2017 IEEE 33rd International Conference on Data Engineering

Influence Maximization in Trajectory Databases

(Extended Abstract)

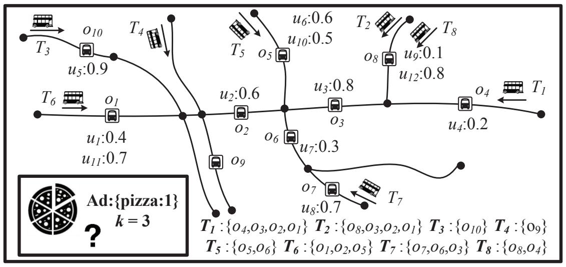
Long Guo $^{\#1}$, Dongxiang Zhang $^{\dagger 2}$, Gao Cong ^{*3}, Wei Wu $^{\S 4}$, Kian-Lee Tan $^{\ddagger 5}$

$^{\#}$School of EECS & Key Laboratory of High Confidence Software Technologies (MOE), Peking University
†University of Electronic Science and Technology of China
*Nanyang Technological University §Visa Inc, Singapore ‡National University of Singapore
^{1}guolong@pku.edu.cn, ^{2}zhangdo@uestc.edu.cn, ^{3}gaocong@ntu.edu.sg, ^{4}hiwuwei@gmail.com, ^{5}tankl@comp.nus.edu.sg

***Abstract*—We study a novel problem of influence maximization in trajectory databases that is very useful in precise location-aware advertising. It finds k best trajectories to be attached with a given advertisement and maximizes the expected influence among a large group of audience. We show that the problem is NP-hard and propose both exact and approximate solutions to find the best set of trajectories. We also extend our problem to support the scenario when there are a group of advertisements. We validate our approach via extensive experiments with real datasets.**

I. Introduction

Influence maximization in a social network is a key algorithmic problem behind online viral marketing. By word-of-mouth propagation effect among friends, it finds a set of k seeds to maximize the expected influence among all the users. It has attracted significant attention from both academic and industry communities due to its potential commercial value [1], [2], [3], [4], [5].



u_1: {NBA 0.8, pizza 0.3}	u_2: {pizza 0.9, NBA 0.4}	u_3: {pizza 0.6, NBA 0.8}
u_4: {pizza 0.5, swim 0.5}	u_5: {pizza 0.9, gym 0.2}	u_6: {pizza 0.3, swim 0.9}
u_7: {pizza 0.9, gym 0.1}	u_8: {pizza 0.5, gym 0.5}	u_9: {pizza 0.1, NBA 0.9}
u_{10}: {NBA 0.9, gym 0.1}	u_{11}: {gym 0.9, swim 0.4}	u_{12}:{gym 0.9, swim 0.4}

Fig. 1. A working scenario

In paper [6], we make the first attempt to transplant the concept of influence maximization from social-aware advertising to location-aware advertising. To facilitate a better comprehension of our new problem, we start with a toy example depicted in Figure 1. Each user or audience u_i in this scenario is associated with an interest profile as well as motion patterns. For instance, we know that u_2 likes pizza and will wait at bus stop o_2 daily with probability 0.6 in a certain time period. On the other hand, we are aware of the trajectory information of all the buses from their schedules. We say an audience is influenced by a bus if they occur at a bus stop at the same time and the interest profile matches the ad. Given an advertisement for "pizza", our goal is to find top-k buses carrying this ad and generating the maximum influence among a huge audience group.

We formulate the trajectory influence maximization problem and show that the problem is NP-hard. Thereafter, we propose both exact and approximate solutions to find the best set of trajectories. In the exact solution, we devise an expansion-based framework that enumerates trajectory combinations in a best-first manner and propose three types of upper bound estimation techniques to facilitate early termination. In addition, we propose a novel trajectory index to reduce the influence calculation cost. To support large k, we propose a greedy solution with an approximation ratio of (1-1/e), whose performance is further optimized by a new proposed cluster-based method. We also propose a threshold method that can support any approximation ratio $\epsilon \in (0, 1]$. In addition, we extend our problem to support the scenario when there are a group of advertisements.

II. Problem Definition

Data Model. In our data model, three types of roles are involved: 1) *Trajectory.* A trajectory is generated by a vehicle which serves as the carrier for an advertisement. We preprocess a trajectory and represent a trajectory by a sequence of POIs along the roads. Let T denote a trajectory and $T = \{\mathtt{o}_i : (o_i, t_i) | 0 \leq i \leq |T|\}$, where o_i is a POI and t_i is the time period when the trajectory passes the POI. 2) *Audience.* An audience is modeled as a text profile associated with spatial-temporal patterns. It captures a user's preference and the likelihood to visit a region. We formally represent each audience as $u = \{\mathtt{t}, \mathtt{m}\}$, where $\mathtt{t}$ is a set of weighted tags and $\mathtt{m} = \{(o_i, t_j, p_{i,j}) | 0 \leq i \leq m \wedge 0 \leq j \leq n\}$. Here o_i is a POI, t_i is a period and $p_{i,j} \in (0, 1]$ refers to the probability that u visits o_i during t_j. 3) *Advertisement.* An advertisement q is represented by a set of weighted tags.

Trajectory Influence Given an advertisement q, we can define the influence score between an audience u and a trajectory T carrying q as follows:

$$\mathcal{I}(q, u, T) = \sigma(q, u) \cdot \rho(u, T). \tag{1}$$

Here $\sigma(q, u)$ measures the textual relevance between q and u which can be captured by any information retrieval model, and $\rho(u, T)$ measures the influence probability that u will "meet" a vehicle that carries q and can be defined as

$$\rho(u, T) = 1 - \prod_{j \in M(u,T)} (1 - p_j) \tag{2}$$

where $\prod_{j \in M(u,T)} (1 - p_j)$ measures the probability that an audience will not "meet" a vehicle running on T. Such an influence score captures the textual relevance, spatial relevance and temporal relevance between u and a vehicle attached with q moving along T.

Similar to Eqn. 1, we measure the influence between an audience u and a set of trajectories $S = \{T_1, T_2, \ldots, T_k\}$

2375-026X/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDE.2017.20

carrying the same advertisement q as follows:

$$\begin{aligned}\mathcal{I}(q,u,S) = &\quad \sigma(q,u)\cdot\rho(u,S)\\ = &\quad \sigma(q,u)\cdot(1-\prod_{j\in M(u,T_1)\bigcup\ldots\bigcup M(u,T_k)}(1-p_j))\end{aligned} \tag{3}$$

Again, $\prod_{j\in M(u,T_1)\bigcup\ldots\bigcup M(u,T_k)}(1-p_j)$ measures the probability that u will not be influenced by any of the k trajectories.

Let U denote a group of audience. We define the expected influence between a group of audience U and a set of trajectories S with advertisement q:

$$\mathcal{I}(q,U,S) = \sum_{u\in U}\mathcal{I}(q,u,S). \tag{4}$$

Problem Definition We formulate the influence maximization problem in a trajectory database as follows.

Definition 1 (Trajectory Influence Maximization): Given a trajectory database $\mathcal{T}$ and a group of audience U attached with profiles and spatial-temporal patterns, for an advertisement q, our goal is to find a trajectory set S where $S\subset\mathcal{T}$ and $|S|=k$ such that the expected influence $\mathcal{I}(q,U,S)$ is maximized.

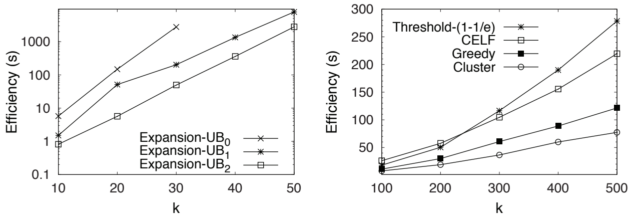
We can reduce the Set Cover problem to the trajectory influence maximization problem and thus prove it to be NP-hard. We propose both exact and approximate methods to solve the trajectory influence maximization problem efficiently.

III. Methods

Exact Methods. To find the exact top-k trajectories, we propose an expansion-based framework that enumerates the trajectory combinations in a best-first manner. The algorithm starts by calculating the influence score of each trajectory w.r.t. to the advertisement. The trajectories are then sorted by the influence and accessed accordingly. In each iteration, combinations with the new trajectory are enumerated. If a combination contains fewer than k trajectories, it is considered *incomplete* and we estimate its upper bound influence from the unvisited trajectories. If a combination is *complete*, we calculate its exact influence. The algorithm terminates when the upper bound influence score of all the incomplete combinations are smaller than the best result ever found. To accelerate the calculation of the influence score, we propose a novel trajectory index which pre-computes a portion of the influence. To facilitate early termination of the algorithm, we propose three types of upper bound influence score, denoted as UB_0, UB_1 and UB_2. UB_0 is computationally efficient but is too loose to facilitate pruning and UB_1 is tight but incurs expensive computation cost, while UB_2 achieves a better tradeoff between efficiency and effectiveness in pruning.

Approximate Methods. However, the expansion-based exact method is not scalable when k is large. To address the issue, we propose a greedy algorithm which finds the trajectory with the maximum incremental influence at each iteration until k trajectories are found and achieves a $(1-1/e)$ approximation ratio. It avoids examining all the unvisited trajectories, which is achieved by scanning the trajectories in order and terminating as early as possible, and avoids calculating the expensive incremental influence for each examined trajectory, which is achieved by utilizing the estimated incremental influence. To further improve its efficiency, we propose a cluster-based algorithm that guarantees the same approximation ratio. It partitions the trajectory database into clusters and allows us to access the clusters in an order such that promising trajectories will be found earlier. Our third approximate solution, named threshold-based method, provides a flexible means to adjust the tradeoff between efficiency and accuracy. It guarantees a ϵ approximation ratio for any $\epsilon\in(0,1]$. In addition, we propose a group greedy method to support the influence maximization for a group of advertisements, which selects the trajectories by considering all the advertisements simultaneously and can guarantee a $(1-1/e)$ approximation ratio.

IV. Experimental Evaluation



(a) Exact methods

(b) Approximate methods

Fig. 2. Evaulation on real trajectory dataset.

Evaluation on Exact Methods. For the exact methods, we compared three variants of the expansion-based framework with different upper bound estimation techniques. Fig 2(a) shows the impact of k, i.e., the number of selected trajectories. As shown, `Expansion-UB`$_2$ achieves the best performance. It takes only about 5 seconds to find the top-20 trajectories. This is because it can achieve a better tradeoff between efficiency and effectiveness in pruning.

Evaluation on Approximate Methods. Fig 2(b) present the impact of k on approximate methods. As shown, `Greedy` and `Cluster` scales well w.r.t. k compared with the exact methods. It takes about 20 seconds to find the top-100 trajectories. In addition, `Cluster` achieves better performance than `Greedy`. This is attributed to its preknowledge about the promising trajectory.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China under Grant No. 61572039 and Grant No. 61602087, 973 program under No. 2014CB340405, Shenzhen Gov Research Project JCYJ20151014093505032 and the NExT Search Centre (grant R-252-300-001-490).

References

[1] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, “A data-based approach to social influence maximization,” *Proc. VLDB Endow.*, 2011.

[2] Q. Jiang, G. Song, G. Cong, Y. Wang, W. Si, and K. Xie, “Simulated annealing based influence maximization in social networks,” in *AAAI*, 2011.

[3] Y. Li, D. Zhang, and K.-L. Tan, “Real-time targeted influence maximization for online advertisements,” *Proc. VLDB Endow.*, 2015.

[4] P. Domingos and M. Richardson, “Mining the network value of customers,” in *KDD*, 2001.

[5] D. Kempe, J. Kleinberg, and E. Tardos, “Maximizing the spread of influence through a social network,” in *KDD*, 2003.

[6] L. Guo, D. Zhang, W. Wu, G. Cong, and K. L. Tan, “Influence maximization in trajectory databases,” *IEEE Transactions on Knowledge and Data Engineering*, 2016.