Evaluating the City Characteristics through Geo-Tagged Tweets

Hajime Murao
Graduate School of Intercultural Studies, Kobe University
hajime.murao@mulabo.org

Abstract

We try to evaluate the city characteristics from emotional expressions in texts on the net which are posted at the city. An emotional dictionary developed at WordNet Affect[1] project is used to extract emotional states from them. We have tested the proposed method on three cities in U.S.A. and two cities in U.K. Principal component analysis (PCA) has been used to investigate the difference of characteristics between the cities.

1 Introduction

Evaluating and clarifying the city characteristic is very important for city development. Clear unique characteristics attract people and which vitalize the city. Political, economic, and social activities in the city cannot neglect the characteristics. In general, the city characteristic is evaluated from environmental, economic, and social point of view. However, we think it is also valuable to identify the city characteristic from the impression of the city by people living there.

To gather people's impression of a city, we utilize the internet instead of doing questionnaires. That is, we collect texts on the internet which are posted at the city and extract emotional expressions. It must be investigated further but we believe that how many and what kind of emotional expression are included in those texts is related to people's impression of the city.

Nowadays, there are several ways to exchange texts on the internet like Facebook, Twitter, etc. In many services, users can add location information where the texts are sent from. We can use any of them but here choose Twitter. There are studies trying to utilize Twitter in a similar way to our studies, where Twitter is treated as social sensors to detect emergent events like earthquake [2, 3].

Collected text from the internet out from a city are divided into words. Each of them are then looked up in an emotional dictionary to find corresponding emotional states. Finally, we can obtain a list of emotional states with occurrence frequency. By which we define the characteristics of the city.

2 Employed Techniques

2.1 Twitter

Twitter is a social networking service (SNS) created in 2006 where users can send and receive text-based messages called 'tweets' of up to 140 characters. Users can attach geolocation tags to tweets when sending them from GPS-enabled devices like smartphones. By using it, we can specify the location where the tweets are sent from.

There is a report saying that only 4.83% out of all tweets are geo-tagged[4]. However, Twitter has over 500 million registered users as of 2012, generating over 340 million tweets per day. 4.83% of 340 million tweets per day are not small.

To collect public tweets from the specific area, we can use application programming interfaces (APIs) officially provided from Twitter service. 'Public tweets' here means tweets from non-locked users, which anyone can retrieve using official APIs.

Twitter APIs are web APIs defined as a set of standard HTTP request. For example, the following request retrieves up to 10 tweets with a keyword '`#brusselsws`' posted at the center of Kobe City.

```
https://api.twitter.com/1.1/search/tweets.json
  ?q=%23brusselsws&geocode=34.7,135.2,1km&count=10
```

Where the location and the area is specified by 34.7 in latitude, 135.2 in longitude, and 1km in radius.

We use a streaming API which enables to track tweets matches one or more filters in real-time. There is a location filter which enables to track

tweets only within a specified bounding box on the earth.

2.2 TreeTagger

Before looking up into the emotional dictionary, we divided obtained tweets into a set of words and convert each word into lemmas using TreeTagger. TreeTagger was originally developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart [5]. Which can be used to annotate texts with part-of-speech and lemma information. A sample output of TreeTagger is like the following:

Word	POS	Lemma
The	DT	the
TreeTagger	NP	TreeTagger
was	VBD	be
developed	VVN	develop
by	IN	by
Helmut	NP	Helmut
Schmidt	NP	Schmidt
.	SENT	.

2.3 WordNet Affect

In this study, we use an emotional dictionary with 1,536 entries developed at the project WordNet Affect[1]. Which is based on the concept of "six emotional states" proposed by Paul Ekman[6]. That is, one of the six emotional states: "joy", "anger", "disgust", "fear", "sadness", and "surprise" is assigned to each entry. For example, the emotional state "joy" is assigned to the entry "cheerful" and "favor", the state "anger" to "malicious" and "hate".

2.4 Python

Python programming language is used to gather all above techniques. Over 500 line program can do the following automatically: retrieving tweets in target area, converting words in the tweets into lemma, looking up into emotional dictionary, and recording the frequency of emotional states.

3 Test Settings

3.1 Target Cities

We have chosen three cities: New York, Los Angels and Salt Lake City from U.S.A., and two cities: London and York from U.K. Detailed bounding boxes shown in Table 1 taken from Google Earth are used in experiments. Figure 1 shows how London was specified.

Table 1: Bounding boxes of observed cities.

City	Upper Left Lat.,Long.	Lower Right Lat.,Long.
New York	40.6,-74.2	40.9,-73.7
Los Angels	33.7,-118.9	34.1,-117.7
Salt Lake City	38.3,-114.0	41.8,-111.2
London	51.3,-0.5	51.7,0.3
York	53.9,-1.2	54.0,-0.9

Figure 1: How London city was specified. Bounding box is drawn as a dark rectangle.

3.2 Observation Period

We have collected tweets in the specified area above for 9 days from Dec. 28, 2012 to Jan. 5, 2013 in JST.

4 Results

4.1 Word Occurrence

By observing the five different area during the periods, we obtained 3,000 to 27,000 tweets depending to the city as shown in Table 2. Figure 2 shows the 20 most occurrent words in the tweets and their emotional states. Since we collected tweets during the year change period, there is a strong bias in the occurrence. As you can see, "love", "good", and "happy" are the 3 most frequent words. This hides difference in characteristics between cities. That is, the main characteristic of all cities become same emotional state "joy".

Table 2: The total number of obtained tweets.

City	# of tweets
New York	16,681
Los Angels	15,944
Salt Lake City	3,205
London	27,722
York	4,088

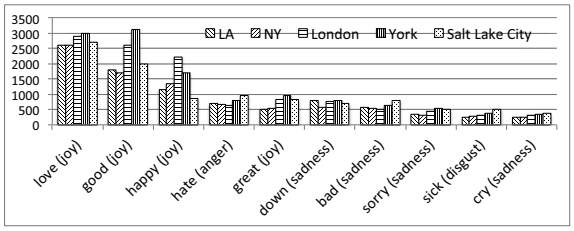

Figure 2: The 10 most occurrent words.

4.2 Principal Component Analysis

PCA has been done by treating cities as original variables and words as observation. As a result, we have obtained a set of composed variables as principal components (PCs). The importance of the first 5 PCs are shown in Table 3. The standard deviations are over 1.0 from PC1 to PC2 and the cumulative proportion becomes more than 92% with the first three PCs, which means it is reasonable to use the first 3 PCs to distinguish the cities.

Figures 3, 4, and 5 are the 10 most important words in terms of the loadings for PC1, PC2, and PC3 respectively. Which can be used to estimate the meanings of each PC. In Figure 3, the words with emotional state "joy" in the first four words have positive loadings, and the word "offense" with emotional state "anger" has negative, accordingly we can say PC1 relates to joyful image especially having somehow the mood of festival. In a similar way, we can say PC2 relates to comfortable and calm image, and PC3 somehow dark or dull image.

Table 3: The importance of principal components.

Index	PC1	PC2	PC3	PC4	PC5
Std. Dev.	1.95	1.24	1.08	0.72	0.00
Cumul. Prop.	0.54	0.76	0.93	1.00	1.00

Std. Dev. = Standard deviation.
Cumul. Prop. = Cumulative proportion.

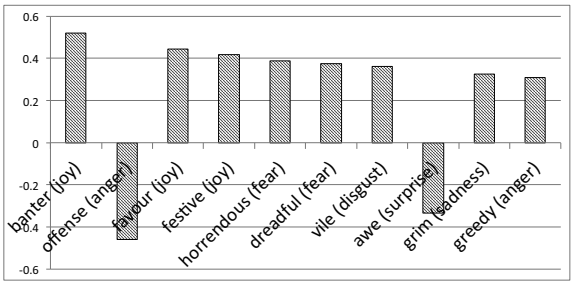

Figure 3: The 10 highest loading words for PC1.

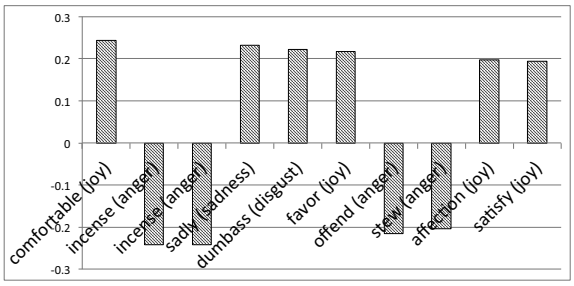

Figure 4: The 10 highest loading words for PC2.

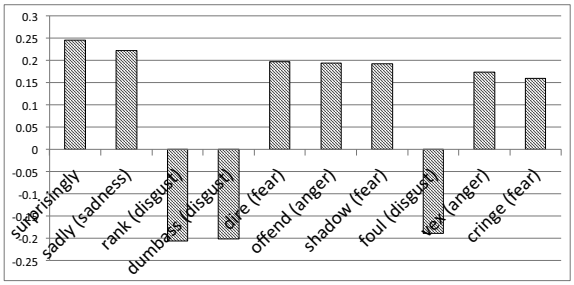

Figure 5: The 10 highest loading words for PC3.

4.3 City Characteristics

Now, we are ready to investigate the characteristics of the cities. Figure 6 and 7 are cities drawn on PC1-PC2 plane and PC1-PC3 plane respectively.

As shown in figures, the cities in U.K. and ones in U.S.A. are separated along the PC1 axis. The cities in U.K. have greater values of PC1 than the cities in U.S.A., it shows U.K. cities have a more joyful mood than U.S. cities.

As shown in Figure 6, Los Angels and New York are located far up from Salt Lake City. This means the former two cities have greater values of PC2 than the latter. It doesn't clear to us but it perhaps means coastal cities like New York and Los Angels have a more comfortable image than inland cities like Salt Lake City. This might give reasonable explanation to the positions of London and

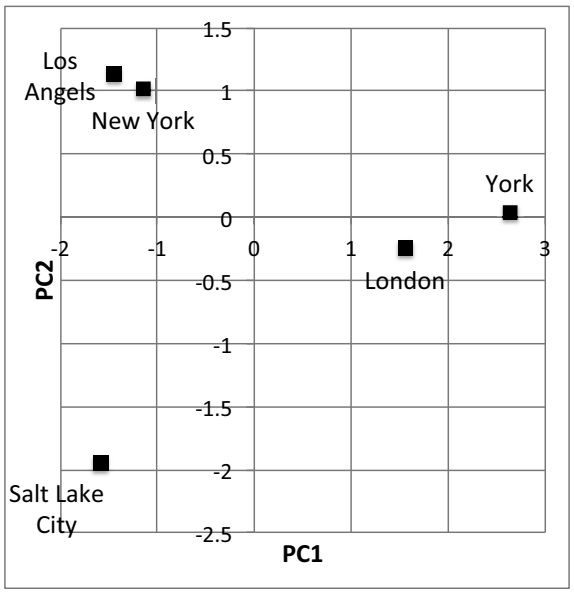

Figure 6: The city characteristics on PC1-PC2 plane.

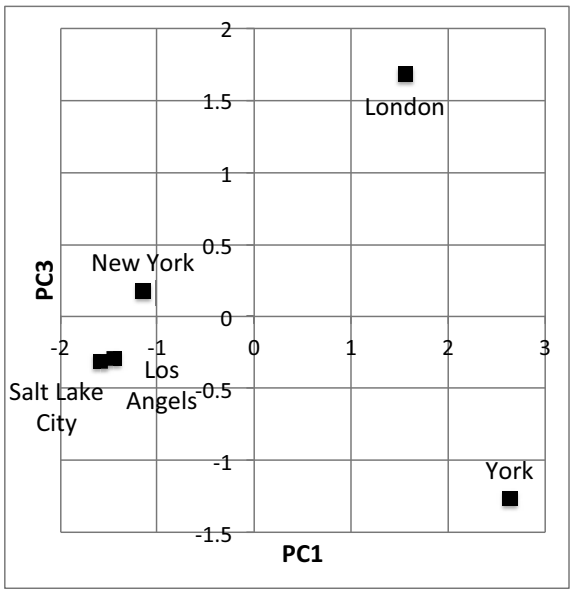

Figure 7: The city characteristics on PC1-PC3 plane.

York. They are neutral in terms of PC2, it perhaps be because these cities are not completely coastal nor inland.

Similarly, London has quite large value of PC3, which is opposite to York. Which means London has a darker or a more dull image than York. It is also unclear but perhaps weather is related.

5 Conclusion

We have proposed a method to evaluate city characteristics by using geo-tagged tweets, text-based messages with geolocation information exchanged over the internet. The method has been tested with 5 cities of U.S.A. and U.K. during the year change period. As a result, the method could reveal the differences in characteristics between the cities. However, it is not yet verified that the observed characteristics match the people's impressions to the cities. It remains as one of tasks to be solved in future works.

References

[1] R. Valitutti, "Wordnet-affect: an affective extension of wordnet," in *Proceedings of the 4th International Conference on Language Resources and Evaluation*, 2004, pp. 1083–1086.

[2] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: real-time event detection by social sensors," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 851–860.

[3] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on twitter based on temporal and social terms evaluation," in *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, ser. MDMKDD '10. ACM, 2010, pp. 4:1–4:10.

[4] Y. Arakawa, S. Tagashira, and A. Fukuda, "Relational analysis between user context and input word on twitter (in japanese)," *IPSJ SIG technical reports*, vol. 2010, no. 50, pp. 1–7, 2010.

[5] H. Schmidt, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of International Conference on New Methods in Language Processing*, vol. 12, no. 4, 1994, pp. 44–49.

[6] P. E. Ekman, "Universals and cultural differences in facial expression of emotion," in *Nebraska Symposium on Motivation*. University of Nebraska Press, 1972, pp. 207–283.