Chen W, Wang W, Yin H et al. User account linkage across multiple platforms with location data. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 35(4): 751–768 July 2020. DOI 10.1007/s11390-020-0250-7

User Account Linkage across Multiple Platforms with Location Data

Wei Chen, Weiqing Wang, Hongzhi Yin,
Junhua Fang, Lei Zhao

Research Objectives

- Our research aim to link accounts belonging to the same user across multiple platforms with high efficiency, effectiveness, and scalability, by designing a search space pruning method to reduce the computational complexity, and similarity measuring method combining spatial and temporal information.

Research Contributions

- We analyze the complexity of the study from a theory perspective, propose the idea “reverse pruning” to reduce computation complexity and ensure the high scalability of the proposed model.
- We develop a novel method incorporating KL Divergence and kernel density estimation to calculate the similarity between user accounts with the goal of linking them more precisely.

Research Problem

- Given m sets of user accounts from m different platforms with $U_1 = \{u_1^1, u_2^1, \cdots, u_1^{n_1}\}, \cdots, U_m = \left\{u_m^1, u_m^2, \cdots, u_m^{n_m}\right\}$ where n_m denotes the number of user accounts in the m-th platform, our goal is to identify all account match $\varsigma\{u_{(m)}\}$ of the same user from $\left\{\left(u_1^i, u_2^i, \cdots, u_m^i\right)\middle|u_1^i \in U_1, \cdots, u_m^i \in U_m\right\}$.

Research Method

- We propose a novel method GTkNN, which is used for pruning search space by efficiently retrieving top-k candidate user accounts indexed by well-designed spatial and temporal index structures. In addition, to precisely measure the similarity between different user accounts, a kernel density estimation based method is presented, where both spatial and temporal information are considered.

Research Results

- User account linkage across multiple platforms has been achieved with high efficiency, effectiveness, and scalability on four-real world datasets. The average time cost of our proposed model on datasets WBC, FQT, ITW, and Gowalla is 4.57s, 1.28s, 0.95s, and 1.52s, respectively. The precision, recall on these datasets is (0.38,0.39), (0.42,0.4), (0.65,0.61), and (0.91,0.905), respectively.

Research Conclusions

- By conducting extensive experiments on four real-world datasets, the results demonstrate the high performance of the proposed model ULMP. The proposed method GTkNN has reduced the complexity, and the kernel density estimation based method has precisely measured the similarity between different user accounts. Following the user account linkage, we can obtain abundant data from different sources, and this data can be used in cross-domain recommendation, cross-domain user behaviors prediction, and so on.