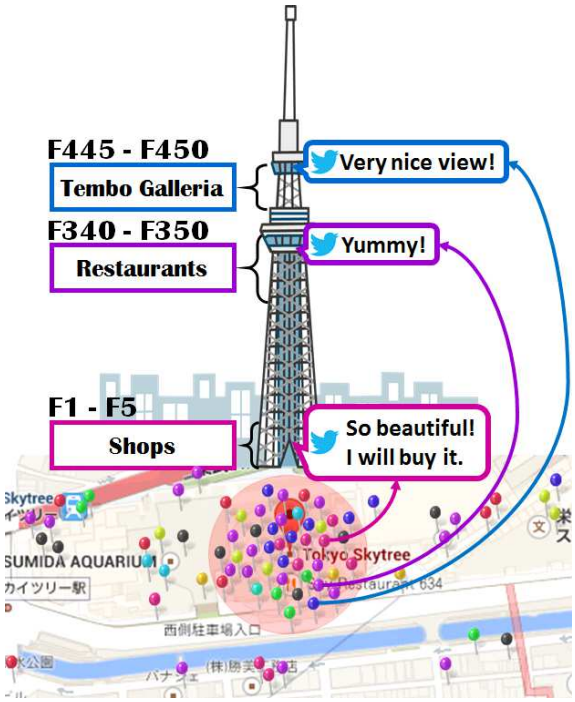

Figure 1: An example of streaming tweets of Tokyo Skytree.

cation names detected from tweets and web pages; and categorizes acquired tweets in different time frames of a day based on category names of floors from web pages. Our system has two features: 1) mapping tweets to web pages based on spatio-temporal analysis of locations, floors, and timestamps; 2) attaching a chat box to web pages support for web and Twitter user communication.

2. SYSTEM OVERVIEW

To use our system, which is on the basis of existing Web services, users are required to simply install a toolbar (a Firefox add-on), in which a streaming tweet list with a chat box is attached to each web page in a Web browser. Twitter users are required to follow an account[3] of our system for communicating with web users. Once a user browses a web page, the system records the information into a server database, which is used for mapping tweets to the web page based on a location name extracted from the tweets and the web page, and categorizing the tweets in different time frames of a day based on category names of floor information from the web page. The functions of our system are described as follows:

- A web user selects a web page to browse, the system then returns a tweet list based on a timeline, in which most related tweets with the web page are presented in a Web browser.
- When the web user sends a message in a chat box, the system presents it in a tweet list, users who browse the same web page, or Twitter users who follow our system can receive it.
- When a Twitter user replies the message of the web user through Twitter service; the system presents the reply relating to the web page in the tweet list in real time.

[3]https://Twitter.com/@RtQAService

Figure 2: Tweets associate with a page of Tembo Galleria.

Figure 3: Tweets associate with a page of The Skytree Shop.

As an example, which depicts a web user browsing an official website of Tokyo Skytree in the Web browser of our system. Streaming tweets, e.g., "Very nice view!" located on Tembo Galleria (top tweet of Figure 1), are associated with a web page of **Tembo Galleria** (see Figure 2) based on a location name, "Tokyo Skytree," and a category name of a floor, "Tembo Galleria," even though the tweets do not include them; and a tweet "So beautiful! I will buy it." located on the shop floor (bottom tweet of Figure 1), are associated with a web page of **The Skytree Shop** (see Figure 3). Likewise, a tweet "Yummy!" located on the restaurant floor can be detected when the web user browses a web page of **Sky Restaurant**. This allows the web user to gain current situation of each floor of Tokyo Skytree from presented tweets in different time periods, and he can also easily know where and when more people are in Tokyo Skytree. In addition, he can send messages to other users who browse the same web page or Twitter users who follow a Twitter account of our system.

3. MAPPING FUNCTION

3.1 Acquisition of Tweets

A conventional method based on content analysis of tweets and hashtag search [4], it can detect tweets of Twitter users who are not in current locations. However, many tweets are still not related to locations; it is difficult to report current situation from detected

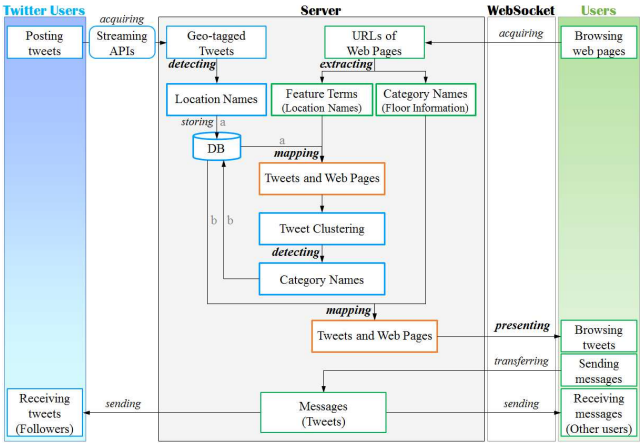

Figure 4: System configuration diagram.

tweets of a precise place. Therefore, we first obtain geo-tagged tweets from a certain region by using the Streaming API v1.1[4] (left part of Figure 4). The certain region is determined by a northeast and a southwest points, then, we obtain tweets in a rectangular region surrounding these two points. Next, we detect location names within a radius r of a region by using Google Places API v3[5], from latitude and longitude of obtained tweets. Then, our server database manages {Twitter user ID, icon URL, latitude, longitude, location name, tweet, word set, acquisition time} (central part of Figure 4).

3.2 Tweet Filtering

For filtering out tweets that have a low relation to detected locations, we analyze the content of tweets by a morphological analysis of nouns and adjectives. Then, we selected tweets that contain many feature terms (high-frequency words) describe locations. For this, we acquire a total amount n of tweets based on a given location, and calculate average frequency of each word i that appears in each tweet t. Moreover, we weight feature terms related to location names with a standard sigmoid function $1/(1+e^{-x})$ if a lot of feature terms appear in the tweet, to increase the weight of them.

$$\sum_{i=1}^{m}\left(\frac{\text{\#tweets with } i}{n}\times\frac{1}{1+e^{-x}}\right)\times\frac{1}{m} \quad (1)$$

$$x=\frac{\text{\#tweets with } i}{n} \quad (2)$$

Here, m denotes the total number of words that appear in tweet t. If Eq. (1) is more than a threshold value, t is related to its location. x as a DF value of i is calculated by Eq. (2).

3.3 Acquisition of Web Pages

To extract location names from web pages, we acquire URLs of web pages (right part of Figure 4) that users are browsing with the installed toolbar in the Web browser. Next, we extract high-frequency words as feature terms of web pages from snippets of the acquired URLs by using Yahoo! Web API[6], and we detect location names of web pages from extracted feature terms by using a morphological analyzer, called JUMAN[7]. Then, we change location names to latitude and longitude information by using Google Places API v3. Also, we extract categorize names of floor information from web pages when the web pages include floor guide information as floor information.

3.4 Clustering of Tweets

In this work, we adopt three types of machine learning algorithms as k-NN (k-nearest neighbor algorithm), naïve Bayes classifier, and SVM (support vector machine), to categorize tweets refer to different time frames of a day (8 time periods are divided by each 3 hours of a day) based on floors of composite facilities from web pages.

3.4.1 k-NN (k-Nearest Neighbor Algorithm)

It is a simple classification algorithm based on a similarity of a target data and a training data by using Euclidean distance. We extract nouns and adjectives from tweets and calculate the DF value of each word by Eq. (2) into a target set, and assign the class (category names) for each tweet into a training set. For fitting DF values of all words in each tweet, if there are n types of words appear in all tweets, vectors of each tweet are represented by an n-dimensional space. With k=8, the similarity of a target set F and a training set L is calculated using vectors of tweets as follows:

$$sim(F,L)=\sqrt{\sum_{i=1}^{n}(F_i-L_i)^2} \quad (3)$$

Therefore, we can extract the class of each training data with the highest similarity. Then, each target data is assigned to the class most common amongst its nearest training data by a majority vote.

$$class=\begin{cases} j & where\ \{c_j\}=max\{c_1,...,c_k\} \\ reject & where\ \{c_i,...,c_j\}=max\{c_1,...,c_k\}\end{cases}$$

Here, k (=8) classes of the training data are acquired by Eq. (3) in descending order.

3.4.2 Naïve Bayes Classifier

It is a simple probabilistic classifier of the supervised learning algorithm. We calculate the probability of a training set of tweets and classes (category names) as follows:

$$P(C|W_t)=\frac{P(C)P(W_t|C)}{P(W_t)}$$

Here, W_t denotes a bag of words is extracted from each tweet, and C denotes a set of classes. Therefore, we can determine the class of each tweet with the highest probability.

3.4.3 SVM (Support Vector Machine)

It is a supervised learning model for classification and regression analysis. We adopt a linear kernel for tweet classification with SVM, and it can classify the tweets by using a training set.

Based on the above, when Twitter users post tweets and a user browses a web page, the system can obtain and present tweets that are relevant based on a location name, and categorize tweets in different time frames of a day based on category names of floor information from the tweets and the web page. In this case, a database stores obtained tweets, obtained web pages, detected location names, and extracted category names (central part of Figure 4).

[4]https://dev.Twitter.com/docs/streaming-apis
[5]https://developers.google.com/place
[6]http://developer.yahoo.co.jp/
[7]http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

Table 1: Category names of floors of LUCUA Osaka.

Floor	Floor Information (Class)	Category Name
10F	LUCUA Dining	Restaurants
9F	Books	Lifestyle Goods
8F	Body Care, Cosmetics	Lifestyle Goods
7F	Men's Fashion	Fashion
6F	Accessories	Fashion
5F	Women's Fashion	Fashion
4F	Shoes, Leg Wear	Fashion
3F	Fashion Goods	Fashion
2F	Women's & Men's Wear	Fashion
1F	Seasonal Products	Fashion
B1F	Sweets, Food, Cosmetics	Sweets
No Relation	Others	

Table 2: RMSE values of categorized tweets.

Classifier	Friday	Saturday	Sunday	Average
k-NN	0.103	0.097	0.089	**0.096**
Naïve Bayes	0.368	0.365	0.199	0.311
SVM	0.096	0.080	0.088	**0.088**

4. EVALUATION

The dataset has been built retrieving 31.6 million tweets between 2014/07/30-12/31 of all Japan. In order to evaluate the accuracy of tweet categorization based on floors when locations are composite facilities, we narrowed down the test dataset in a large shopping mall nearby Osaka station, called "LUCUA Osaka," with a radius r=200m. This was totally 1,721 tweets during 2014/12/05-28 within 17:00-22:00 on Friday, 11:00-16:00 on weekend, in which many people have been shopping and eating. Table 1 shows category names based on floors were extracted from the web page of LUCUA Osaka. Since a composite facility generally has some floors are the same genre, we grouped some floors into the same categories, i.e., 1F to 7F can be grouped into "Fashion," and 9F and 10F can be grouped into "Lifestyle Goods."

There were 13 subjects identified if the tweets were related to category names or not. If the tweet was less related to its category name, subjects gave a score of 1; if the tweet was related to its category name, subjects gave a score of 2; if the tweet was strongly related to its category name, subjects gave a score of 3; if the tweet was not related to its category name, subjects gave a score of 0. Categories of tweets were defined if the average score of each tweet was the maximum value. We compared accuracies of tweet categorization with k-NN, naïve Bayes classifier, and SVM by calculating RMSE (root mean square error)[8] of #tweets and the average scores of tweets in each category. Table 2 shows RMSE values of categorized tweets, and they are explained as follows:

- The average RMSE values of k-NN was 0.096, naïve Bayes classifier was 0.311, and SVM was 0.088. k-NN and SVM are generally good results, however, many tweets were classified into "No Relation" by SVM, we need to remove them for training the learning data by using SVM.
- k-NN could identify adjectives, e.g., delicious, but not only nouns, e.g., shop names appear in the tweets. For example, a tweet "Very satisfied this amount in normal size at 650 yen! Yummy!!!" could be classified into "Restaurants."

[8]https://www.kaggle.com/wiki/RootMeanSquaredError

Through the whole results, several tweets were often wrongly categorized, when specific shop names or chain store names appear in the tweets. For instance, tweets contain a chain store name "Umeda Store" of various categories, but they were wrongly classified into the category "Sweets" only. Another problem is orthographic variants, because Japanese could be written in both Kanji and hiragana.

5. DISCUSSION

Enhanced spatio-temporal analysis. The addition of time expressions and location mentions appear in tweets to the analysis with natural language processing (NLP) techniques should help to more precisely detect and categorize tweets.

Cross-media user communication. As mentioned before, tweets are synchronized with web pages, it can support simultaneous communication between web users and Twitter users in real time to extend a function of our proposed TWinChat [5].

Indoor navigation. Based on categorized tweets of each floor of composite facilities, our system could better suggest indoor navigation to help users find popular areas or avoid crowded places based on current situations refer to [2].

Visualization. It should be possible to visualize a summary of tweets with visualization tools to portray time perspectives [3]. This should be useful for customers to understand the social data easily as a store visualization system for shopping malls.

Recommendation. As future work we plan to extend our system to accept any datasets, e.g., services and products, for discovering topics over tweets. This should be useful to recommend particular activities, products, services, events, or places to visit.

6. ACKNOWLEDGMENTS

This work was supported by SCOPE of the Ministry of Internal Affairs and Communications of Japan, and JSPS KAKENHI Grant Numbers 26280042, 15K00162, 25280039.

7. REFERENCES

[1] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: a content-based approach to geo-locating twitter users. In *CIKM 2010*.

[2] N. Fallah, I. Apostolopoulos, K. Bekris, and E. Folmer. Indoor human navigation systems: A survey. *Interacting with Computers*, 25(1):21–33, 2013.

[3] M. Musleh. Spatio-temporal visual analysis for event-specific tweets. In *SIGMOD 2014*.

[4] O. Tsur and A. Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *WSDM 2012*.

[5] Y. Wang, G. Yasui, Y. Hosokawa, Y. Kawai, T. Akiyama, and K. Sumiya. Twinchat: A twitter and web user interactive chat system. In *CIKM 2014*.

[6] G. Yasui, Y. Wang, Y. Hosokawa, Y. Kawai, T. Akiyama, and K. Sumiya. A simultaneous user communication system between microblogs and web pages [in japanese]. *DBSJ Japanese Journal*, 13-J(2):7–12, 2015.