User Identification in Cyber-Physical Space: a Case Study on Mobile Query Logs and Trajectories

Tianyi Hao[§†], Jingbo Zhou[§], Yunsheng Cheng[§], Longbo Huang[†], Haishan Wu[§*]
[§]Baidu Research, Big Data Lab
[†]IIIS, Tsinghua University
{haotianyi,zhoujingbo,chengyunsheng01,wuhaishan}@baidu.com
longbohuang@tsinghua.edu.cn

ABSTRACT

User identification across domains draws lots of research effort in recent years. Although most of existing works focus on user identification in a single space, in this paper, we first try to identify users by fusing their activities in cyber space and physical space, which helps us obtain a comprehensive understanding about users' online behaviours as well as offline visitation. Out profound insight to tackle this problem is that we can build a connection between the cyber space and the physical space with the stable location distribution of IP addresses. Thus, we propose a novel framework for user identification in cyber-physical space, which consists of three key steps: 1) modeling the location distribution of each IP address; 2) computing the co-occurrence with an inverted index to reduce the space and time cost; and 3) a learning-to-rank tactic to fuse user's features shared in both spaces to improve the accuracy. We conduct experiments to identify individual users from mobile query logs (generated in cyber space) and trajectory data (generated in physical space) to demonstrate the efficiency and effectiveness of our framework.

Categories and Subject Descriptors

H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Data mining; Spatial databases and GIS*

Keywords

User identification; spatial data mining; cyber-physical space; spatial index.

1. INTRODUCTION

Obtaining a deep and comprehensive understanding of each individual user from the big data is an intriguing problem which brings benefits to users and service providers. The problem has been studied in two different spaces in recent years, including user identification in cyber space [19, 10, 18, 12, 9, 23] as well as that from heterogeneous trajectory generated in physical space [22, 16, 2].

In this paper, we investigate another challenging problem to identify a user from heterogeneous data generated in cyber and physical spaces to obtain an enriched understanding of the user, which has a wide range of applications. For example, authors in [14] demonstrate that their recommendation model with fusing the online and offline data can increase the customer purchase ratio significantly.

Especially, in this paper, we make a case study on linking anonymous users in mobile query logs and trajectories. For mobile queries, the users may search something with a mobile browser without login and we can only get a cookie ID. For trajectories from the mobile applications such as Google Maps and Baidu Maps, we can get a device ID for each trajectory. The aim is to link the cookie ID and device ID for the same user when he has not logged in.

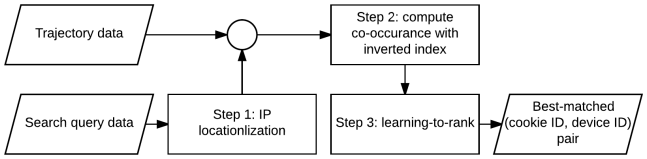


Figure 1: Flowchart of the UNICORN framework

Here we propose a novel and practical user identification framework in cyber-physical space, called UNICORN, for linking users in heterogeneous dataset in cyber-physical space. The flowchart is shown in Figure 1 which consists of three key steps. The first step is the IP locationlization, where we map the IP addresses into location areas, transferring the query data into location data. The second step is to compute the co-occurrence between the trajectory and the IP locations to find the candidate pairs for the linkage. Finally, for the candidate pairs, we use a learning-to-rank (LTR) [13] model to utilize the user's all features shared in the cyber space and physical space to obtain a finalized best-match pair between a cookie ID and a device ID.

In summary, our main research contributions include:

- We first study the user identification problem in the cyber-physical space. We investigate the problem on the mobile query logs and the trajectory data, which are typical data generated in the cyber-physical space.

*Haishan Wu is the corresponding author.

SIGSPATIAL'16, October 31-November 03, 2016, Burlingame, CA, USA

DOI: http://dx.doi.org/10.1145/2996913.2997017

- We propose a novel and practical framework, called UNICORN, to solve the problem. With an inverted index, the framework can efficiently process real-world large dataset on the Map-Reduce platform.
- We conduct evaluation on a real-world dataset to demonstrate the effectiveness and efficiency of our framework.

2. PROBLEM STATEMENT

The purpose is to identify the same users from two datasets, which include:

- Trajectory data: A set of user trajectories, which is

$$S_{\mathrm{L}} = \{\langle id_i^{\mathrm{L}}, \{p_{i,1}, \cdots, p_{i,\mathrm{end}}\}\rangle\}.$$

  id_i^{L} is the device ID of the i-th user, and there is usually one ID for each mobile device. $p_{i,j}$ is the j-th node in the trajectory of u_i, and each node is $p_{i,j} = (x_{i,j}^{\mathrm{L}}, y_{i,j}^{\mathrm{L}}, t_{i,j}^{\mathrm{L}})$, where $x_{i,j}^{\mathrm{L}}, y_{i,j}^{\mathrm{L}}$ are the longitude-latitude coordinates, and $t_{i,j}^{\mathrm{L}}$ is the timestamp.

- Mobile query log data: The set of query logs from the search box of a search engine on mobile phones. By merging the searching records for the same ID, we can get a set of IDs and query logs related to the IDs:

$$S_{\mathrm{Q}} = \{\langle id_i^{\mathrm{Q}}, \{r_{i,1}, \cdots, r_{i,\mathrm{end}}\}\rangle\}.$$

  id_i^{Q} is the cookie ID, and there may be multiple cookie IDs for a user. Each record $r_{i,j}$ is from a query sent by u_i^{Q}, and $r_{i,j} = (\mathrm{IP}_{i,j}^{\mathrm{Q}}, t_{i,j}^{\mathrm{Q}}, s_{i,j}^{\mathrm{Q}})$. $\mathrm{IP}_{i,j}^{\mathrm{Q}}$ is the IP address, and $t_{i,j}^{\mathrm{Q}}$ is the timestamp. $s_{i,j}^{\mathrm{Q}}$ is made up of some extra information, such as the query string, the device OS and the mobile phone model. $s_{i,j}^{\mathrm{Q}}$ may also contain the location $(x_{i,j}^{\mathrm{Q}}, y_{i,j}^{\mathrm{Q}})$, but different from the trajectory data, locations may be inaccessible for some queries.

For each pair of IDs from the two datasets, the aim is to find whether they belong to the same user. We will propose a metric to measure the weighted co-occurrence between the spatial distribution of them. Our solution consist of three parts. Firstly, we model the location distribution of the IP addresses and locationlize the query data. Secondly, we use the inverted index to compute the candidate ID pairs. Finally, we use a learning-to-rank (LTR) approach to considering more features to get a more accurate prediction.

3. THE UNICORN FRAMEWORK

3.1 Extending IP Addresses to Locations

In order to link the query data in cyber space and the trajectory data in physical space, our novel idea is to find out the location distribution for each available IP address, so that we can predict the location where a query was sent by the IP address if there were no location coordinates attached with the query record.

After analyzing the location distribution of some IP addresses, we find there are two types of distributions (Figure 2), including the wide-ranged distribution and the centralized distribution. If an IP address follows a centralized distribution, the locations of the users with this IP address can be predicted as the centers of these central areas. To find out the IP addresses with centralized distribution, we use the DBSCAN algorithm [8] to cluster on the IP locations.

(a) (b)

Figure 2: Wide-ranged (a) and centralized (b) distributions

Let $I_1, I_2, \cdots, I_k$ be the k different clusters which we obtain, and O be the remaining location points not in any clusters. For each cluster I_j, the center $g(I_j)$ is the mean point of I_j. The mean radius $r(I_j)$ of I_j is defined as the root mean square of the distances between $g(I_j)$ and all points in I_j. Then for cluster I_j, we define its weight in S_{IP} to be:

$$w(I_j, S_{\mathrm{IP}}) = \frac{|I_j|}{|S_{\mathrm{IP}}|} \cdot \frac{1}{1 + \alpha r^2(I_j)},$$

which is the confidence that the query is sent from g_{I_j}. In practice, clusters with $w(I_j, S_{\mathrm{IP}})$ less than some threshold w_0 will be ignored.

Following this approach, for a mobile query record without precise location information, we can predict several possible locations for it by IP clustering, together with the prediction confidence for these locations.

3.2 Computing Similarities

In order to match the user IDs from trajectory and online query records, in this section, we propose to use a TF-IDF-based metric to compute weighted co-occurrence similarities between each pair of these cookie IDs and device IDs. Since the computation of all the user similarities takes up a complexity of $O(n^2)$, we use an inverted index based on the Map-Reduce framework to retrieve the top-K similar device IDs for each search query cookie ID.

The mobility trajectories will not be appropriate to be used directly. According to [11, 2], there may be a large number of moving points and noise points in the trajectories, which would be lack of effective information and cause a waste of computation resources. Before using the trajectory data, we will remove the moving points with high speeds.

In this framework, we turn the location records into vectors by the TF-IDF model [17], and use the cosine similarity to represent for the similarity of two trajectories. We can split the city into small grids, each point (x, y) mapped into a grid $(\lfloor \frac{x}{s_g} \rfloor, \lfloor \frac{y}{s_g} \rfloor)$. Consider each grid g_j to be a word, and each trajectory T_i to be a document. The appearance of the user in g_j can be regarded as the appearance of the word in the document. Then we define the term frequency (TF) and the inverse document frequency (IDF) to be:

- $\mathrm{tf}(g_j, T_i) = \log(1 + f_{i,j})$, where $f_{i,j}$ is the frequency that grid g_j appears in the trajectory T_i.
- $\mathrm{idf}(g_j, S) = \frac{1}{\log(1 + |\{i | f_{i,j} > 0\}|)}$.

Then the TF-IDF value for grid g_j in trajectory T_i is $\text{tf-idf}(g_j, T_i) = \mathrm{tf}(g_j, T_i) \cdot \mathrm{idf}(g_j)$. Given the trajectory T_i, we can represent is as a vector

$$\boldsymbol{v}(T_i) = (\text{tf-idf}(g_1, T_i), \text{tf-idf}(g_2, T_i), \cdots, \text{tf-idf}(g_{|G|}, T_i)),$$

and we use its normalized vector $\boldsymbol{v}^*(T_i) = \frac{\boldsymbol{v}(T_i)}{\|\boldsymbol{v}(T_i)\|}$. After that we can use the cosine similarity to represent for the similarity of two trajectories.

Since it takes an $O(n^2)$ complexity to compute the similarity between each pair of vectors for the users from the two datasets (location data and mobile query data), we have built an inverted index to reduce the complexity. Now that we have two sets of vectors:

- $\mathcal{D} = \{v_1, v_2, \cdots, v_{|\mathcal{D}|}\}$, where $v_k = (v_{k,1}, \cdots, v_{k,N})$ is a nonnegative vector representing for the location distribution for some device ID represented by id_k^{L}.
- $\mathcal{K} = \{u_1, u_2, \cdots, u_{|\mathcal{K}|}\}$, where $u_j = (u_{j,1}, \cdots, u_{j,N})$ is a nonnegative vector representing for the location distribution for some cookie ID represented by id_j^{Q}.

The goal of this step is: For each cookie ID id_j^{Q}, we want to find K different devices IDs such that their TF-IDF vectors have the top-K largest cosine similarities with u_j among all device IDs.

According to the uniqueness bound of human mobility introduced in [6], the probability that a single point could be used to identify a unique person is small. Then we will only consider the case that the two user IDs have co-occurrences in at least two grids. We could build an inverted index in which each key is a grid cell. Here we only consider the pairs of IDs who have appeared in at least two different grid cells, and the total time and space requirement can be reduced a lot in this way. We have used a solution similar to [15], which consist of two steps. In the first step, we build an inverted index from the vectors for the user IDs. For each of nonzero entry in one user's vector v_k, we generate a key-value pair, where the key is the grid cell and the value is made up of the user ID and the entry weight. Then we merge all the users that appears in the same grid cell. In the second step, we compute the similarity between each pairs of IDs with co-occurrences. We find all the user pairs that shares the same grid, and then for each pair of users, we find all the grids that they have co-appeared, and then compute the cosine similarity for the pairs of users which share at least two grids. At last, for each cookie ID, we keep K device IDs with the largest similarities with the cookie ID.

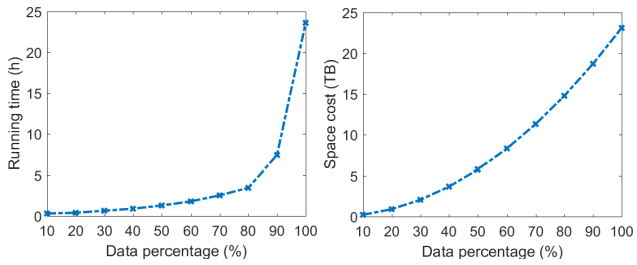
3.3 Learning-to-rank

Besides the location information, we have some extra information shared between the cyber space and the physical space, such as IP addresses, operating systems and phone models. After getting the top-K candidates via TF-IDF similarities, it will be better to take these features into consideration than only using the TF-IDF similarities. In our framework, in order to achieve this purpose, we use the Gradient Boosting Decision Tree (GBDT) algorithm [4] to implement a learning-to-rank (LTR) model [13], in order to get a general ranking for all the top-K device IDs related to the same cookie ID by utilizing different features. The features that we use include the TF-IDF similarity, the number of co-occurrence grids, the maximum distance between co-occurrence grids, and the similarities between the sets of the extra information we have described. After considering all these features in the LTR model, we can get a more accurate prediction for the best matched IDs.

4. EXPERIMENTS

Our experiments are based on the mobile query logs and trajectory data in Harbin, China, during December 2015. In our experiments, we only use the IDs which are active enough. The mobile query log data is obtained from the search query logs of the Baidu search box for the mobile phones. 40.7% of all the query records have location information. There are totally $2,341,283$ active cookie IDs and $270,570,204$ search query records. The trajectory data is obtained from the location records from a mobile application "Mobile Baidu"[1]. There are totally $825,596$ active device IDs and $387,544,953$ location records. The ground truth we used in the evaluation comes from records of account logins, such that if a device ID and a cookie ID have logged in with the same user account, we consider this pair as the same user, which could be used as a ground truth.

Our framework is implemented as a series of Hadoop streaming jobs written in Python 2.7, and the experiments have been running on the Hadoop Map-Reduce cluster [7] with $3,000$ nodes running in parallel for each job.



(a) (b)

Figure 3: Running time (a) and space cost (b) for different indexes

We sample one fifth of the users who were active in the whole year of 2015 to evaluate the running time and space cost, which is shown in Figure 3. From Figure 3 (a), we can see that the running time increases with the amount of data. The main reason that the running time grows faster than $O(n^2)$ when the dataset gets larger (e.g., 90% and 100%) is that the problem of extreme length of some posting lists leads to imbalance problem, and these huge lists cause instability to the Hadoop cluster such that some subtasks will keep crashing and restarting, which takes more time. From Figure 3 (b), we can see that the maximum space requirement grows quadratically with the size of the dataset.

By setting up different values for the threshold of the objective value of LTR, we can get a precision-recall curve for the prediction. We have compared the precision-recall curve of our framework with some comparators, which is shown in Figure 4. These include some modifications of UNICORN, such as: (1) UNICORN with location-based LTR (UNICORN-L): The UNICORN framework without using IP addresses, mobile phone OSes and phone models in LTR. (2) UNICORN without LTR (UNICORN-S): Directly using the TF-IDF similarities without using LTR in the UNICORN framework. There are also some state-of-the-art framework for matching the users in two trajectory datasets, such as: (3) Signal-Jaccard co-filtering introduced in [2], matching the users by co-filtering the ID pairs with signal-

[1]http://xbox.m.baidu.com/wuxian/

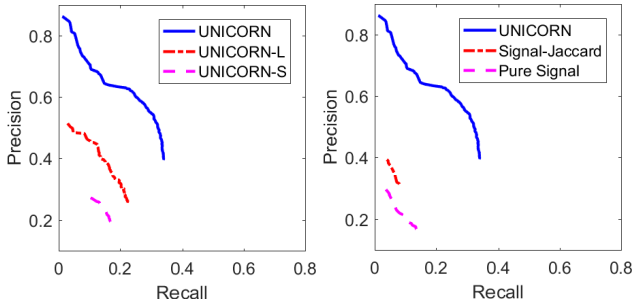


Figure 4: Comparing UNICORN with its competitors by precision-recall curve

based similarity and weighted Jaccard similarity; (4) Signal-based filtering which is similar to (3), but Jaccard similarity is not used. From Figure 4, we can see that our framework has a significant advantage over these comparators, which indicates the effectiveness of our framework.

5. RELATED WORK

Our work in this paper is closely related to the topic of the user identification problem and similarity search technique based on human mobility data. The method for user identification by matching up similar user trajectories from datasets from different resources has been introduced in [2], which has used multi-layer grid indexing in filtering and co-occurrence signals in similarity computation. Authors in [16, 5, 3] have described methods for user identification based on similarity search among trajectories. The author in [20] raised an idea of attaining the lower bounds of similarities during the similarity searching process. Our algorithm is also related with some work on spatial indexing such as [21]. For user identification in cyber space, the most popular problem is how to link and identify the same users from multiple social networks [19, 10, 18, 12, 9, 23].

Another related topic is all-pair similarity searching for vectors. The authors in [1] have proposed an optimization algorithm for all-pair similarity searching, but it is not suitable to run on distributed systems. Metwally et al. [15] have proposed a "V-SMART-Join" framework for the Map-Reduce system, which is a common method for all-pair similarity problems. However, in our problem, the step of matching pairs of IDs in the same inverted index will cause too much output data, and simply dropping big indexes will cause serious information loss.

6. CONCLUSIONS

In this paper, we develop a framework to process the user identification problem between the datasets from the cyber and the physical spaces. At first, for the online mobile query logs without exact location information, we have developed a method to enrich them with the location data by deducing the IP locations after clustering. Then we measure the co-occurrence by TF-IDF metric between the location distributions of the query records and trajectories by building the inverted index. At last we use a learning-to-rank approach to figuring out the matched ID among several possible similar IDs. Our experiments have demonstrated the efficiency and effectiveness of our framework on real-world mobile query log data and trajectory data.

7. ACKNOWLEDGMENTS

The work of Tianyi Hao and Longbo Huang was supported in part by the National Basic Research Program of China Grant 2011CBA00300, 2011CBA00301, the National Natural Science Foundation of China Grant 61033001, 61361136003, 61303195, and the China youth 1000-talent grant.

8. REFERENCES

[1] R. J. Bayardo, Y. Ma, and R. Srikant. Scaling up all pairs similarity search. In *WWW*, pages 131–140, 2007.
[2] W. Cao, Z. Wu, D. Wang, J. Li, and H. Wu. Automatic user identification method across heterogeneous mobility data sources. In *ICDE*, 2016.
[3] L. Chen and R. Ng. On the marriage of lp-norms and edit distance. In *VLDB*, volume 30, pages 792–803, 2004.
[4] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. *CoRR*, abs/1603.02754, 2016.
[5] Z. Chen, H. T. Shen, X. Zhou, Y. Zheng, and X. Xie. Searching trajectories by locations: An efficiency study. In *SIGMOD*, pages 255–266, 2010.
[6] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, 3, 2013.
[7] J. Dean and S. Ghemawat. Mapreduce: Simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, Jan. 2008.
[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, volume 96, pages 226–231, 1996.
[9] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi. On the reliability of profile matching across large online social networks. In *KDD*, pages 1799–1808, 2015.
[10] X. Kong, J. Zhang, and P. S. Yu. Inferring anchor links across multiple heterogeneous social networks. In *CIKM*, pages 179–188, 2013.
[11] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *SIGSPATIAL*, pages 34:1–34:10, 2008.
[12] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan. Hydra: Large-scale social identity linkage via heterogeneous behavior modeling. In *SIGMOD*, pages 51–62, 2014.
[13] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.
[14] P. Luo, S. Yan, Z. Liu, Z. Shen, S. Yang, and Q. He. From online behaviors to offline retailing. In *KDD*, 2016.
[15] A. Metwally and C. Faloutsos. V-smart-join: A scalable mapreduce framework for all-pair similarity joins of multisets and vectors. *PVLDB*, 5(8):704–715, 2012.
[16] L. Rossi, J. Walker, and M. Musolesi. Spatio-temporal techniques for user identification by means of gps mobility data. *EPJ Data Science*, 4(1):1, 2015.
[17] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1986.
[18] S. Tan, Z. Guan, D. Cai, X. Qin, J. Bu, and C. Chen. Mapping users across networks by manifold alignment on hypergraph. In *AAAI*, pages 159–165, 2014.
[19] J. Vosecky, D. Hong, and V. Y. Shen. User identification across multiple social networks. In *NDT*, pages 360–365, 2009.
[20] M. Werner. Bacr: Set similarities with lower bounds and application to spatial trajectories. In *SIGSPATIAL*, pages 29:1–29:10, 2015.
[21] R. T. Whitman, M. B. Park, S. M. Ambrose, and E. G. Hoel. Spatial indexing and analytics on hadoop. In *SIGSPATIAL*, pages 73–82, 2014.
[22] H. Zang and J. Bolot. Anonymization of location data does not work: A large-scale measurement study. In *MobiCom*, pages 145–156, 2011.
[23] X. Zhou, X. Liang, H. Zhang, and Y. Ma. Cross-platform identification of anonymous identical users in multiple social media networks. *TKDE*, 28(2):411–424, 2016.