1S-2

Inferring Users' Intentions of Visits from Twitter

Chengcheng Zhang* Haruhiko Sato Satoshi Oyama Masahito Kurihara
(Graduate School of Information Science and Technology
Hokkaido University) †

ABSTRACT

In this study, we propose a method that uses machine learning to infer the intentions of visiting places or attending events from Twitter. Our method aims at distinguishing similar but different intentions such as "Want to go" and "Confirm to go." We describe the outline of our proposed method. We also discuss several research challenges in preparing training data, extracting useful features, and building accurate classification models.

Keywords: Twitter, intentions, events, spatio-temporal data, machine learning

1. INTRODUCTION

In a microblogging service Twitter, people tweet about various aspect of their real-world experiences such as visiting places or attending events. Extracting such information from Twitter is an important technique for estimating the popularity of places or events, which will be used in recommendation and navigation systems.

So far, there have been several studies for extracting people's real-world experiences from blogs or Twitter [1,2,3]. These studies mainly focused on past information such as reports from people actually visited certain places or events. Such information is useful for estimating the popularity of places or event in the past or present. However estimating the popularity of places or event in the future will be more useful when visitors make travel plans or owners/organizers prepare to accommodate their visits. In Twitter, people frequently tweet about their ongoing activities including future plans. Therefore, in this study, we try to extract people's intentions about visiting places or attending events from Twitter.

We are interested in detecting different kind of intentions of visits, such as "Confirm to go", "Want to go" and "Have been there before."

Fig.1 shows the outline of our method. Tweets related to a certain place or event are collected from Twitter. Each tweet is determined whether it belong to one of the three classes, "Confirm to go", "Want to go" or "Have been there before", using SVM-based classifiers [4,5].

We adopt a machine-learning-based approach to extracting intentions of visits from Twitter. Sample tweets are collected from Twitter and manually classified into the four classes (three intention-related classes and the "No intention" class). These labeled data are used as training data for machine-learning-based classifiers such as SVMs.

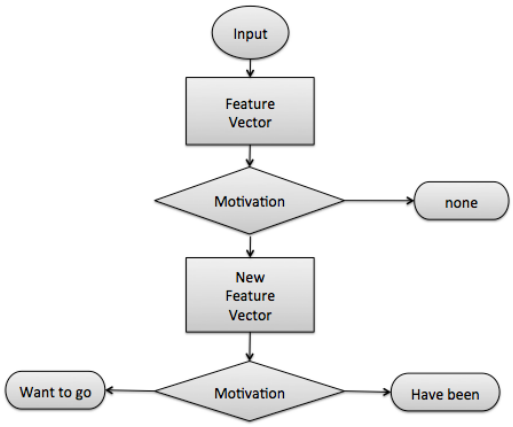

Fig.1 Outline of the proposed method

2. PROPOSED METHOD

2.1 Collecting tweets from Twitter

In our research, we use the function search of Twitter4j [6] to extract the data that match the input query, which is the name of a specific place or event, such as "Tokyo Sky Tree" or "Sapporo Snow Festival."

2.2 Using 5 features instead of the text to construct feature vectors

The characteristics of tweets are as follows:

1. The text of the tweet is short
2. There may be some mistakes in the text
3. It has a lot of new words that created by the users

Because of the characteristics of tweets, in our research we use 5 features that extract from the text of tweets instead of the whole text of tweets.

We find out that the term frequency of the auxiliary words “たい” and “た” can reflect the intentions in our research. While the text may contains more than one auxiliary word “たい” ,we have to confirm that whether the verb which is modified by the auxiliary word has the relationship to our input. To solve this problem, we calculate the distance between our input and the auxiliary word. If there is just one auxiliary word “たい”, we will calculate the distance between the input and the auxiliary word directly. Sometimes there are more than one auxiliary words in the text of tweet. Under this situation, we will calculate the average distance between the auxiliary word and the input as our feature distance.

Then, we find out that there may be some time expressions in the text of tweet which can reflect the tense of the contents that can help us infer the intentions. So we use the tense as a feature to conduct the feature vector.

Finally, the feature vector is as follows:

$$v_i = (tf_1, D_1, tf_2, D_2, tense)$$

Which i is the ID of the tweet, tf_1 is the term frequency of the auxiliary word ”たい”, D_1 is the distance from the

*zhangchengcheng@complex.ist.hokudai.ac.jp
†Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido, Japan.

input to the auxiliary word "たい", tf_2 is the term frequency of the auxiliary word "た", D_2 is the distance from the input to the auxiliary word "た", tense is the tense of the contents that we use 1 to represent the future tense and -1 to represent the past tense. If there is no tense expression in the text, the value of the tense is 0.

2.3 Using SVMs to classify the intensions

In this part, we will first label the data manually to build training data. We prepare a set of place names and event names and build training data for each place name or event.

We can confider two different classification models. One is place/event dependent classification models and the other is place/event independent classification models. The former is used to infer the intentions of visits to known places or events, while the latter is used to infer the intentions of visits to new/unknown events.

To build a place/event dependent model, an SVM is trained for each place/event using the training data collected for the place/event. To build a place/event independent model, a common SVM is trained using training data collected for different places and events.

3.EXPERIMENT

In this part, we will talk about several experiments to see whether the features work or not. We designed 3 experiments.

We used the different combinations of the features to conduct the feature vectors. Then used the SVM-Light to do the classification and observed the accuracy on the test set.

3.1 Experiment 1

In this experiment, we used the term frequency of the auxiliary word "たい" and "た" to conduct the feature vector. The training data consist of 200 tweets from twitter that have intentions and 200 tweets that do not have intentions. The test data are 50 tweets about "富士山" which are made of 25 tweets that have intentions and 25 tweets that do not have intentions. The result of the experiment is that the accuracy on test set is 81.33%. Then we chose the training data as 200 tweets from twitter that have the intention of "want to go" and 200 tweets that have intention of "have been there". The test data are 50 tweets about "富士山" which are made of 25 tweets that have the intention of "want to go" and 25 tweets that have intention of "have been there". The accuracy on test set is 92%.

3.2 Experiment 2

In this experiment, we used the term frequency of the auxiliary word "たい" and "た" and also the average distance from the name of the event to the auxiliary word "たい" and "た" to conduct the feature vector. The training data and the test data are the same as the data in experiment 1.

3.3 Experiment 3

In this experiment, we used the combination of the features in experiment 2 and the feature tense to conduct the feature vectors. The training data are the same as the data in experiment 1 and experiment 2.Then we chose 50 tweets that have the intention of "want to go" and 50 tweets that have the intention of "confirm to go" as the training data. The test data are 10 tweets about "富士山" that have the intention of "want to go" and another 10 tweets about "富士山" that have the intention of "confirm to go".

4. FUTURE WORK

In Twitter, there are few tweets related to intentions of visits while tweets with no intentions of visits are abundant. To obtain a sufficient number of tweets containing intentions, we have to sample a large number of tweets. This causes a large cost in manually labeling the training data. To reduce this cost, we have to first filter out tweets that have low possibilities of containing intentions. This filter can be build by using hand-coded heuristics or using machine learning techniques.

Even though the filter can reduce the number of tweets to be labeled, manual labeling will be still labor-intensive. We consider using crowdsourcing such as Amazon Mechanical Turk (AMT) [7] to reduce the cost of labeling.

There are several research challenges to enabling the proposed method described above. We will use the temporal difference between an event and a tweet to infer the intention of attending an event . The difference between the time of the event and the time of the tweet will be useful information. When the events or the places do not have a start time, we consider the t as 0.

5.CONCLUSION

In this paper, we proposed a method that uses machine learning to infer the intentions of visiting places or attending events from Twitter. We described the outline of our proposed method and discussed several research challenges.

ACKNOWLEDGEMENT

This research is supported in part by a grant from the Artificial Intelligence Research Promotion Foundation.

REFERENCES

[1] T. Tezuka, T. Kurashima, and K. Tanaka, Toward Tighter Integration of Web Search with a Geographic Information System, WWW 2006.

[2] 池田佳代，田邊勝義，奥田英範，奥雅博：Blog からの体験情報抽出，情報処理学会論文誌，Vol.49，No.2，2008

[3] 佐々木 建．ソーシャルメディアからの実体験情報の抽出と地図表示によるその応用．北海道大学大学院情報科学研究科修士論文，2011

[4] B. E. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. COLT 1992.

[5] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, 2001.Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[6] Twitter4j, http://twitter4j.org/ja/index.html

[7] Amazon Mechanical Turk, https://www.mturk.com/mturk/welcome