# Semantic Enrichment of Interesting Regions with POI data

Jianan Shen*[1] and Tao Cheng†[1]

[1] SpaceTimeLab, Department of Civil, Environmental and Geomatics Engineering, University College London

January 12, 2017

**Summary**

We propose a quantification algorithm for with categorical description information from POI data, to identify the characteristics and meanings of interesting places in human urban activities. The framework is composed of two major steps, namely, Region of Interest (ROI) detection and semantic enrichment. Interesting places and time periods are detected with density based method in the ROI detection process and semantic enrichment in the second step is based on a spatial coverage and text mining algorithm for the meaning categorization of the ROIs. As a case study, we describe how it can be applied to London police patrol activities to extract the interesting places for police officers. The extraction process is done using TF-IDF based weighting algorithm to rebalance the importance of different categories of POIs. It is particularly dedicated to avoid the meaning of a place from being overwhelmed by large amount of meaningless and unimportant POIs.

**KEYWORDS:** Region of Interest, Semantic Enrichment, POI, DBSCAN, human dynamics

## 1. Introduction to guidelines

With the widespread use of new geospatial technologies, huge amount of human dynamics data have been collected. These technologies include GPSs, sensor networks and mobile phones that track people's movements by recording where and when people move and stop. Space-time activity patterns drawn from these mobility datasets indeed reflect social-demographic characteristics and personal interests of the people. Especially, Location-based service (LBS) has been a popular industry with the widespread the above-mentioned technologies in recent years. Some of the LBS applications, such as Foursquare and twitter, have well penetrated into all aspects in daily life and provided data recording the POI "check-in" and place visiting behaviours of millions of users.

In this article, we apply ST-DBSCAN on the stay points of police movements for ROI detection and define the spatial coverage of ROI with convex hulls. All POIs that fall in the convex hull area are used to interpret the semantic meaning of the Region of Interest. To alleviate the influence of drastic number differences between different categories of POIs, the importance weight of the POIs need to be rebalanced. Therefore TF-IDF, an algorithm originally developed for text mining is introduced to weight the POIs not only considering the number but also taking the spatial distribution of the POIs into account. This method can enrich the description of places and movements with public Point-of-Interest data and enable us to analyse the dynamism in a large city-scale area with high heterogeneity in term of semantic meanings.

## 2. Methodology

This method can be realised by 2 steps as follows:

* jianan.shen.13@ucl.ac.uk

† tao.cheng@ ucl.ac.uk

(1) **ROI Detection:** Extracting Region of Interests (ROIs) of movements of all police officers. This step is identical to the first step in previous work (Shen and Cheng, 2016) but is applied on the lasted version of the police movement dataset which is larger in size and with higher sampling rate;
(2) **Semantic Enrichment of ROIs:** Using and weighting POI data within the convex hull of ROIs to analyse the semantic meaning of the ROIs;

## 2.1. ROI detection and bounding convex hulls of ROI

ST-DBSCAN is firstly used for ROI detection. Figure 1 shows a demonstration of ROI detection results in the borough of Camden, London based upon APLS (Automatic Personnel Location System) dataset provided by the London metropolitan police. As can be seen from the figure, the ST-DBSCAN results are clustered points in space and time and we need a method that allows us to characterise the stay points in an ROI sets through their shapes before explaining the semantic meaning within the area. Minimal convex hull bounding method can create a polygon area enclosing all stay points of each ROI, making it the most ideal way to serve this purpose (Andrew, 1979). We use parallel spatial retrieving method to find the convex hulls that define the spatial boundaries of the ROIs. One example of the generated convex hulls is shown in the right side of figure 1. The semantic analysis in done within the convex hull of every ROI.

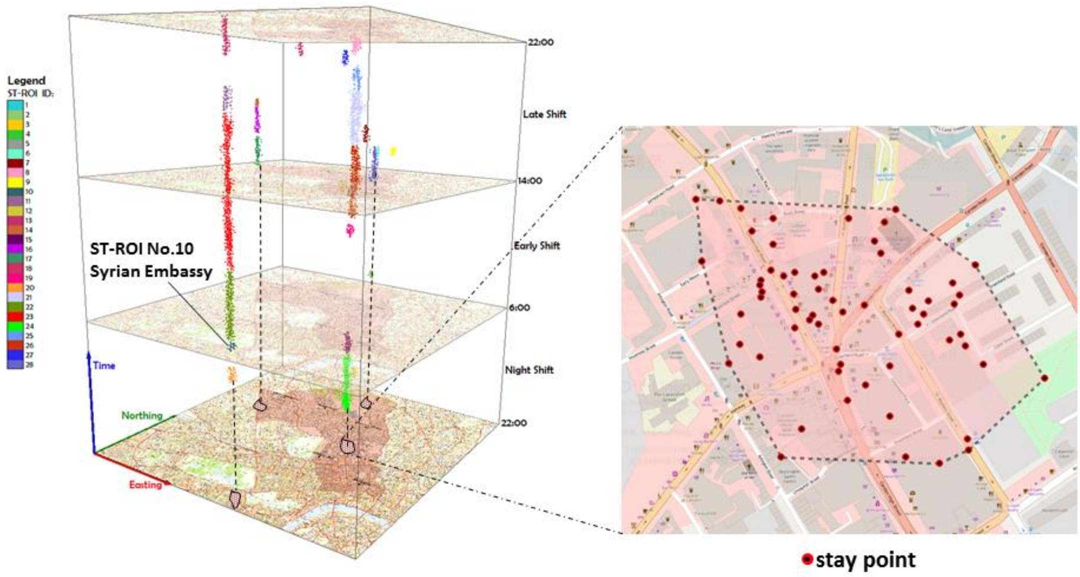


**Figure 1** The 28 ROIs found by ST-DBSCAN method in our previous prototype framework (Shen and Cheng 2016). The bounding convex hull on the right enclosing the stay points is generated to define the spatial area covered by the ROI.

## 2.2. Semantic analysis of ROIs

POI dataset contains the information of all the functional facilities (POIs) that can be summarised to interpret the semantic meaning of a place they are in (Alvares et al., 2007). In order to understand the staying behaviour within each ROI area, we use POI data to depict the functional images of the ROIs and enrich the semantic meaning of users' visits to these ROIs. The Point of Interest (POI) dataset (Ordnance Survey, 2016) contains the information of a wide range of finely categorised infrastructures and buildings that offer different services and utilities.

**Table 1** The reclassified POI categories based on the Ordnance Survey POI classification scheme.

| Customised Classification Scheme | | | |
|---|---|---|---|
| **01 Accommodation, eating and drinking** | | **06 Public infrastructure** | |
| 01 | Accommodation | 34 | Infrastructure and facilities |

| | |
|---|---|
| 02 | Eating and drinking |
| **02 Commercial services** | |
| 03 | Construction services |
| 04 | Consultancies |
| 07 | Contract services |
| 05 | Employment and career agencies |
| 06 | Engineering services |
| 60 | Hire services |
| 08 | IT, advertising, marketing and media services |
| 09 | Legal and financial |
| 10 | Personal, consumer and other services |
| 11 | Property and development services |
| 12 | Recycling services |
| 13 | Repair and servicing |
| 14 | Research and design |
| 15 | Transport, storage and delivery |
| **03 Attractions** | |
| 58 | Bodies of water |
| 16 | Botanical and zoological |
| 17 | Historical and cultural |
| 19 | Landscape features |
| 18 | Recreational |
| 20 | Tourism |
| **04 Sport and entertainment** | |
| 22 | Gambling |
| 23 | Outdoor pursuits |
| 21 | Sport and entertainment support services |
| 24 | Sports complex |
| 25 | Venues, stage and screen |
| **05 Education and health** | |
| 26 | Animal welfare |
| 28 | Health practitioners and establishments |
| 29 | Health support services |
| **07 Manufacturing and production** | |
| 37 | Consumer products |
| 38 | Extractive industries |
| 39 | Farming |
| 40 | Foodstuffs |
| 41 | Industrial features |
| 42 | Industrial products |
| **08 Retail** | |
| 46 | Clothing and accessories |
| 47 | Food, drink and multi item retail |
| 48 | Household, office, leisure and garden |
| 49 | Motoring |
| **09 Transport** | |
| 53 | Air |
| 59 | Bus transport |
| 57 | Public transport, stations and infrastructure |
| 54 | Road and rail |
| 55 | Walking |
| 56 | Water |
| **10 Education** | |
| 27 | Education support services |
| 31 | Primary, secondary and tertiary education |
| 32 | Recreational and vocational education |
| **11 Government and organisations** | |
| 33 | Central and local Government |
| *Annotation:* | |
| **XX Major category code** | |
| xx Sub-category code | |

POIs and buildings of identical functions are likely to aggregate in the same neighbourhood. Based on these phenomena, Polisciuc et al. (2015) used the quantity of POIs in convex hulls to explain the major semantic meaning of places. However, quantities of different categories of POIs vary dramatically in urban space. For instance, the large amount of iconic public telephones and red pillar mailboxes can be found everywhere in London but they have considerably small influence on the function of an area. On the contrary, if there is only one museum in the entire study area, the influence of this museum to the place where it locates should be magnified to outmatch the many telephone boxes. Hence, directly using the quantity of the POIs for semantic enrichment is not enough. The bias caused by unbalanced quantities of different POIs should be subdued. A similar case can be found in text mining researches where article words like “the” and “a” appear far more frequently than the truly meaningful words in most sentences. The importance of a given word in one sentence increases proportionally to the number of times this word appears in the sentence, but is offset by the frequency of the word in the whole context.

Term Frequency–Inverse Document Frequency (TF-IDF) is originally used to measure the importance (weighting factor) of a word in information retrieval and text mining (Salton and Buckley, 1988). Therefore, we borrow TF-IDF to weight the importance of different POIs within the ROIs.
The influence of TF_IDF can be seen in figure 2 shows, 33% of the POIs in ROI No.10 belongs to major POI category No.6 (i.e. public infrastructure). As we know, public infrastructures such as electricity poles and traffic lights can be found literally everywhere in the city. Therefore, after the TF-IDF weighting process, the weight of public infrastructure POIs is significantly lowered from 0.333 to 0.043.

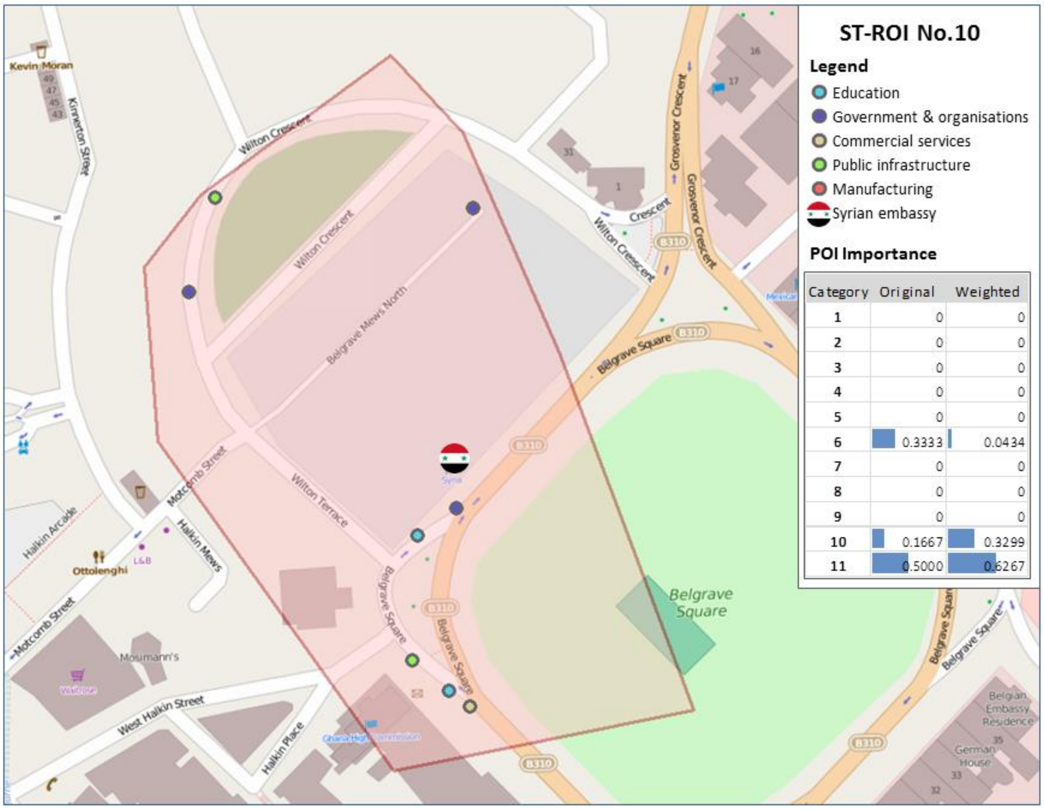


**Figure 2** Summarising the POIs locating in the bounding convex hull of ROI No.10. The POIs' categories are represented with different colours and the difference the weighting process made on the 11 categories in show on the right.

Table 2 shows the semantic weights of POIs in the 5 ROIs in London after TF-IDF weighting. The name of the ROIs are also labelled. These weights reflected the semantic meaning of the place.

**Table 2** The TF-IDF semantic weights and names of 5 chosen ROI.

| ST-ROI ID | 2 | 6 | 7 | 27 | 34 |
|---|---|---|---|---|---|
| Name of the place | **Backingham Palace** | **Soho** | **Trafalgar Square** | **Camden Station** | **White Hall** |
| **Accommondation, eating, drinking** | 0.0000 | 0.3233 | 0.0642 | 0.1041 | 0.1375 |
| **Commercial services** | 0.0487 | 0.1676 | 0.1910 | 0.2370 | 0.1818 |
| **Attractions** | 0.7438 | 0.0375 | 0.2413 | 0.0119 | 0.0156 |
| **Sport & entertainment** | 0.0000 | 0.1339 | 0.0000 | 0.1047 | 0.0288 |
| **Health** | 0.0000 | 0.0411 | 0.0000 | 0.0812 | 0.0000 |
| **Public infrastructure** | 0.0233 | 0.0387 | 0.0644 | 0.0236 | 0.0268 |
| **Manufacture & production** | 0.0000 | 0.0426 | 0.0203 | 0.0267 | 0.0000 |
| **Retail** | 0.0523 | 0.1412 | 0.0419 | 0.3027 | 0.0000 |
| **Transport** | 0.0000 | 0.0332 | 0.2551 | 0.0375 | 0.0357 |
| **Education** | 0.0000 | 0.0296 | 0.0300 | 0.0534 | 0.0000 |
| **Government & organisations** | 0.1319 | 0.0114 | 0.0918 | 0.0170 | 0.5738 |

## 3. Conclusion

It is clear that, in order to improve current and future location based services, more information must be associated to the interesting regions detected in human movements. Location representation needs more semantic information explaining the meaning of the places and the purpose of visit. Our work presented a methodology to identify places characteristics from the human movement trajectories and POI data sources. Our evaluation has shown the feasibility and validity of the methodology beside the quality of the knowledge resource used in the implementation.

## References

Andrew A (1979). Another efficient algorithm for convex hulls in two dimensions. *Information*

*Processing Letters*, 9(5), 216–219.
Alvares O, et al. (2007). A model for enriching trajectories with semantic geographical information. *Proceedings of the 15th annual ACM international symposium on advances in geographic information systems*.
Ordnance Survey, 2016. Points of Interest [online]. Available from: https://www.ordnancesurvey.co.uk/business-and-government/products/points-of-interest.html
Salton, G. and Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5), 513–523.
Shen, J. and Cheng, T., 2016. A framework for identifying activity groups from individual space-time profiles. *International journal of geographical information science*. 30(9), 1785-1805.