Identifying the City Personality from Text Messages transmitted over SNS with Location Information

Hajime Murao
hajime.murao@mulabo.org

Introduction

This study tries to develop a novel method to identify city personality, which characterizes a city to distinguish it from other cities. It is important for city development. Activities in cities fitting their personality will attract people and vitalize cities themselves. The city personality is also important from the political, economic, and social point of views. However, it is always underlying as a principal problem how to identify the city personality. A promising method is taking survey by questionnaire, but is labor-consuming.

I have tried to utilize social networking services (SNSs). That is, text messages transmitted over SNSs originated from a city are used to identify the city personality instead of taking survey by questionnaire. A score of each questionnaire item is automatically calculated from messages according to how many words relating to the questionnaire item appeared in the messages. I would like to note here that the proposed method is not limited to calculate city personality but can be used in other cases of taking survey by questionnaire.

A Big Five personality test has been used as a base questionnaire, which measures degrees of five factors representing human personality. It is well studied in psychology and there have been a variety of questionnaire items proposed. I have used 32 popular questionnaire items in experiments. The proposed method automatically calculates scores for each of the 32 questionnaire items from text messages in a city, which are used to evaluate degrees of five factors of personality.

Resulting traits do not represent personality of single person but people staying in the city where text messages have been exchanged. It might be called the city personality.

There are several major SNSs to be able to choose, such as Facebook, Twitter, LinkedIn, Instagram, Google+, etc. In many services, users can add location information to text messages to indicate where the messages are sent from. I can use any of them but here have chosen Twitter. There have been studies trying to utilize Twitter to detect special events in a specific area like earthquake [1,2] or festivals [3,4].

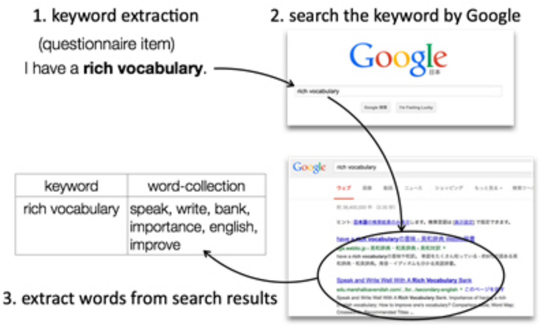
The Proposed Method

Generating Word-Collection using Google

I define a score of a questionnaire item corresponding to text messages as a function of the number of words related to the questionnaire item appeared in the text messages. To calculate the score from text messages, a collection of words related to the questionnaire item is necessary in advance. I utilize Google search system to generate the collection.

The procedure to collect words related to questionnaire items is shown in Fig. 1. Keywords are extracted from questionnaire items of the Big Five personality test firstly. For example, "rich vocabulary" is chosen as a keyword for a questionnaire item "I have a rich vocabulary." Each keyword is then searched through Google. Obtained search results are segmented and converted into a set of lemmas. A part-of-speech utility named TreeTagger is used for segmenting a text into words and converting them into lemmas.

As a result of the procedure, keywords extracted from and therefore characterizing questionnaire items, and corresponding word-collections can be obtained.

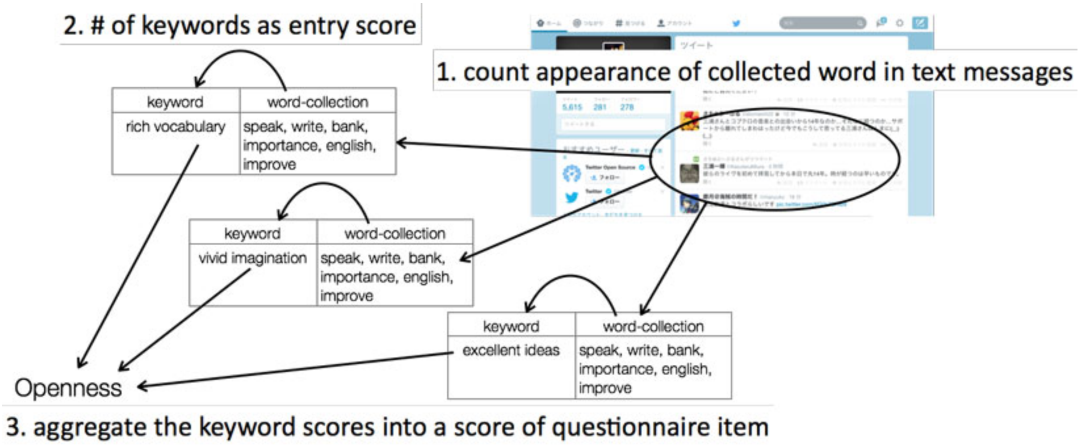

1. keyword extraction
(questionnaire item)
I have a **rich vocabulary**.
2. search the keyword by Google
Google
3. extract words from search results

keyword	word-collection
rich vocabulary	speak, write, bank, importance, english, improve

Fig. 1 The procedure to generate a word-collection related to a questionnaire item.

Calculating Personality Scores for Text Messages

The procedure to deploy the word-collections generated above to calculate scores for questionnaire items for the Big Five personality test is as follows. Text messages transmitted over Twitter originated from a specified area are collected using Twitter API explained below. These are segmented and converted into a set of lemmas using TreeTagger. Each lemma is then searched in pre-generated word-collections. A point is given to a keyword when the lemma is found in a word-collection corresponding to the keyword. In this way, scores of keywords are accumulated. A score of the each factor is calculated as an average of scores of keywords corresponding to the factor. Figure 2 shows the procedure.


2. # of keywords as entry score
1. count appearance of collected word in text messages
Openness
3. aggregate the keyword scores into a score of questionnaire item

keyword	word-collection
rich vocabulary	speak, write, bank, importance, english, improve

keyword	word-collection
vivid imagination	speak, write, bank, importance, english, improve

keyword	word-collection
excellent ideas	speak, write, bank, importance, english, improve

Fig. 2 The procedure to calculate personality scores from text messages.

Employed Techniques

Twitter

Twitter is a SNS created in 2006 where users can send and receive text-based messages called "tweets" of up to 140 characters. Users can attach geo-location 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[5]. 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 are 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.

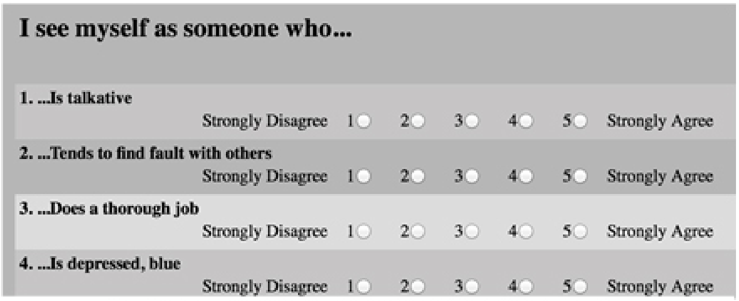
TreeTagger

A part-of-speech utility named TreeTagger is used to divide obtained tweets into a set of words and convert each word into lemma. TreeTagger was originally developed by Helmut Schmidt in the TC project at the Institute for Computational Linguistics of the University of Stuttgart[6]. It 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	.

Big Five Personality Test

A Big Five personality test is a questionnaire to measure degrees of five factors describing human personality. The five factors are "openness to experience", "conscientiousness", "extraversion", "agreeableness", and "neuroticism". It is well studied in psychology and is known that each factor is correlating specific traits of personality without overlapping. For example, "extraversion" is related to traits of "sociability", "activity", "warmth", and "positive emotions". It is also known that the five-factors structure can be found across a wide range of human in different ages and in different cultures.



I see myself as someone who...

1. ...Is talkative
Strongly Disagree 1○ 2○ 3○ 4○ 5○ Strongly Agree

2. ...Tends to find fault with others
Strongly Disagree 1○ 2○ 3○ 4○ 5○ Strongly Agree

3. ...Does a thorough job
Strongly Disagree 1○ 2○ 3○ 4○ 5○ Strongly Agree

4. ...Is depressed, blue
Strongly Disagree 1○ 2○ 3○ 4○ 5○ Strongly Agree

Fig. 3 An example of a Big Five personality test.

Thousands of different questionnaire items in different languages have been proposed so far, each of which is corresponding to one of the five factors. An example of questionnaire items is shown in Fig. 3. Usually, there are dozens of questionnaire items in a single test and a respondent gives a grade to each item in five levels according to its fitness. Resulting score of each of the five factors is calculated as an averaged sum of grades of corresponding questionnaire items.

Python Programming Language

Python programming language is used to gather all above techniques. Over 1,000 line programs can do the following automatically: retrieving tweets in target area, converting words in the tweets into lemma, looking up into word-collections preliminary generated using Google, and calculating scores of Big Five personality factors. They have also been used to search keywords extracted from the Big Five test through Google to generate the word-collections in advance.

Settings of Experiments

Target Cities

I have chosen eight cities: New York, Los Angels, Chicago, Salt Lake City (U.S.A.), London, Oxford (U.K.), Paris (France), and Brussels (Belgium). Detailed bounding boxes shown in Table 1 taken using Google Map are used in experiments. Figure 4 shows how Brussels was specified.

Table 1 Bounding boxes of observed cities.

City	Upper Left		Lower Right	
	Latitude	Longitude	Latitude	Longitude
New York	40.6	-74.1	40.9	-73.8
Los Angels	33.7	-118.5	34.2	-118.0
Chicago	41.6	-87.9	42.2	-87.5
Salt Lake City	40.7	-112.1	40.8	-111.8
London	51.3	-0.5	51.7	0.3
Oxford	51.7	-1.3	51.8	-1.2
Paris	48.8	2.2	48.9	2.5
Brussels	50.8	4.3	50.9	4.4

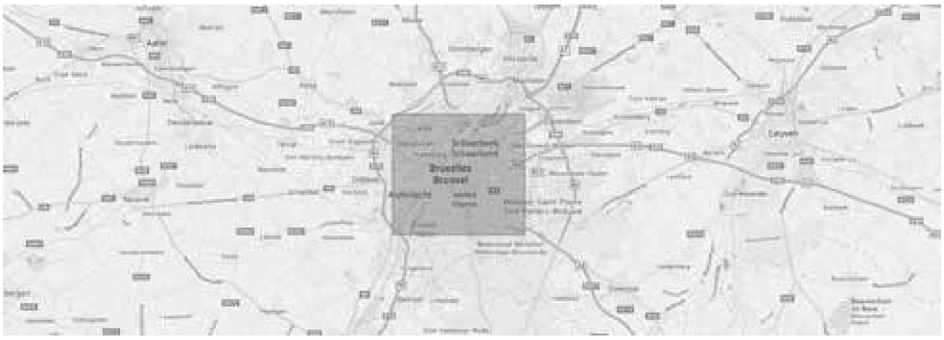


Fig. 4 How Brussels was specified. Bounding box is drawn as a dark rectangle.

Observed Period

I have collected tweets in the specified area above for about 4 months from Oct. 22, 2013 to Feb. 25, 2014 in JST. I have obtained 2,176 to 860,817 tweets depending to the city shown in Table 2. In total, 2,205,804 tweets have been obtained.

Table 2 The total number of obtained tweets

City	# of tweets
New York	180,428
Los Angels	828,180
Chicago	97,972
Salt Lake City	5,686
London	224,746
Oxford	5,799
Paris	860,817
Brussels	2,176
Total	2,205,804

Results

Big Five Personality of the Cities

There have been multiple languages used in collected tweets. Above all, English has been mostly used in U.S.A. and U.K, and French in Paris and Brussels. So, in the experiments, English and French versions of Big Five personality test are used to calculate degrees of five personality factors. Resulting degrees of five personality factors normalized between [0,5] are shown in Table 3. There are differences between cities but characteristics of cities are not so clear.

Table 3 Resulting degrees of Big Five personality factors for selected cities.

	Agree.	Cons.	Extraver.	Neuro.	Openness
Chicago	2.96	2.60	1.78	4.81	3.36
Salt Lake City	3.18	2.80	1.98	5.00	3.51
New York	2.72	2.36	1.62	4.27	3.07
Los Angels	2.17	1.88	1.19	3.66	2.48
London	3.00	2.66	1.83	4.74	3.43
Oxford	2.80	2.49	1.67	4.57	3.31
Paris	1.28	1.01	0.32	3.32	3.27
Brussels	0.85	0.55	0.00	2.49	2.53

Personality Differences between Cities

Principal Component Analysis (PCA) has been applied to clarify differences of personality between cities, where cities as original variables and personality factors as observations. Cumulative importance of principal components reveals that using the first three principal components are reasonably enough to distinguish cities.

Loadings of each of the five personality factors to the three principal components are shown in Fig.5. It can be used to estimate the meanings of each component. For example, "openness" has a large negative loading in the first principal component (PC1) and PC1 can be thought to relate to "openness". In the same way, PC2 can be thought to relate to "neuroticism". PC3 is somehow difficult to resolve its meaning where "agreeable" has a large negative loading and "conscientiousness" a large positive loading. I interpret PC3 as relating to "stubbornness".

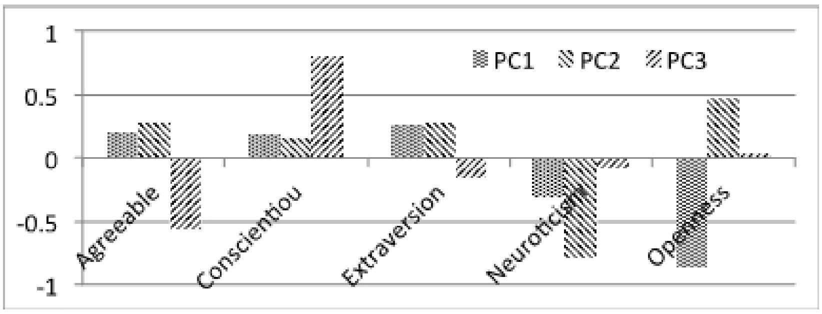


Fig. 5 Loadings of each principal component.

Figure 6 is cities drawn on PC1-PC2 plane and PC1-PC3 plane respectively. As shown in Fig. 6 (left), Los Angels is located at far right while Paris at far left. This implies Paris is more open while Los Angels less. Looking into the vertical direction, New York, London, and Brussels are located at upper position while Chicago and Los Angels at lower. This implies Chicago and Los Angels are more neurotic or nervous cities while New York, London, Brussels are more optimistic. Figure 6 (right) shows rather clear segmentation between two groups of cities. Oxford, London, Los Angels and Paris belong to the same group. They are located at upper position while other cities at lower. This implies Oxford, London, Los Angels, and Paris are more stubborn than others.

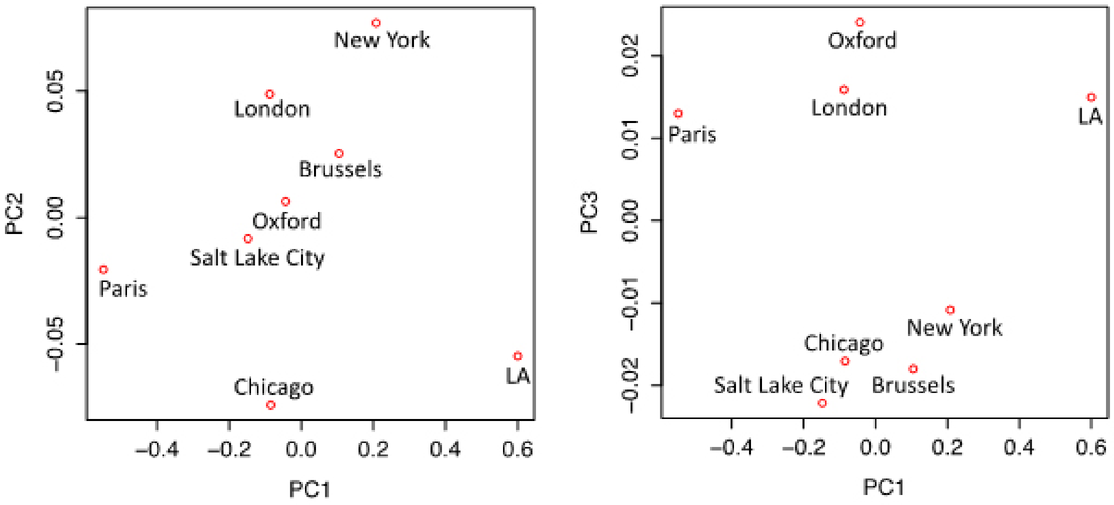


Fig. 6 The city personalities drawn on principal components planes.

Conclusion

I have proposed a method to apply questionnaire to text messages transmitted over SNS originated in specified area. Which can be used to evaluate city characteristics instead of taking survey by the questionnaire. I would like to emphasize that the method is independent to languages. That is, the proposed method can be applied to text messages in different languages simultaneously to evaluate any areas in the world.

The proposed method has been applied to evaluate city personality for 8 cities of U.S.A., U.K., and two continent cities. English and French versions of questionnaires from well-known Big Five personality test have been used for the purpose. Experiments clarify the differences between the cities while deeper investigation is till required to figure the cultural differences out.

References

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

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

[3] R. Lee, and K. Sumiya, "Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection," in Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks, 2010, pp. 1–10

[4] K. Watanabe, M. Ochi, M. Okabe, and R. Onai, "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs," in Proceedings of the 20th ACM international conference on Information and knowledge management, 2011, pp. 2541-2544.

[5] Y. Arakawa, S. Tagashira, and A. Fukuda, "Re- lational analysis between user context and input word on twitter (in japanese)," IPSJ SIG tech- nical reports, vol. 2010, no. 50, pp. 1–7, 2010.

[6] H. Schmidt, "Probabilistic part-of-speech tag- ging using decision trees," in Proceedings of In- ternational Conference on New Methods in Lan- guage Processing, vol. 12, no. 4, 1994, pp. 44–49.