Topic-Based Place Semantics Discovered from Microblogging Text Messages

Eunyoung Kim
Department of
Computer Science,
KAIST, Daejeon, Korea
ey_kim@kaist.ac.kr

Hwon Ihm
Division of Web Science
and Technology
KAIST, Daejeon, Korea
raccoon@kaist.ac.kr

Sung-Hyon Myaeng
Division of Web Science
and Technology
KAIST, Daejeon, Korea
myaeng@kaist.ac.kr

ABSTRACT

Location-based social network services (LBSNS) such as Foursquare are getting the highlight with the extensive spread of GPS-enabled mobile devices, and a large body of research has been conducted to devise methods for understanding and clustering places. However, in previous studies, the predefined set of semantic categories of places play a critical role in both discovery and evaluation of the results, despite its limited ability to represent the dynamics of the places. We explore beyond the predefined semantic categories of the places and discover topic-based place semantics through the use of Latent Dirichlet Allocation, by extracting topics from the text which people post on site. We also show the proposed method allows for understanding the temporal dynamics of the place semantics. The finding of this study is intended for, but not limited to, context aware services and place recommendation systems.

Categories and Subject Descriptors

H.1.2 [**User/Machine System**]: Human Factors and Human Information I.2.7 **[Natural Language Processing]**: Text Analysis

General Terms

Human Factors, Experimentation, Languages

Keywords

Place Semantics, Location-Based Social Networks, Latent Dirichlet Allocation, Topic Modeling

1. INTRODUCTION

In the application domains, it has been well known that rather than obtaining only the latitude and the longitude of a user's location, it is more useful to employ the concept of place where people actually impart a meaning: social meaning, conventions, cultural understanding about the roles, function and nature [1]. Owing to the prevalence of location-based social network services (LBSNS) such as Foursquare and Facebook Places, researchers can now easily access and utilize data generated from such services for many bodies of research.

However, only few previous studies use the textual content accompanied by the geospatial data as in Foursquare, for example. Several studies explored how topics were different for geographical locations, in terms of the latitude and the longitude of the places [3, 5], but only presented city-level, or even country-level analyses, conducted in units much larger than what people generally perceive as places.

In our research, we attempt to elicit place semantics, the conceptualization of a place, derived from the crowd by analyzing the text data gathered from Foursquare. Our main goal is to collect the text messages written from different places in different times, analyze them automatically for salient topics, and determine the corresponding place semantics. We believe that the Latent Dirichlet Allocation, a generative probabilistic model for collections of discrete data, is a reasonable choice for modeling such latent variables. To the best of our knowledge, our approach is the first attempt at mining topic-based place semantics from the crowd-generated text and understanding the temporal dynamics.

2. DATASET

In this work, we use Foursquare check-ins and shouts for the places in New York City, collected indirectly through Twitter for eight months between March 19th and November 5th, 2012, in order to secure a sufficient number of shouts and venues in the dataset while limiting the scope at the same time. We filtered out the tweets consisting only of a text automatically generated by Foursquare. Our final dataset consists of 453,150 check-ins containing shouts, and a complete set of attributes for the 72,415 venues that appears in our dataset.

3. DISCOVERING PLACE SEMANTICS

3.1 Topic Discovery

We discover place semantics by analyzing the topicality of the Foursquare shouts. Shouts posted from a venue are aggregated to form a "mega-document", which is assumed to contain various descriptions and expressions about the places. That is, we end up building a collection of documents for all the venues in our dataset, which belong to one or more of the nine categories at the top level in Foursqaure. For topical analysis, we applied Latent Dirichlet Allocation (LDA), one of the most widely used topic modeling method [2], to assign each document a probability distribution over the topics identified from the entire collection. Each topic is in turn expressed in terms of a word probability distribution. The topic probability distribution associated with each document is considered place semantics for the corresponding venue. For LDA, we set the hyper-parameters α and β to 0.1 and 0.01 respectively, which are commonly used [4], and the number of topics to 50, which is also common in this type of work.

The LDA-based analysis re-discovered salient topics such as 'music event', 'leisure time', 'nightlife', and 'drinking beer' for some venues, which are almost equivalent to the manually labeled categories. On the other hand, it captured hidden semantics. In the case of Pier 59 Studio, for example, the original category is *building*, but the analysis elicited *fashion show*, which is reasonable because Pier 59 is in fact a photo studio where fashion show related activities take place.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576955

4. UTILIZING PLACE SEMANTICS

4.1 Venue Similarity

Since our place semantics are expressed in terms of probability distributions over topics, similarity between two venues can be measured by Jensen-Shannon (JS) divergence [4], most appropriate for computing similarity based on topic distributions.

Fig. 1 shows an example of similarity calculation results. The more the similar two venues are, the darker the color of the box. In our dataset, *John F. Kennedy (JFK) airport* and *LaGuardia (LGA) airport* have the highest similarity.

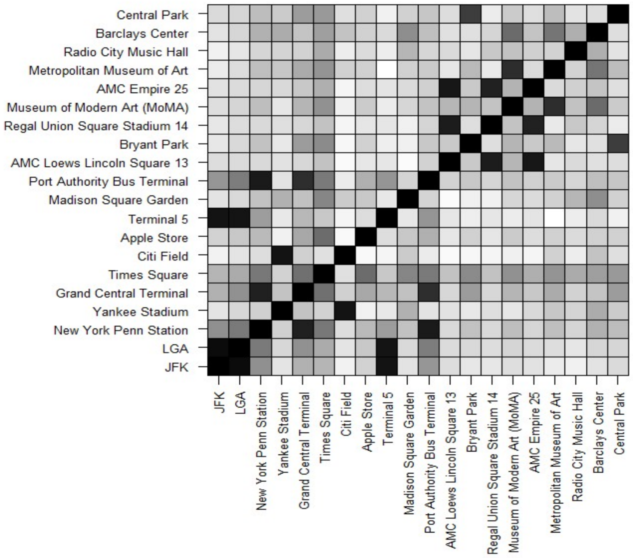

Figure 1. An Example of a Similarity Matrix

While the pair-wise similarity value comparisons reveal functionally similar places, topic-based place semantics can also reveal hidden semantics that are not obvious in the predefined venue categories. For example, *Barclays Center*, a multi-purpose indoor arena in Brooklyn, is categorized as a *basketball stadium* in Foursquare. In our result, however, *Barclays Center* is considered most similar to *Museum of Modern Art* and *Metropolitan Museum of Art*. Given that the arena is actually used for concert, conventions, and exhibitions, it suggests that such analysis can capture the semantics not readily available using only the manually assigned categories.

4.2 Temporal Dynamics

We also investigated how topic-based place semantics changed along the two different time lines: 'seasons of a year' and 'days of a week'. For the seasonal dimension, we split the entire shouts into three time spans corresponding to 'spring' (March to May), 'summer' (June to August), and 'fall' (September to November). Similarly the shouts were also divided into the 'weekday' and 'weekend' buckets based on their posting dates.

The analysis provided different place topic distributions corresponding to the seasonal change as shown in Fig. 2 where x-axis and y-axis represent the IDs for the 50 topics and probabilities. For spring, the top-ranked topics are 'workout', 'back', 'gym', 'day', and 'work'. On the other hand, different top ranked topics emerged for 'summer' and 'fall'. They are 'party', ' show', 'music', 'love' and 'time' for 'summer' and 'brooklyn', ' run', 'walk', 'ride', and 'bridge' for fall

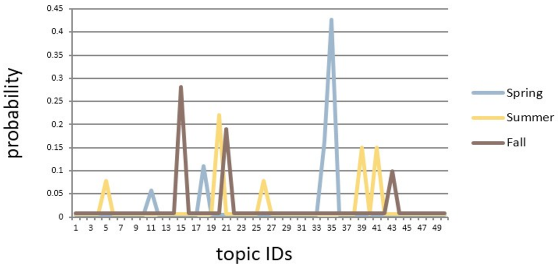

Figure 2. Temporal Dynamics of a Venue

In order to observe the proportion of venues whose topics change over the seasons, we computed JS divergence for all season pairs of each venue. The mean average and the standard deviation are 0.39 and 0.11. We consider places whose average JS divergence is larger than 0.3 as dynamic. A total of 48% of the venues are shown to incorporate temporally dynamic topics.

5. CONCLUSIONS

We proposed a method for capturing topic-based place semantics from the textual descriptions in the Foursquare check-ins. Topic-based place semantics were shown to capture the hidden as well as obvious semantics associated with the venues. As such, it becomes possible to discover semantically or functionally similar venues within and across the categories attached to them. Furthermore, we demonstrated the proposed method makes it possible to recognize the changes in place semantics over time. While this study shows the feasibility of analyzing the textual data in SNS to reveal hidden and temporal semantics associated with venues, it is only the beginning. Our text analyses were quite primitive, just enough to apply a topic modeling method, but we believe more sophisticated methods than just interpreting a set of words can separate activities from objects in the messages more clearly.

6. ACKNOWLEDGMENTS

This research was partially supported by Microsoft Research Asia (MSRA)'s University Relation Project and partially supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2013.

7. REFERENCES

[1] Alves, A., Pereira, F., Rodrigues, F., and Oliveirinha, J. 2010. Place in perspective: extracting online information about points of interest. Ambient Intelligence, 61-72.

[2] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. Latent Dirichlet Allocation, Journal of Machine Learning Research, v.3, 993-1022.

[3] Hong, L., Ahmed, A., Gurumurthy, S., Smola, A. J., and Tsioutsiouliklis, K. 2012. Discovering geographical topics in the twitter stream. In Proceedings of the 21st international conference on World Wide Web, 769-778. ACM.

[4] Kim, D., and Oh, A. 2011. Topic chains for understanding a news corpus. Computational Linguistics and Intelligent Text Processing, 163-176.

[5] Wang, C., Wang, J., Xie, X., and Ma, W. Y. 2007. Mining geographic knowledge using location aware topic model. In Proceedings of the 4th ACM workshop on Geographical information retrieval, 65-70. ACM