Location2Vec: Generating Distributed Representation of Location by Using Geo-tagged Microblog Posts

Yoshiyuki Shoji[1][0000−0002−7405−9270],
Katsurou Takahashi[2][0000−0001−8418−8853],
Martin J. Dürst[1][0000−0001−7568−0766],
Yusuke Yamamoto[3][0000−0001−9829−6521], and
Hiroaki Ohshima[2][0000−0002−9492−2246]

[1] Aoyama Gakuin University, 5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi,
Kanagawa 252-5258, JAPAN {shoji, duerst}@it.aoyama.ac.jp
[2] University of Hyogo, 7-1-28 Minatojima-minamimachi, Chuo-ku, Kobe-shi,
Hyogo 650-0047, JAPAN {ab18y501, ohshima}@ai.u-hyogo.ac.jp
[3] Shizuoka University, 3-5-1 Johoku, Naka-ku, Hamamatsu-shi,
Shizuoka 432-8011, JAPAN yamamoto@inf.shizuoka.ac.jp

Abstract. This paper proposes a method to represent the characteristics of a place (*i.e.,* use of the venue, atmosphere of the area) by using geo-tagged microblog posts around the place. It enables a vector representation of a location similar to the distributed representation of a term in Word2Vec. Our method uses a simple neural network that is trained through the task of estimating the terms that appear in tweets posted from the area. The effectiveness of our method is illustrated through an experiment of a comparison of similar locations in Tokyo and Kyoto.

Keywords: Geo-tag · Social Sensing · Word2Vec · Social Media Analysis

1 Introduction

When we are visiting a certain location, we are concerned about its atmosphere. It is difficult to guess how the atmosphere of a location will be, and it is more difficult to search for locations with an atmosphere similar to that of a well-known location. One reason of this difficulty is lack of data; how visitors felt there or what visitors did in that place does not appear in official information or traditional Web sites. Also, we often face the situation that we need to find a location based on its usage or its atmosphere in daily life. For instance, when you move to a new city, you may look for a coffee shop that has an atmosphere similar to a familiar coffee shop in your previous home town. In such a situation, there is no point in searching for a shop with a similar name, or to search for a shop with a menu similar to the one at your familiar shop. It is more important to focus on the environment around the shop, and how other customers are feeling at the shop. There are also marketing needs; if you are the owner of a thriving shop, you want to look for a place that has similar atmosphere to your current shop's location in order to open a second shop. What kind of shops will become more popular in the area depends on the atmosphere or usage of the area.

Posts in social media are one of the most useful information sources to know the atmosphere of a location. For instance, the contents of tweets in downtown areas and residential areas are quite different. By analyzing them, we can find the difference of atmosphere between those areas.

We propose a method of generating arbitrary dimensional feature vectors of locations by using geo-tagged microblog posts. Every vector contains human information such as how people feel about the location and what people did there because it is generated from common daily social media posts. By using such feature vectors as input to machine learning methods, or for direct vector operations (*e.g.,* similarity calculation, addition and subtraction), we can compare, analyze, or search for locations by atmosphere. For instance, similarity of vectors is likely to be a good clue for information retrieval tasks that search for a location with a similar atmosphere to a given location. It is also usable for clustering areas by their atmosphere, or for area visualization. As an application, the vector is suitable to be used as input of machine learning methods, similar to embedding of terms. It will enable a more advanced social analysis of locations by using modern machine learning techniques.

Our method uses a simple neural network to create a feature vector for an object. This approach follows the basic idea of Word2Vec [15]. Word2Vec and its derivative methods train their networks to estimate linguistic contexts of terms with those surrounding terms. Finally they use the weight of the hidden layer as the feature vector of the term. The vector is called "distributed representation" of the term, and it represents the meaning of the term. This approach is based on the hypothesis "the meaning of a term is defined by the terms around it". We formulated a similar hypothesis: "the meaning of a certain location is defined by the terms appearing around it". Here, "terms appearing around it" is related to physical distance. It means terms included in geo-tagged tweets posted close to the location. Figure 1 shows an outline of the neural network that our method uses. The weight of the edges from a dimension that represent a certain place to the middle layer can be used as a feature vector of the place, in analogy with Word2Vec.

The rest of this paper is organized as follows. Section 2 describes related work. Section 3 explains our method named "Location2Vec". Section 4 describes our experiment and its results. In Section 5, we examine the experimental result. We conclude with a summary of the key points in Section 6.

2 Related Work

The purpose of our method is sensing the use or atmosphere of a location from social media posts by using a shallow neural network model. We discuss existing prior research related to our method from two view points; Social Sensing and Word2Vec-like Methods.

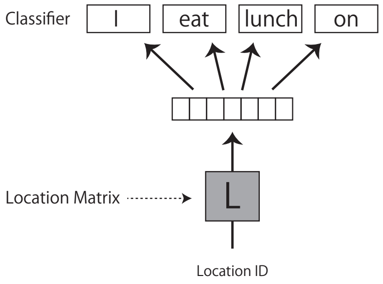

Fig. 1. Simple neural network estimating terms appearing around a given location.

Table 1. Locations in Tokyo and Kyoto sampled for the experiment.

	statistics
Size of the Area	7,854m^2
# data in full dataset	256,289
# data sampled	35,211
Sampling rate	14%
Average # tweets	33.09
Median # tweets	11
Mode # tweets	0
Max # tweets	1995

There is a lot of previous research in the field of Social Sensing. Social media posts are used to detect real place information or events [13, 1, 12, 21, 7, 9]. Geo-tagged social media posts are good resources for subjective or empirical information. Some of research uses geo-tagged posts to find physical phenomena related to the location, such as earthquakes [19], weather events [6], intention of moving [23], and so on. Our work is also an instance of a Social Sensing approach. Our method incorporates both term features and coordinates of tweets into its algorithm.

There also exist some papers which expand topic modeling methods with location-based features. These methods use the physical distance as the relationship between terms or topics. Canh *et al.* [4] proposed an extended model of LDA (Latent Dirichlet Allocation) named “Spatial LDA”to find regional communities. Kurashima *et al.* [11] proposed the “Geo Topic Model”, which is a new topic model that uses information about users and locations to estimate topics. Ahmed *et al.* [2] proposed a hierarchical geographical modeling method of combining location and text content in social networking services. Yin *et al.* [26] proposes LGTA (Latent Geographical Topic Analysis), which combines both location-driven methods and text-driven methods to detect topics appearing in geo-tagged social media posts. The method that we propose also calculates the feature vector of a place. Our vectors will have properties similar to vectors made by these topic modeling methods; containing subjective information of the place such as usage, opinion or atmosphere. The main difference is that we chose a Word2Vec-based neural network method to make a vector. Our method can accept various improvements of Word2Vec applications.

Word2Vec is a family of methods to generate distributed representations of words, proposed by Tomas Mikolov in 2013 [15]. Word2Vec has many derivative methods and applications. All of them have in common that they use a shallow neural network model. The most famous derivatives are Doc2Vec and Paragraph2Vec proposed by Mikolov himself [16]. These methods are frequently used for vectorization of words or documents. There exist many methods combining Word2Vec with other natural language processing methods in order to improve performance and create additional applications (*e.g.,* LDA2Vec [17]).

Nowadays, the Word2Vec-based approach is a popular embedding method for deep learning. There are many derivative methods which accept other kinds of data; not only terms or documents, but also graphs, multimedia data, and so on. RDF2Vec [18] creates distributed representations of graph entities. It converts sub-graphs into sequential tokens. Node2Vec [8] is another Word2Vec-based graph vectorization method. A characteristic of this method is how to sample a node's neighborhood feature; it uses random walk as a negative sampling method. As an application for multimedia data, Madjiheurem *et al.* [14] proposed Chord2Vec, which converts musical chords to a feature vector. Alcorn [3] proposed a method named "(batter—pitcher)2Vec", a unique derivative method of Word2Vec, which vectorizes major league baseball players. Place2Vec [25] is also a model which is very similar to our method. It estimates a location by its surrounding locations. The main difference is that we use physical distance instead of conceptual distance such as co-occurrence of terms or hop count in a graph.

There are some methods applying Word2Vec to social media posts. Dhingra *et al.* [5] proposes "Tweet2Vec", an approach for vectorizing non-geo-tagged tweets. It uses hash tags in Twitter as an objective of the estimation. There is another algorithm called "Tweet2Vec" [22] that does not use hash tags, but adopts a method similar to LSTM (Long short-term memory) and its auto encoder. Word2Vec is typically used to summarize sentiments and opinions from social media posts. One of the most common usages of Word2Vec in social media analysis is vectorization of the posts in the same way as other topic modeling methods or dimensionality reduction methods. Wang *et al.* uses Word2Vec to estimate the usage of the area from geo-tagged posts from Sina Weibo [24]. Seki *et al.* [20] also uses Word2Vec to estimate the sentiment of the area by analyzing Twitter posts which are posted by users who live in the area. These methods use standard Word2Vec to vectorize social media posts by using co-occurrence of terms. Our method incorporates geographical factors into the calculation itself.

3 Location2Vec Algorithm

In this section, we describe the details of our method named "Location2Vec". The novelty of our method is the data used in the estimation. Word2Vec uses surrounding terms and a central term as input and output of their estimation. Our method uses the location instead of a central term, and uses the terms appearing in the tweets posted in the area around that location instead of surrounding terms.

3.1 Training Data Creation

Our method generates a feature vector of a location from geo-tagged tweets around the location. This idea can be considered as an analogy of the concept of "context window" in Word2Vec. That is, it uses "terms that appeared in a geo-tagged tweet that was posted within 10 meters from the location" instead of "terms that appeared within 10 words around a term".

A specific location, such as a shop, an institution or a restaurant has coordinates represented as longitude and latitude. The most primitive method is to use tweets posted at a location within a radius of n meters from the target venue. We calculated the distance between the venue and the location of a tweet by the Hubeny formula. Although it is a reasonable method for representing a place by using the tweets around it, there is a problem in the training of the network. Sometimes several venues are located at the same coordinate when they are in a multistory building or in a commercial complex. In addition, in dense areas, different kinds of venues will share their surrounding tweets. The training of the neural network often fails under these circumstances. For instance, it happens that the classifier has to estimate different outputs from the same input. A reasonable way to eliminate the duplication of tweets is isolating areas uniformly by a grid. In this case, instead of a certain venue, the area is expressed using the tweets included in the grid. The size of the grid will have the same effect as the window width of Word2Vec. A fixed size grid will work well if the density of tweets in the dataset is homogeneous. Yet, it should be considered how to divide the map. For example, by changing the cell size according to the density of tweets, or using area division methods such as a Voronoi diagram. There is a possibility of improving the accuracy of estimation.

3.2 Calculation

Our method follows the skip-gram model of Word2Vec. When a document ID is given, the method estimates the list of words included in the given document. Our algorithm uses physical areas on the earth, separated by certain conditions, in the same way the original algorithm uses documents. It uses words included in tweets actually tweeted within a certain area instead of words in a certain document.

The model is shown in Fig. 1. The input is a square matrix consisting of one-hot vectors. Dimensions of each vector and the number of vectors are similar to the number of locations in the dataset. The output is a set of vectors that represents which terms were used in each location. The dimension of each vector is similar to the number of terms used in the dataset, and the number of the vectors is similar to the number of locations. Thus the m-th dimension of the n-th vector is 1 if the m-th term appeared in posts around the n-th location.

4 Experiments

We evaluated our method with an application similar to analogy-based information retrieval. The aim of this experiment is to verify the usefulness of our method for an actual location search system.

We performed an evaluation task that analyzes the similarity between locations in two different cities. This task is related to analogy-based search. Analogy-based search [10] is an Information Retrieval model which accepts an example in a known domain as a query, and returns similar entities in an unknown domain. For instance, if a user wants to know “The location in Singapore that is similar to the Statue of Liberty in New York”, then system should return “Merlion”. This kind of search model is reasonable for search tasks with ambiguous information needs, such as the atmosphere of a place.

We conducted a experimental task that discovers pairs of areas with similar atmosphere in Tokyo and Kyoto in Japan. The reason why we choose Tokyo and Kyoto is because both are large cities in Japan (first and fourth). Therefore, it is easier to collect a large number of geo-tagged tweets, sufficient to train a neural network-based machine learning method. In addition, it is easy to find evaluators for the experiment who know both Tokyo and Kyoto.

4.1 Dataset

To compare the two areas, we collected tweets posted in the two cities, and extracted some of them. Tweets were collected from April 2016 to March 2017 from the Twitter Streaming API. Table 1 shows the statistics of the dataset. This dataset contains sampled foursquare venues located in Tokyo or Kyoto. First, we extracted venues in the rectangles of Tokyo area and Kyoto area. As both Tokyo and Kyoto are big cities, 31 percent (256,289 / 826,266) of the venues in Japan were included in the dataset. To reduce calculation cost, we randomly sampled 14 percent of the venues from the whole dataset. Next, we extracted tweets posted around each venue. We set a radius $r = 50m$. Tokyo and Kyoto are high density cities. Our dataset contains many venues which share surrounding tweets with other venues. Finally, we cleansed data, as follows. Our dataset contains many useless geo-tagged posts; such as posts by bots, “Check-in” posts of geo-social networking services, and replies for friends. We removed tweets that meet any of four conditions below, and keeping 5.15 percent (1,165,109 / 22,640,460) of tweets:

- tweets that include URIs, or other users’ IDs (*i.e.,* replies),
- tweets that do not contain any Japanese characters,
- tweets that contain terms characteristic for bots, and
- tweets by a user who posted more than 20 times in the same location.

4.2 Methodology

We analyzed pairs of locations in Tokyo and Kyoto. Areas have their own atmosphere, such as atmosphere of exclusive residential area, atmosphere of desolate town and so on. First of all, we generated 200 dimensional vectors for each venue. We used “keras”, the python TensorFlow library for neural network calculation. We used the default parameters of the library for estimation.

Next, we compared pairs of locations by hand. We created a ranking of pairs consisting of a location in Tokyo and a location in Kyoto by their similarity. Euclidean distance was adopted as the similarity measure in 200 dimensional vector space. Our dataset contained 31,772 venues in Tokyo, and 3,439 venues in Kyoto.

Table 2. Top 10 "similar" pairs in Tokyo and Kyoto. The pair in each line was estimated as places that have same atmosphere .

Place in Tokyo	Place in Kyoto
Dining district in a small station	Peaceful residential area
Dining district in high-end area	Dining district in high-end area
Commercial building in big station	Commercial building in midsize station
Dining district in a old town	Out of downtown
Road-side in commuter town	Residential area in suburb
Backstreet of shopping street	A famous temple
University	Event areas near Kyoto station
Luxury office area	Road-side in suburbs
Station in a commuter town	Interchange in suburban area
Residential area along a river	Residential area next to imperial palace

Table 3. Top 10 "dissimilar" pairs in Tokyo and Kyoto. The pair in each line was estimated as dissimilar combination.

Place in Tokyo	Place in Kyoto
Bar district	Shrine in a mountain
Shopping mall	College town
Dining area in deserted station	Famous shrine
In front of famous market	Marine museum
Subway station in residential area	Famous shrine
Suburban residential area	History museum
Anime vocational school	Marine museum
Park	Center of downtown
Station in old town	Commercial complex
Dining area in deserted station	Guest house at mountain

We calculated the similarity of 109,263,908 pairs. Two evaluators analyzed the top 100 pairs and bottom 100 pairs one by one. They discussed locations' atmosphere, other venues around the location, usage of areas, and analyzed tweets around there.

4.3 Result

We compared pairs of two locations in Tokyo and Kyoto. The ranking of pairs is shown in table 2. We translated the concrete location to a short description. We were able to find some pairs of dining districts at the top of the ranking. People often tweet at restaurants; they post what they ate, and the evaluation of the restaurant. In dining district, it is likely to easily find clues to calculate similarity. Our method also finds many pairs of residential areas. In these areas, we were able to find daily tweets posted by residents. Through the top part of ranking, pairs of suburban residences frequently appeared. The reason may be that they are just numerous in our dataset, and all of them are similar to each other. Conversely, Table 3 represents the ranking of the most dissimilar pairs of locations. It contains many incomprehensible combinations of locations.

5 Discussion

There is some room for improvement in the proposed method. First, a vector generated by the proposed method is affected by the amount of documents for the target place. It was assumed that tweets within a range of 50 meters from the target place characterize the place itself. However, if the number of tweets is too small, we cannot calculate the characteristics of the location well. Since the appropriate distance would be different for each area, we should consider dynamic distance configuration.

The quality of the tweets is also an important factor. We deleted unnecessary tweets such as ones by bots. However, there are still many tweets that do not represent features of the target place (*e.g.,* daily reports, everyday conversations between users). A clean dataset is important for good results, even if strong data cleansing reduces the amount of available training data.

Applications using Location2Vec can also be improved. We conducted an experiment to discover similar places among different regions. We calculated the distances of vectors at two places from different regions to realize such an application. The vector of a certain place in Tokyo vaguely contains the feature of Tokyo as a whole. If we can acquire the characteristics of Tokyo as a vector, we can express the unique features of the target place in Tokyo more clearly by subtracting the vector representing the characteristics of Tokyo from the vector at the target place. One way to obtain such a vector is to acquire the average vector at every place in Tokyo in the proposed method. Alternatively, in machine learning, it is conceivable to use a neural network that takes regional hierarchy into account.

6 Conclusion

We proposed a method of generating a feature vector of a place from tweets posted around the place. Since the feature vector is made from social media data, this vector may contain subjective information such as use of the venue or atmosphere of the area. We followed the Word2Vec algorithm and its basic idea. Our method uses a simple neural network which is trained through the estimation of terms that appear in tweets posted around the place.

We conducted an experiment: a case study of analogy-based search by using geo-tagged tweets in Japan. It showed the possibility of our method for social analysis.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers JP18K18161, JP17K17832, JP18KT0097, JP16H02906, JP16H01756, JP17H00762, JP18H03243.

References

1. Aggarwal, C.C., Abdelzaher, T.: Social sensing. In: Managing and mining sensor data, pp. 237–297. Springer (2013)
2. Ahmed, A., Hong, L., Smola, A.J.: Hierarchical geographical modeling of user locations from social media posts. In: Proceedings of the 22Nd International Conference on World Wide Web. pp. 25–36. WWW '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2488388.2488392, http://doi.acm.org/10.1145/2488388.2488392
3. Alcorn, M.A.: (batter—pitcher)2vec: Statistic-free talent modeling with neural player embeddings. In: MIT Sloan Sports Analytics Conference. p. 5435 (2016)
4. Canh, T.V., Gertz, M.: A spatial lda model for discovering regional communities. In: 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013). pp. 162–168 (Aug 2013). https://doi.org/10.1109/ASONAM.2013.6785703
5. Dhingra, B., Zhou, Z., Fitzpatrick, D., Muehl, M., Cohen, W.W.: Tweet2vec: Character-based distributed representations for social media. In: The 54th Annual Meeting of the Association for Computational Linguistics. p. 269 (2016)
6. Doran, D., Gokhale, S., Dagnino, A.: Human sensing for smart cities. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 1323–1330. ASONAM '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2492517.2500240, http://doi.acm.org/10.1145/2492517.2500240
7. Giridhar, P., Wang, S., Abdelzaher, T., Al Amin, T., Kaplan, L.: Social fusion: Integrating twitter and instagram for event monitoring. In: Autonomic Computing (ICAC), 2017 IEEE International Conference on. pp. 1–10. IEEE (2017)
8. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 855–864. ACM (2016)
9. Kamath, K.Y., Caverlee, J., Lee, K., Cheng, Z.: Spatio-temporal dynamics of online memes: a study of geo-tagged tweets pp. 667–678 (2013)
10. Kato, M.P., Ohshima, H., Oyama, S., Tanaka, K.: Query by analogical example: relational search using web search engine indices. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 27–36. ACM (2009)
11. Kurashima, T., Iwata, T., Hoshide, T., Takaya, N., Fujimura, K.: Geo topic model: joint modeling of user's activity area and interests for location recommendation. In: Proceedings of the sixth ACM international conference on Web search and data mining. pp. 375–384. ACM (2013)
12. Lee, R., Sumiya, K.: 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. pp. 1–10. ACM (2010)
13. Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., Shi, L.: Social sensing: A new approach to understanding our socioeconomic environments. Annals of the Association of American Geographers **105**(3), 512–530 (2015)
14. Madjiheurem, S., Qu, L., Walder, C.: Chord2vec: Learning musical chord embeddings. In: Proceedings of the Constructive Machine Learning Workshop at 30th Conference on Neural Information Processing Systems (NIPS2016), Barcelona, Spain (2016)

15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
17. Moody, C.E.: Mixing dirichlet topic models and word embeddings to make lda2vec. CoRR **abs/1605.02019** (2016), http://arxiv.org/abs/1605.02019
18. Ristoski, P., Paulheim, H.: Rdf2vec: Rdf graph embeddings for data mining. In: International Semantic Web Conference. pp. 498–514. Springer (2016)
19. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes twitter users: real-time event detection by social sensors pp. 851–860 (2010)
20. Seki, Y.: Use of twitter for analysis of public sentiment for improvement of local government service. In: 2016 IEEE International Conference on Smart Computing (SMARTCOMP). pp. 1–3 (May 2016). https://doi.org/10.1109/SMARTCOMP.2016.7501726
21. Sheng, X., Tang, J., Xiao, X., Xue, G.: Sensing as a service: Challenges, solutions and future directions. IEEE Sensors journal **13**(10), 3733–3741 (2013)
22. Vosoughi, S., Vijayaraghavan, P., Roy, D.: Tweet2vec: Learning tweet embeddings using character-level cnn-lstm encoder-decoder. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 1041–1044. ACM (2016)
23. Wakamiya, S., Jatowt, A., Kawai, Y., Akiyama, T.: Analyzing global and pairwise collective spatial attention for geo-social event detection in microblogs. In: Proceedings of the 25th International Conference Companion on World Wide Web. pp. 263–266. WWW '16 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland (2016). https://doi.org/10.1145/2872518.2890551, https://doi.org/10.1145/2872518.2890551
24. Wang, Y., Wang, T., Tsou, M.H., Li, H., Jiang, W., Guo, F.: Mapping dynamic urban land use patterns with crowdsourced geo-tagged social media (sina-weibo) and commercial points of interest collections in beijing, china. Sustainability **8**(11), 1202 (2016)
25. Yan, B., Janowicz, K., Mai, G., Gao, S.: From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. pp. 35:1–35:10. SIGSPATIAL'17, ACM, New York, NY, USA (2017). https://doi.org/10.1145/3139958.3140054, http://doi.acm.org/10.1145/3139958.3140054
26. Yin, Z., Cao, L., Han, J., Zhai, C., Huang, T.: Geographical topic discovery and comparison. In: Proceedings of the 20th international conference on World wide web. pp. 247–256. ACM (2011)