Influence maximization problem to find the influential location

Sun Yehong
515030910593
syhroy7@gmail.com

Abstract

Influence maximization problem is a hot research direction in social network.network.Influence maximization in social networks is a classic and extensively studied problem that targets at selecting a set of initial seed nodes to spread the influence as widely as possible. In this report, I try to create a model of influence maximization problem to find some influential places. Mining the most influential k-location set finds k locations, traversed by the maximum number of unique trajectories, in a given spatial region. These influential locations are valuable for resource allocation applications, such as selecting charging stations for electric automobiles and suggesting locations for placing billboards.We usually solve this problem in which we choose the most frequent place. In this paper, we propose a new method to find the most influential points set in real places. The greedy heuristic is efficient with performance guarantee. We evaluate the performance of our proposed method based on a taxi dataset of Shanghai.

1 Introduction

1.1 Influence maximization in social network

A social network is the network of relationships and interactions among social entities such as individuals, groups of individuals, and organizations.With the rise of the Internet and the World Wide Web developing, We able to investigate large-scale social networks. There has been growing interest in social network analysis.

Information can propagate from one node to another node through a link on a social network, which makes it an important research issue to find influential nodes for the spread of information through a network represented by a directed graph. These combinatorial optimization problem was called the influence maximization problem. This is the problem of extracting a set of k nodes to target for initial activation such that it yields the largest expected spread of information for a given integer k. The two widely-used fundamental information diffusion models is the *independent cascade (IC) model* and the *linear threshold (LT) model.* And in this project, we use *IC model* to test the algorithm.

1.2 Model the Influence maximization problem in real place

Advances in location acquisition technology have resulted in massive trajectories, representing the mobility of a diversity of moving objects, e.g., human, vehicles, and animals.GPS-enabled devices, like GPS-phones, are changing the way people interact with the Web by using locations as contexts. With such a device, a user is able to acquire present locations, search the information around them and design driving routes to a destination.Finding k most influential locations using the GPS trajectories of taxi in a given spatial region is vital to many resource allocation problems:

Algorithm 1 Framework of Greedy Heuristics

Input: Vertex-trajectory index $\mathcal{I}_{vt}$, spatial index $\mathcal{I}_{spatial}$, spatial range R, and k value.
Output: k vertices V_{gdy}
1: $V_s := SpatialRangeSearch(\mathcal{I}_{spatial}, R)$
2: **for** i = 1 to k **do**
3: $V_{gdy} \leftarrow V_{gdy} \bigcup v_{cur}, v_{cur} \in V_s \setminus V_{gdy}$ with max coverage.
4: Update the coverage values in the *vertex coverage table*.
5: return V_{gdy}

Figure 1: The greedy hill-climbing algorithm

The first application is selecting charging stations for electric vehicles according to their GPS trajectories. We usually choose the locations based on the number of occurrences. But from the perspective of information propagation, I don't think it's a wise choice.Someone put forward a method called "MAX-k-COVER". They use the trajectories to solve a k-cover problem[2][3]. This method also ignore the diffusion of the information.So I tried to create a influence maximization model to solve these questions. And we can extract the start locations and end locations of the trajectories from our dataset, which can be treat as the points set V. The edge set E can be get from the relations of start locations and end locations. And in this project, I will create an *independent cascade (IC) model* using the dataset of taxi trajectories in Shanghai.Our main contributions are summarized as follows:

1. I introduce a novel problem, i.e., mining the most influential k-location set, with many potential applications.
2. I solve the problem in a new perspective. I create a model of *independent cascade (IC)* and treat it as a question of information propagation.
3. Evaluation results on real datasets demonstrate the efficiency of our proposed solution.

2 Overview

2.1 Problem Definition

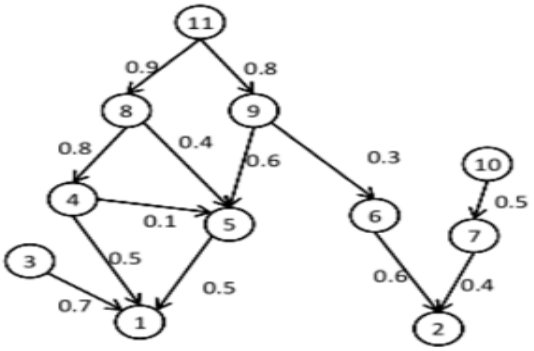
Given a user-specified spatial region R, a k value and a set of trajectories Tr, we denote the spatial network in R as G = (V,E). The most influential k-location set in R finds k locations in Vs, such that the total number of locations being activated by the k locations is maximized.

2.2 Independent Cascade Model

We define the *IC* model. In this model, we must specify a real value pu,v $\in$ [0, 1] for each directed link (u, v) in advance. Here, p(u,v) is referred to as the propagation probability through link (u, v). When an initial set A of active nodes is given, the diffusion process proceeds in the following way. When node u first becomes active at step t, it is given a single chance to activate each currently inactive child v, and succeeds with probability p(u,v). If u succeeds, then v will become active at step t+1. If multiple parents of v first become active at step t, then their activation attempts are sequenced in an arbitrary order, but performed at step t. Whether or not u succeeds, it cannot make any further attempts to activate v in subsequent rounds. The process terminates if no more activations are possible.

2.3 Greedy hill-climbing algorithm

To approximately solve this optimization problem, I consider the following greedy hill-climbing algorithm[1]:



h

Figure 2: A simple IC model

Figure 3: The mapping of the trajectories

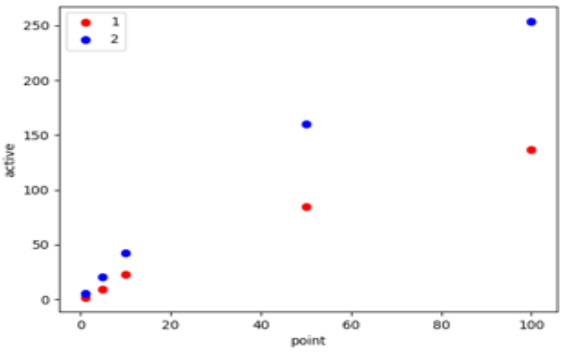
2.4 create the model

The dataset we used is a dataset of taxi trajectories in Shanghai.And I extracted the start locations and end locations of the trajectories from our dataset, which can be treat as the points set V. The edge set E can be get from the relations of start locations and end locations. The weights are calculated by the graph model. With this model, I test the efficiency of my algorithm.The result is shown in the figure 2.

3 Design And Implementation

3.1 Pre-process

This step contains two tasks:
1)trajectory mapping, I maps the raw trajectories onto the corresponding spatial network, e.g., using a map matching algorithm.And the result is shown on the figure 3:
2)spatial network construction, The initial points set has too many points which are not representative then I clustered the locations based on the trajectories, and then constructs the spatial network;Graph partitioning is used to divide a graph into several chunks while satisfying certain constraints and objectives .The most common constraint is to produce partitions having similar chunk sizes, while the most common objective is to minimize the number of edges between the divided chunks.And we finally make the points set simpler and more representative.



h

Figure 4: The result of my experiment

3.2 Experiment

After my pre-process, the points set has been reduced to 274 points. With the points set V, edges set E and the weights set W. I create the *IC* model. And in the experiment, I compared the two algorithm to verify the accuracy of my opinion. The first algorithm is the number of occurrences first. Another algorithm is the greedy hill-climbing algorithm. I test these algorithm with the initial graph and the initial number of points k. And we choose different k 1,5,10,50,100. And the result is in the following:

3.3 Result analysis

We can see that my algorithm is much better than the first algorithm. In the real life, this will be useful for the resource allocation and the chain store layout. So in my opinion, this work is useful and the result is pretty good.

4 Future work

I get a pretty good result,but the complexity of this algorithm is O(knRm).It is too complex. And the clustering is not the best.So I will make my project better in the following aspects:

1. There are many new model in the influence maximization problem, the complexity of some model algorithm even can be O(m) called Degree Heuristic.
2. I want to improve the clustering algorithm and add the influence of the POI[2].

5 Conclusion

This work presents a model on mining most influential k-location set over massive trajectory data. It has many potential applications in resource allocation applications. I use the greedy heuristic algorithm to support interactive queries. Extensive experiments on real datasets demonstrate that our proposed solution is efficient.

Thanks to TA and Teacher Fu for her help on my project and suggestion!

References

[1] Masahiro Kimura,Kazumi Saito*Extracting Influential Nodes for Information Diffusion on a Social Network* www.aaai.org

[2] Yuhong Li,Jie Bao,Yanhua Li,Yingcai Wu,Zhiguo Gong, Yu Zheng *Mining the Most Influential k-Location Set From Massive Trajectories.* http://dx.doi.org/10.1145/2996913.2997009
[3] Yu Zheng, Lizhu Zhang, Xing Xie, Wei-Ying Ma*Mining Interesting Locations and Travel Sequences from GPS Trajectories.*