Mining Points of Interests with Popular Travel Patterns and Spatial Guidance from Social and Credible Sources

Erum Haris and Gan Keng Hoon
{he15_com037@student.usm.my,khgan@usm.my}

School of Computer Science, Universiti Sains Malaysia,
11800 Penang, Malaysia

Abstract. This work aims to utilize social and credible tourism content in order to develop a gazetteer of worth seeing points of interest (POIs) in a region along with mining most popular visit patterns of these POIs followed by experienced travelers. It proposes a new insight to frequent travel pattern mining by enriching these routes with spatial relations among the POIs to facilitate navigation.

Keywords: trip planning, frequent pattern mining, spatial relation extraction

1 Introduction

Tourism weblogs are a good source for mining travel patterns as they are written in a diary-style, so the order of appearance of place names in the text can be interpreted as blogger's actual travel pattern [1]. Furthermore, they usually provide hints regarding spatial connectivity between these places. These spatial relations can be topological (across, next to etc.), directional (north, in front, behind etc.) or distance (nearby, far etc.). Consider the snippet below from a travelogue with highlighted travel pattern and spatial information:

KLCC is the shiny new centre of Kuala Lumpur best known for its iconic Petronas Towers….The nearby Syakirin Mosque is worth a visit, the shopping and food courts of Suria KLCC mall, the surrounding KLCC Park…

1.1 Research Questions and Contribution

Mining popular travel routes from blogs has been touched by [1] where a multimedia tour guide is developed based on the text, images and video content posted on blogs to generate travel routes. Ref [2] extracted hot tourism locations and travel patterns using frequent pattern mining and location correlation analysis whereas [3] used structured tourism blogs for identifying frequent spots using compact pattern mining. Now, the questions of concern are: which piece of textual content can be useful apart from popular travel routes to facilitate novice travelers in choice of next destination along the way? Secondly, compared to mathematical solutions for trip planning that target to maximize objectives such as distance, timings etc. [4], what parameters one can target to learn from human behaviour by analyzing travel blogs? This

work proposes the extraction of spatial relations between POIs thereby generating knowledgeable routes.

2 Proposed Methodology

Problem Statement: 1) Given a dataset $\mathbb{B}$ of n blog entries, with pre-processing and gazetteer matching, transformed to a set of vectors X' = $\{x_{1'}, x_{2'}, \ldots, x_{n'}\}$ where vector xi' = $\{xi_1, \ldots, xi_n\}$ and a defined minimum support threshold (s_min), find all *ordered sequences* with *n* items $\{xi_j, xi_{(j+1)} \ldots xi_n\}$ having relative support no less than s_min.
2) Given a spatial relations annotated corpus $\mathbb{C}$ obtained by pre-processing dataset B, extract instances of spatial triplet (POI_i, SR_k, POI_j) where POI_i defines the *trajector* which means the POI whose location is to be described with reference to POI_j which defines the *landmark or relatum*. SR_k denotes the *spatial relation* between POI_i and POI_j.

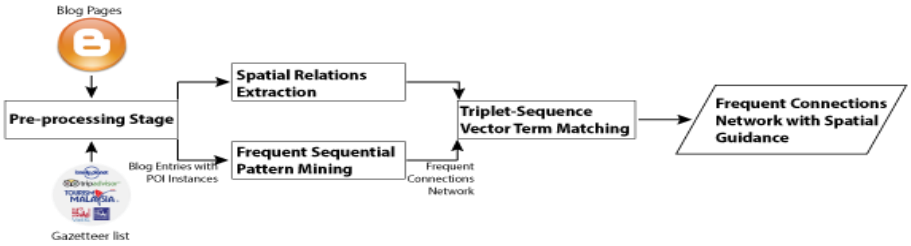

Fig. 1. Proposed Methodology

Description: Figure 1 shows the flow of process. Popular POIs are accumulated first from credible travel websites. This results in construction of a gazetteer list GL that will be used for pre-processing of blog entries to eliminate non-geographic terms while retaining the favorite POIs' mentions. Referring to the problem statement, Frequent 1-itemset $F^1(X')$ is mined from the set of blog entries' vectors X', with support count higher than s_min. $F^1(X')$ will then be used to generate candidate frequent *n*-itemsets. To illustrate the scenario, let *GL = {KLCC, Petronas Towers, Syakirin Mosque, Suria KLCC mall, KLCC park}*. Table 1 depicts the vectors obtained by applying GL based filtering to five blog entries. With s_min=60%, we obtain $F^1(X')$ = {*KLCC, Petronas Towers, Syakirin Mosque, Suria KLCC mall, KLCC park*}. Setting s_min = 40% and *n=2*, we get sequence of popular POIs. See Table 2 and Figure 2.

Table 1. Pre-processed Blog Entries

Vector	POIs Sequence
$x_{1'}$	{KLCC, Petronas Towers, Syakirin Mosque, Suria KLCC mall, KLCC park}
$x_{2'}$	{ KLCC, Petronas Towers, Suria KLCC mall, Syakirin Mosque, KLCC park}
$x_{3'}$	{ Syakirin Mosque, KLCC, Petronas Towers, Suria KLCC mall}
$x_{4'}$	{ Petronas Towers, Syakirin Mosque, KLCC, KLCC park}
$x_{5'}$	{ KLCC park, Syakirin Mosque, Petronas Towers, Suria KLCC mall, KLCC}

For spatial relation extraction, a set of syntactical rules is developed based on geo named entity recognition and a corpus annotated with spatial relations. Recalling second part of problem statement, the *trajector – landmark* (T-L) pairs in the extracted triplets are now matched with the

sequential POIs pairs to associate them with corresponding spatial relation. The order of target and reference objects in spatial relation may conflict with the sequential travel order of POIs. So, while matching, this order need not to be distinguished in case of some topological or distance relation in a T-L pair. For example, a triplet extracted like *(Batu Caves, 13 Kilometers, KL)* is to be matched with a sequential pattern *{KL, Batu Caves}*. However, a triplet like *{Batu Caves, North, KL}* indicating a direction relation, while matching with a sequential pattern, T-L order is to be retained.

Table 2. Popular Travel Sequences

POIs Sequence (n=2)	Correlation Weight / Route Popularity
{KLCC, Petronas Towers}	60%
{Petronas Towers, Syakirin Mosque}	40%
{Syakirin Mosque, KLCC}	40%
{Petronas Towers, Suria KLCC mall}	40%

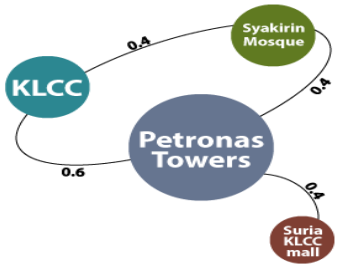

Fig. 2. A frequent connections network

3 Preliminary Results and Future Work

To this point, gazetteer has been constructed from *TripAdvisor* and *LonelyPlanet* containing POIs related to Kuala Lumpur. The system managed to extract 80 unique POIs and their attributes from *TripAdvisor* and 50 records from *LonelyPlanet*. For future work, the results need to be integrated along with the extracted POIs from the official state tourism website *Tourism Malaysia* [5]. Also, assessment of pattern mining and triplet matching is remained to be carried out on crawled blog pages.

Acknowledgments. This research was supported by USM Research University Grant (1001/PKOMP/811335: Mining Unstructured Web Data for Tour Itineraries Construction). The authors would also like to thank Tourism Malaysia.

References

[1] Kori, H., Hattori, S., Tezuka, T. and Tanaka, K.: Automatic Generation of Multimedia Tour Guide from Local Blogs. In: 13th International Multimedia Modeling Conference (2007)
[2] Xu, H., Yuan, H., Ma, B. and Qian, Y.: Where to go and what to play: Towards summarizing popular information from massive tourism blogs. Journal of Information Sciences. 41(6), (2015)
[3] Guo, L., Li, Z. and Sun, W.: Understanding Travel Destinations from Structured Tourism Blogs, Wuhan International Conference on e-Business (2015)
[4] Yahi, A., Chassang, A., Raynaud, L., Duthil, H. and Chau, D.H.: Aurigo: An Interactive Tour Planner for Personalized Itineraries. In: Proceedings of the 20th International Conference on Intelligent User Interfaces (2015)
[5] Tourism Malaysia, http://www.malaysia.travel/en/my