POI Type Matching based on Culturally Different Datasets

Li Gong [1, 2] Song Gao[2] Grant McKenzie[2]

1 Geosoft Lab, Institution of Remote Sensing and GIS, Peking University, Beijing, China

2 STKO Lab, Department of Geography, University of California, Santa Barbara, CA, USA

1 Introduction

The development of mobile social media networks has changed our daily life. More and more people tend to share their locations, emotions and activities with their friends, which are called 'check-in records'. These geo-tagged data create an unprecedented opportunity for researchers to reveal spatio-temporal activity patterns of citizens [1,2], capture usages of urban public facilities [3], and understand the interactions between citizens and urban built environment. Points of interest (POI) data play an important role when analyzing check-in records, because they contain categories like restaurant, school to indicate people's activities. Considering the existing various social media data sources, POI type matching is a basic problem if we intent to do research on two or more check-in datasets with different type schemata, which is also a major challenge for POI data conflation in existing map navigation products like *Apple Maps*. Currently, typical approaches POI type matching focuses on comparing types through schema. For example *Schema.org* is often used as an organized vocabulary representing a hierarchy of place types (e.g., a *Restaurant* is a type of *Food Establishment*). Much of the ongoing work in assessing the similarity of POI types happens at a top-down schema level rather than a bottom-up data-driven approach. This work is very much focused on a bottom-up approach to assessing the similarity between POI types in two culturally different datasets adhering to two different schemata. Much existing literature has found that the temporal population variation, named '*temporal signature*', of POIs has strong relationship with the category of POIs [3], but there are few attempts to take the temporal signatures of POIs into consideration when making POI type matching. In this work, we try to match points of interest between culturally (and linguistically) different datasets based on the temporal signatures. We conducted an experiment on two different location-based social networking check-in datasets from different countries. Both linguistic meaning and temporal signatures of POI types were adopted to address the POI type matching problem. We summarized four kinds of POI type matching patterns and measured the similarity between matching POI types, which supports the aforementioned research hypothesis.

2 Data and Methods

Two check-in dataset are used in this experiment: 1) four month timespan check-in records from Foursquare, 2) one-year check-ins from Jiepang. Foursquare and Jiepang are famous LBSN providers in U.S. and China, respectively. These records are also a part of the check-in dataset that has been used in [2] and [3]. We select Shanghai (SH),

China and Los Angeles, U.S. (LA) as two research cities. The venues in Jiepang are categorized into 165 POI types, while the number of POI types in Foursquare is 421. Given 24 hours over 7 days, 16 hourly bands of temporal signature are aggregated by POI type.

In our experiment, POI types in Jiepang are defined as the standard one, and we attempt to find out the candidate POI types from Foursquare. We translate the Chinese POI type name into English firstly, and then take POI types that contain the key words or similar semantics as the candidate POI types. In order to determine the differences between the temporal signatures of object POI type and its candidate POI types, we chose two dissimilarity measures, *Jensen-Shannon Distance* (*JSD*) and *Earth Mover's Distance* (*EMD*) to calculate the variability. The *Jensen-Shannon Distance* is a one-to-one bin comparison approach to calculating the dissimilarity between two distributions [4]. The *Earth Mover's Distance* also calculates the dissimilarity between two distributions but instead of doing a one-to-one bin comparison, each bin is compared to all other bins. Cost and Flow matrices are introduced and a Distance value is returned based on the coast of converting one distribution into the other distribution [5].

3 Four POI Type Matching Patterns

Due to the inconsistency of the numbers of POI types as well as the semantic uncertainty, it's impossible to mapping all POI types one by one. In this section, we summarize four kinds of matching patterns and make an example-based analysis for each pattern.

3.1 Equivalent Alignment

Equivalent Alignment means we can find only one candidate POI type whose name is completely consistent with the object POI type. In this case, we can directly mapping these two POI types, for example, Korean Restaurant (韩国餐厅) and Karaoke Bar (KTV) in Jiepang and Foursquare. The *JSD* of Korean Restaurant and Karaoke Bar are 0.372 and 0.251, respectively. It indicates that the temporal signatures between two dataset are similar.

3.2 Semantic Ambiguity

Equivalent alignment is the most perfect patterns in POI type matching process. As the existence of semantic uncertain, we have to face the dilemma that one object POI type would have several candidates. For example, there is no corresponding POI types in Foursquare which match to Supermarket (超市) in Jiepang, but we can find some results that are similar to Supermarket in the term of semantics, including Mall, Grocery Store and Convenience Store. According to the *EMD* values shown in Table 1, Mall in Foursquare is the best matching result for Supermarket in Jiepang, while Convenience Store seems not a good candidate. We also calculate the dissimilarity for Shopping Mall and Convenience Store in Jiepang, because the semantics of these

three POI types are easy to be confused. We found that Mall and Grocery Store in Foursquare can match the Shopping Mall in Jiepang well.

Table 1 The Normalized Earth Mover's Distance of pairs of shopping related POI types cross Jiepang and Foursquare

	Mall	Grocery Store	Convenience Store
综合商场 (Shopping Mall)	0.150	0.227	0.701
超市 (Supermarket)	0.000	0.181	1.000
便利店 (Convenience Store)	0.851	0.731	0.019

3.3 One-Many Alignment

The third pattern is caused by hierarchical mismatch, which means the object POI type can find a series of its sub-class POI types. The most typical example is Western Restaurant (西餐) in China. As the culture difference, there are a wide variety of restaurants from different counties in U.S., while in China we usually aggregate them into Western Restaurant. We calculated the EMD value between Western Restaurant and each sub-class POI type and the ranked results are shown in Table 2. The EMD represents the relative similarity for all pairs of POI types. We think the POI types less than 0.5 are reasonable candidates.

Table 2 The Normalized Earth Mover's Distance of pairs of Western Restaurants cross Jiepang and Foursquare

POI Type name	EMD
Cuban Restaurant	0.000
New American Restaurant	0.073
Caribbean Restaurant	0.145
French Restaurant	0.218
Middle Eastern Restaurant	0.236
Mexican Restaurant	0.282
Latin American Restaurant	0.282
German Restaurant	0.336
Vietnamese Restaurant	0.364
American Restaurant	0.536
Thai Restaurant	0.564
Indian Restaurant	0.664
Italian Restaurant	0.673
Tapas Restaurant	0.791
Southern / Soul Food Restaurant	0.827
Mediterranean Restaurant	0.964
Greek Restaurant	1.000

3.4 Many-One Alignment

In contrast to One-Many alignment situation, we can define the fourth matching patterns. More than one object POI type belongs to one super-class POI type in Foursquare. Take Chinese Restaurant as an example. Since Chinese local dishes have their own typical characteristics, the restaurants are roughly label as province name according to the origin place of food. According to the EMD values of Chinese Restaurant, the first ten POI types in Jiepang have similar performance with the Chinese Restaurant in Foursquare. Not surprisingly, *Tea Restaurant which is one type of the most popular Chinese* Restaurants across cultures ranks the top by similarity.

Table 3 The Normalized Earth Mover's Distance of pairs of Chinese Restaurants cross Jiepang and Foursquare

POI Type name	EMD
茶餐厅 (Tea Restaurant)	0.000
江浙菜 (Jiangzhe Restaurant)	0.066
四川菜 (Sichuan Restaurant)	0.127
台湾菜 (Taiwan Restaurant)	0.189
家常菜 (Homely Restaurant)	0.193
广东菜 (Guangzhou Restaurant)	0.245
上海菜 (Shanghai Restaurant)	0.316
湖南菜 (Hunan Restaurant)	0.335
新疆/清真 (Xinjiang Restaurant)	0.368
西北菜 (Xibei Restaurant)	0.377
东北菜 (Dongbei Restaurant)	0.594
云南菜 (Yunnan Restaurant)	0.599
北京菜 (Beijing Restaurant)	0.623
澳门菜 (Macao Restaurant)	0.816
贵州菜 (Guizhou Restaurant)	0.863
湖北菜 (Hubei Restaurant)	0.943
山东菜 (Shandong Restaurant)	1.000

4 Conclusions

In this work we have discussed the POI type matching problem between culturally different datasets. Instead of transferring the POI category schema to the golden standard one, we directly mapped POI types purely based on their linguistic meaning and temporal signatures. We also summarize four kinds of matching patterns in terms of semantic uncertainty and hierarchical mismatch. Future work will be focused on realizing automatic matching procedure based on machine leaning algorithms and prior experience.

References

[1] Wu, L., Zhi, Y., Sui, Z., & Liu, Y. (2014). Intra-Urban Human Mobility and Activity Transition: Evidence from Social Media Check-In Data. *PloS one*, *9*(5), e97010.

[2] Liu, Y., Sui, Z., Kang, C., & Gao, Y. (2014). Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. PloS one, 9(1), e86026.

[3] McKenzie, G., Janowicz, K., Gao, S., Yang, J-A., & Hu, Y. POI Pulse: A multi-granual, semantic signatures-based information observatory for the interactive visualization of big geosocial data.Cartographica: The International Journal for Geographic Information and Geovisualization (In Press).

[4] Lin, J. (1991). Divergence measures based on the Shannon entropy. Information Theory, IEEE Transactions on, 37(1), 145-151.

[5] Rubner, Y., Tomasi, C., & Guibas, L. J. (2000). The earth mover's distance as a metric for image retrieval. International Journal of Computer Vision, 40(2), 99-121.