City Happenings into Wikipedia Category: Classifying Urban Events by Combining Analyses of Location-based Social Networks and Wikipedia

Shoya Sato
Takuro Yonezawa
Jin Nakazawa
Keio University
Kanagawa, Japan
{shoya,takuro,jin}@ht.sfc.keio.ac.jp

Satoshi Kawasaki
Ken Ohta
Hiroshi Inamura
NTT Docomo, Inc.
Kanagawa, Japan
{satoshi.kawasaki.vx, ootaken, inamura}@nttdocomo.com

Hideyuki Tokuda
Keio University
Kanagawa, Japan
hxt@ht.sfc.keio.ac.jp

ABSTRACT

Recently, many researchers have been focusing on the detection and classification of urban events by information analysis on social networks. Previous works mainly use text analysis of users' posts on social networks for detecting urban events. However, this approach has a limitation that the users' posts must mention the event for the analysis to be conducted. We propose a new method for classifying urban events by extracting user interest from the location-based social network information without text analysis. The proposed method includes analyzing common friends of users in the vicinity of the event venue and extracting common friends' attributes by referring to related Wikipedia information. We designed and implemented the proposed method, and conducted an experiment for evaluating our method. Our experimental result shows that our method can classify events well in cases where participants have similar interests.

CCS Concepts

- Human-centered computing → Social network analysis;

Keywords

Urban event; classification; SNS; Wikipedia

1. INTRODUCTION

Various events are conducted in cities, including sports games, festivals, etc. Recognizing these urban events real-time is one of the key functionalities for making cities smarter in terms of disaster/emergency management, disease/health management, traffic management, and further urban planning. To detect and classify urban events, many researchers have been focusing on analyzing user-generated content-information from location-based social networks, such as geo-tagged tweets from Twitter [7]. The analysis approaches are mainly classified into the following two categories (or their cross analysis): semantic analysis [3, 9, 2] and spatiotemporal signal analysis [5, 1, 8]. Semantic analysis leverages the fact that the participants in urban events post text information about the events. By extracting hot topics or burst information from the content, urban events can be detected and classified. However, since this approach relies on the assumption that there are a number of text contents that mention an urban event, applying this approach to the urban events not mentioned considerably is difficult. In addition, even if there is considerable content, it always contains noisy information, which is not related to the events. However, spatiotemporal signal analysis focuses on the geolocation and timestamp of content, instead of semantic information. The approach analyzes the spatio-temporal pattern of signals from the location-based social network to detect abnormality in cities. However, though this approach works well for detecting abnormalities, achieving classification-inferring the type of urban events is difficult. Therefore, both semantic and spatiotemporal approaches still have limitations in the detection and classification of urban events based on location-based social network information.

In this paper, we propose an alternative simple approach for detecting and classifying urban events to resolve this problem. We focus on common followee information from a location-based social network. We assume that people participating in an event have similar interests, and these interests are reflected on their followees. For example, participants of a football game are generally interested in football, and they would also follow football related users on social networking sites. Therefore, by analyzing common followees of participants on social networks, the characteristics of the event could be inferred. Our classification method first involves obtaining common followees from the social network of users who tweet from the vicinity of an urban event site. It then weighs each common followee. Highly weighted common followees usually include famous users, who already appear in Wikipedia entries. Therefore, we
adopt the Wikipedia category of the common followees for urban event classification.

We designed and implemented an interactive analysis tool that grabs and classifies events based on the common followees information. We classified sixteen urban events by using this method and compared the results with the ground truth, which was achieved by manual classification of those events. From the result, we could confirm that classification by our method achieves a higher correlation with ground truth for the events that gather selective participants, such as a baseball game, which is attended by the fans of two teams, and a comic market, where otakus gather, as compared to generic events, e.g., a nature viewing event. The contributions of this paper are as follows:

- Presenting a new method for urban event classification based combining the analysis of common followees of event participants and Wikipedia entries
- Designing and implementing an interactive web tool that easily captures and classifies urban events
- Evaluating the effectiveness of our approach by comparing classification results of our method with event classification of human impression of the events

2. EVENTS CLASSIFICATION BASED ON COMMON FOLLOWEES ANALYSIS

2.1 Overall Process

As location-based social network data sources, we use geo-tagged tweet information from Twitter. Figure 1 shows the overall process of urban event classification. First, we extract SNS users who post textual messages from the vicinity of an event venue when the event is being conducted. Second, we extract the common followees weighting factor from the user set. Common followees imply user accounts being followed by more than two users from a user set. Then, we classify each common followee by referring to a Wikipedia article related to the common followee. Finally, we use the overall result of classified information as classification of the urban event. Details of each process are explained in the following sections.

2.2 Common Followees Weighting

This process extracts common followees that are uniquely exposed in the events, and make their weighting. A more straightforward approach is simply counting the number of users from the user set following each common follower. However, this simple approach hides important information regarding the common followees being uniquely followed by the users of the event. For example, through Twitter, we found that @masason user always appears as a popular common followee from user sets of many events. The @masason user has most followed account in Japan (2,595,630 followers). Such famous users are recommended by Twitter when a user registers on Twitter. These users may possibly show up as noise in the classification of events. Therefore, picking up common followees who are uniquely followed by users of the events is necessary. To remove noise, we calculate the ratio of following by the user sets from exact events and entire Twitter space. With the comparison information, we extract the common followee weighting factor for the events.

For detailed processing, we put a set of common followees as \( F = \{ f_0, f_1, \ldots, f_n \} \). \( \text{EventFollow}(f_x) \) represents each common followee’s total follow number by Twitter users observed in the event. We apply \( \text{EventIndex}(f_x) \) as popularity of the common followee among twitter users observed in the event:

\[
\text{EventIndex}(f_x) = \frac{\text{EventFollow}(f_x)}{\sum_{i=0}^{n} \text{EventFollow}(f_i)}
\]

Then, we also calculate \( \text{BasicIndex} \), which represents the original popularity of each common followee among a set \( F \) in entire Twitter space. \( \text{BasicFollow}(f_x) \) represents the number of followed of each common followee by all Twitter users.

\[
\text{BasicIndex}(f_x) = \frac{\text{BasicFollow}(f_x)}{\sum_{i=0}^{n} \text{BasicFollow}(f_i)}
\]

Finally, we calculate \( \text{FeatureIndex} \) of each common followee, a weighting factor of common followees, which is uniquely exposed from Twitter users observed in the event:

\[
\text{FeatureIndex}(f_x) = \text{EventIndex}(f_x) - \text{BasicIndex}(f_x)
\]

If \( \text{FeatureIndex} \) of a common followee is greater than zero, it implies that the common followee is obtaining more followers than usual. Through this process, we obtain the list of common followees with unique weighting information.
when the urban events are conducted. Then, we classify each common followees’ character.

2.3 Common Followees Classification by Wikipedia Category

Next, we infer the categories of the extracted common followees. To investigate characteristics of twitter users, several research works focused on analyzing textual information from tweets and additional metadata. For example, Marco et al. [6] conducted a linguistic Twitter analysis to infer the Twitter users’ political affiliation, ethnicity identification, etc. These approaches work effectively in detecting certain topics; however, classified categories are still limited for application to urban events. To classify common followees in various categories, we used a Wikipedia Category related to the article about common followees. We found that most common followees are influencer accounts having more than 2,000 followers. In addition, there are also exact or related Wikipedia articles for most of the common followers. Therefore, we assume that Wikipedia information related to common followees, especially the category of the article, can be utilized for classifying common followees.

Wikipedia is an online encyclopedia created and managed by the power of collective wisdom. It has more than 30 million articles in multiple languages maintaining fairness of policy as much as possible. Each Wikipedia article is associated to certain Wikipedia categories. Wikipedia categories are structured as a hierarchical category tree. Though details of a category (e.g., name of category, or level of tree at which the category exists) differ depending on different language versions of Wikipedia, any Wikipedia version has a similar higher-level category that covers various fundamental genres/keywords. For example, Japanese Wikipedia has nine categories at the top-level category (e.g., society, nature, technology, or culture). Lower-level tree categories include more details of categories, such as political system, biology, computer, or animation. By stepping up from a lower-level category to an upper-level category, categories attached to any article can be classified into the same higher-level category. We use the number of categories at a certain higher-level tree as the dimension of vector spaces for classification of the article.

To classify a common followee, specifying the related Wikipedia article of the common followee is necessary. Finding the related article is not difficult because there are several online databases that provide the relationship between famous Twitter accounts and a Wikipedia article. After specifying the article, we first collect categories associated with the article. Since these collected categories are located at a lower tree level, the upper level (parent) of category, which is preliminarily defined to be used for urban classification, is discovered by breadth first search. Some lower-level categories have multiple parents, so breadth first search is adapted to all parent trees. If the search hits a defined level of a category, our algorithm adds a score to the category.

As an example, @efm_miku the Twitter account of a virtual idol called Hatsune Miku\(^1\), which is followed by more than 40,000 users, can be associated to the Wikipedia article of Hatsune Miku. The article is classified into following categories: music software, virtual singer, vocaloid, music in Sapporo, and Crypton Future Media (a company name). These categories are located at a lower level of the category tree; thus, each category is transformed into an upper category by breadth first search. In this example, all categories are expressed, such as music software = {music:2, movie:1, computer:4} and virtual singer = {art:1, entertainment:3}. Finally, the result of totaling categories is used as a feature vector for each common followee’s classification.

2.4 Urban Events Classification

Based on common followee weighting factor and classified categories of each followee, we finally classify urban events. We define \( \vec{f}_x \) as a feature vector of an arbitrary common followee \( f_x \in F \). Then, the feature vector of the urban event is calculated as the total feature vectors of common followees while considering each of their weighting factors:

\[
\vec{e} = \sum_{i=0}^{n} \sum_{j=0}^{\text{FeatureIndex}(f_i)} \vec{f}_i
\]

3. DESIGN AND IMPLEMENTATION

We designed and implemented an interactive system that enables users to grab and classify urban events semi-automatically (see Figure 2 for overall system architecture). The system has the following functions:

3.1 Collecting and visualizing geo-tagged tweets

We collected geo-tagged tweets in Japan by using Twitter Streaming API\(^2\) (A in Figure 2). The collected tweets are stored into a database, and it can be visualized on Google Maps with date and area specifications (B in Figure 2).

3.2 Registering urban events from the map to be classified

By using the interface, users can register target events to be classified (C in Figure 2). Figure 3 shows the registration interface. Each pin represents geo-tagged tweets, and users

\(^1\)https://en.wikipedia.org/wiki/Hatsune_Miku

\(^2\)http://dev.twitter.com
can register a target event by specifying the date, time, and area (see red circle on Figure 3).

3.3 Classifying registered urban events automatically

The registered events are then classified using our method described in Section 2 (D in Figure 2). First, we extract common followees from users present where target events happened. Considering that analyzing all common followees requires high calculation costs (e.g., too many queries to Twitter API beyond its limit and also Wikipedia access), we analyzed the top ten weighted common followees for urban event classification.

Most of the geo-tagged tweets we collected are generated by Japanese users, and the extracted common followees are also related to Japan. Therefore, we used the Japanese version of Wikipedia. In terms of category tree level for urban event classification, we used a category level that has 91 categories (see Table 1). This is because this category level is recommended as a comprehensive category level by the Japanese version of Wikipedia, and we consider that it can provide enough classification category to urban events. The information at each classification process (i.e., the original categories of each Wikipedia article of common followees, level of 91 category of each Wikipedia article and result classification of the urban event) is visualized with the tool (see also Figure 3). As an example, Figure 4 shows the classification results of two different events held at the same venue at different dates. We visualized the classification of urban events, which is 91 dimension of feature vector, as a word cloud for recognizing its feature easily. The size of each keyword represents the weight of each keyword, and zero weighted keywords are not shown in the word cloud. On April 13th, 2013, an idol group’s live music concert was held at the national stadium. The analyzed categories show that classification keywords of music and entertainment were represented as strong categories. On the contrary, on January 13th, 2014, a soccer game was held at the same venue. At that time, we confirmed that classification keyword of sports was represented as strong category.

4. Evaluation

We evaluate how our classification method is similar to people’s actual impression about the event classification. We selected 13 events for our evaluation. Then, we collected the ground truth of the event classification by conducting a questionnaire survey.

4.1 Target Events

JACE (Japan Association for the Promotion of Creative Events) defines 16 types of events, including traditional festivals, art exhibitions, music exhibitions, parades, sports, contents, etc. From these, we found 12 types of events with geo-tagged tweets from their venue and classified them by using our tool. We analyzed all Twitter users who posted geo-tagged tweets when/where the event was being held.

4.2 Questionnaire Survey for Ground Truth

To validate the accuracy of our classification method, we also prepared the ground truth of the target event classification. We conducted a questionnaire survey for 27 participants (18-25 ages) to classify the events with same classification keywords as that of our method. Figure 5 shows the actual questionnaire interface. The interface provides a name, image, and simple explanation of events that were available from the event’s official website. In addition to this information, classification keywords (as same as Table 1) are also presented. When participants click on each keyword, the keyword is associated to the event. We asked participants to select multiple keywords considered as related keywords of each event. The classification result for each event from all participants is aggregated as values of the 91 dimension of vector space. We used this result as the ground truth because this result shows the extent to which people are impressed by the events.

4.3 Comparison Result

We compared the classification results of our method with the ground truth through a questionnaire survey. Both classifications are expressed as the 91 dimension of feature vector, so we evaluate the similarity of both classifications with cosine similarity, as follows.

\[
\cos(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{v} x_i y_i}{\sqrt{\sum_{i=1}^{v} x_i^2} \sqrt{\sum_{i=1}^{v} y_i^2}}
\]
introduced “popularity” of urban events that are composed of diverse participants and suggested a method of classifying events based on “popularity”. These low similarity events have a possibility of becoming high popularity events. If so, we can estimate a certain accuracy of event classification by adapting the popularity analytics at the same time.

In addition to above observations, we confirmed several features from low similarity events. Sapporo Snow Festival is one of Japan’s most popular winter events that exhibits a dozen large snow sculptures. Thus, the ground truth includes several keywords such as culture, fineart, or Hokkaido (name of the island where the event is held). However, our method classifies the events by using different keywords such as music and entertainment. These keywords are derived from high-weighted common followees of Hatsune-Miku (a virtual singer software idol). In fact, there was a special sub-event called Hatsune-Miku Snow Festival as a part of the snow festival in 2014. Hatsune-Miku is popular among internet users. This was done so that the effect on Miku’s fans may be more reflected on Twitter rather than other users. Similarly, though Hakata Dontaku is a traditional festival, our method derives entertainment and music as keywords. This is also because the idol group performed live on that day at the event, and their fans’ tweets were mainly observed. These results suggest that our classification method is significantly affected by active Twitter users. On the contrary, though the above two results are different from the public impression of the events, our result has a possibility of showing more real aspects of the snow festival.

5. CONCLUSION AND FUTURE WORK

In this paper, we propose a novel method of classifying urban events by using a Wikipedia category. Previous related works mainly focused on textual analysis of location-based SNS for event classification; the quality and number of textual information to be analyzed are always important issues. Our method analyzes common followees in users in the vicinity of the event venue and extracts common followee attributes by referring to Wikipedia information. Our experimental result showed that the proposed method can classify the events well where participants have similar interests. Therefore, we can provide an alternative classification method even if SNS texts from the area do not mention the events.

As future work, first, the accuracy of urban event classification requires enhancement. In this paper, we regard all Twitter users from an area as the event’s participants. However, some users might not be participants. Removing such noise will enhance our classification accuracy. Combining our method with textual analysis may provide a more
accurate and robust method of urban event classification. Second, we will apply our method also for event detection by constant monitoring. If the analyzed classification keywords change for an area, it implies that an event has happened. This urban event detection and classification plays an important role in the era of cities with IoT.

6. ACKNOWLEDGMENTS
This research is partially supported by National Institute of Information and Communications Technology.

7. REFERENCES


