Research Article

International Journal of
Distributed
Sensor Networks

International Journal of Distributed Sensor Networks
2016, Vol. 12(10)
© The Author(s) 2016
DOI: 10.1177/1550147716671256
ijdsn.sagepub.com
SAGE

TCB: A feature transformation method based central behavior for user interest prediction on mobile big data

Chen Zhou[1], Hao Jiang[1,2], Yanqiu Chen[1], Jing Wu[1,2], Jianguo Zhou[1] and Yuanshan Wu[3]

Abstract

Although traditional spatial-temporal features, such as gyration, probability, and the intervals between consecutive records, have contributed to model human dynamics, the importance of these basic spatial-temporal features in predicting mobile user interest is not fully investigated. Moreover, these typical features ignore the fact that human behaviors are highly predictable and centralized. Specifically, human mobility is constrained in a small area depicted by several hotspots, and users tend to access mobile Internet intensively on several particular timeslots, which are defined as hot-times in this article. Thus, this article proposes a feature transformation method based central behavior to construct informative feature sets. Transformation method based central behavior only requires small amount of records to extract hotspots/hot-times information for every user, and projects original records into a relative vector space, of which coordinates represent the effects suffered from corresponding centralities (hotspots/hot-times). Then, the new space is further enriched by statistical summaries related to hotspots/hot-times. Based on the state-of-the-art classification algorithms, the proposed transformation method based central behavior is validated on a large Usage Detail Records dataset generated in real physical world. Results show that features generated by transformation method based central behavior surpass traditional spatial-temporal features and preference in the terms of precision, recall, and f1-score.

Keywords

User behavior, hotspots, hot-time, user interest prediction

Date received: 18 March 2016; accepted: 28 August 2016

Academic Editor: Michelangelo Ceci

Introduction

With the popularity of mobile Internet, the technology of mobile user interest prediction has attracted extensive researches both from industry and academia. On one hand, user interest prediction provides exciting opportunity to develop many personalized applications, such as search and recommendation. On the other hand, by figuring out users' interest, service providers can make the best of network resources, such as the construction of content distribution network (CDN).

Despite various datasets[1–3] generated in mobile Internet, *who* (*users*), *where* (*location*), *when* (*timestamp*), and *what* (*interest*) are four essential elements in human behavior description.[4] The abundant spatial-temporal information hidden in these elements is widely used in the analysis of human dynamics, such as mobility and interest. Nevertheless, the importance of these

[1]School of Electronic Information, Wuhan University, Wuhan, China
[2]Collaborative Innovation Center of Geospatial Technology, Wuhan, China
[3]School of Mathematics and Statistics, Computational Science Hubei Key Laboratory, Wuhan, China

Corresponding author:
Hao Jiang, School of Electronic Information, Wuhan University, Wuhan 430072, China.
Email: jh@whu.edu.cn

basic spatial-temporal features and how much they matter in predicting mobile user interest are still not fully investigated. Thus, this article focuses on how to predict mobile user interest with those four basic elements. Note that user interest in this article refers to his activity represented by the website extracted from Uniform Resource Identifier (URL). Let $f(\cdot)$ denote a known classification algorithm. In general, four strategies can be taken into account:

(1) $f(what) \rightarrow what$: utilizing preference[5,6] obtained from historical records in predicting interest *what*. Preference refers to the probability of historical activities;
(2) $f(what) \rightarrow what$: utilizing temporal-related features to predict interest *what*. Regardless of the timestamp captured in original records, typical temporal features can be inferred from the field *when* include time intervals between two successive records,[7] the intervals between two successive in same category,[6] and the dwelling time of record[8];
(3) $f(where) \rightarrow what$: utilizing spatial-related features to predict interest *what*. Regardless of the coordinates of a location captured in original records, typical spatial features can be extracted from the field *where* include, gyration,[9] and the distance between two successive records[5,9];
(4) $f(where) \rightarrow when$: utilizing both spatial-temporal related features mentioned in (2) and (3) to predict activity *what*, such as EW[4] in Yuan et al.[4]

However, typical spatial-temporal features that can be used in the strategies above ignore the fact that human behaviors are highly predictable and centralized in spatiality and temporality. On one hand, people spent most of their time at a few locations,[9–11] such as home and work places. We define these places as hotspots in this article. Most individuals have regular mobility pattern: commuting to workplace in the morning, spending most of their daytime at workplace, taking some leisure activities after work, and returning home in the evening.[12] An example of a typical user in a single day is illustrated in Figure 1. On the other hand, bursty is a nature of human behavior.[13] The analysis of human temporal behaviors reveals that both memory effect[14,15] and periodic characteristics[16] can be found in human behaviors, indicating that regular behavior pattern may exist. We then define these hours containing heavy network usage as hot-times. Thus, human behavior shows centrality both on temporality (hot-times) and spatiality (hotspots). Then, an intuitive question is, can these centralities improve the prediction of mobile user interest?

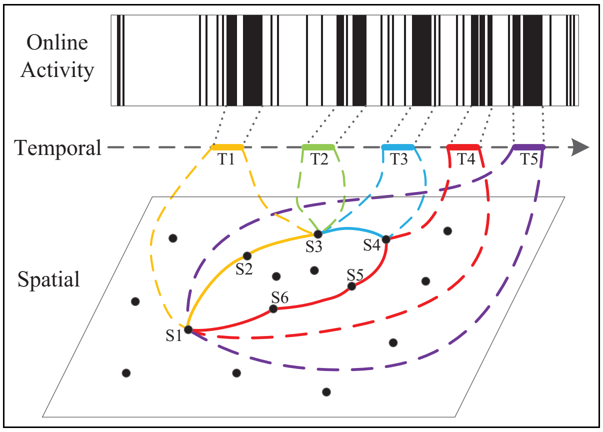

Figure 1. Behavior of a typical user in a single day. The upper panel shows the online activity records. Each bin represents a network usage, and the width indicates the length of corresponding record. The middle panel shows the heavy mobile Internet usage on timeline. The lower panel is spatial plane. Each dot specifies the location of base station. S1, S3, and S4 are hotspots attracting most of the network usages. In particular, S1 and S3 are home and work places, respectively. S4 is a frequently visited location for leisure activities, such as bar or shop. T1–T5 are hot-times and correspond to five heavy network usage time periods. Meanwhile, they also correspond to status transition on spatiality. Both memory effect and periodic characteristics contribute to the predictability of state transition on temporality and spatiality.

Therefore, this article provides a new solution $f(TCB(where, when)) \rightarrow what$, and proposes a feature transformation method based central behavior (TCB) for mobile user interest prediction. Specifically, we utilize centralities both on spatiality and temporality. For each record, the original spatial-temporal information in raw data is projected into a relative vector space, of which coordinates represent the effects that corresponding centralities (hotspots or hot-times) introduce to current online activity. Moreover, TCB collects statistical summaries at hotspots and hot-times to enrich the relative vector space for better user description. Since TCB only requires a small amount of records to obtain users' hotspots and hot-times information, it is quite suitable for user interest prediction in mobile big data.

The main contributions of this article are listed as follows:

- Based on the typical fields that are widely existent in various datasets, we systematically compare the importance of traditional spatial-temporal features and how much they matter in predicting mobile user interest using standard classification algorithms.
- Integrating hotspots and hot-times information, we propose a novel and effective feature

transformation method for interest prediction in the era of mobile big data. Validated by the state-of-the-art classification algorithms, namely DecisionTree (DT) and RandomForest (RF), extensive experiments show that feature sets generated by TCB have great advantages over traditional spatial-temporal feature sets in terms of precision, recall, and f1-score.

- Meanwhile, extensive experiments show that statistical summaries related to hotspots and hottimes can make significant contribution to the prediction of mobile users' interests, which provides new insight into human dynamics related to interest and mobility.

The rest of this article is organized as follows. Related work is presented in the next section followed by the section that offers the characteristics of mobile network usages. The next section details about the feature TCB followed by the section that details the performance and comparison of various feature sets and we conclude this article in the final section.

Related work

This article focuses on predicting the interests of mobile users using multidimensional contextual information, and concerns human dynamics in the following aspects:

Temporal features based

In Yuan et al.,[4] timeslot t was used to depict the temporal pattern in the proposed model EW[4]. Zhao et al.[6] analyzed the time interval Δt_s that user returns to the same interest category, and used a fat-tailed distribution to fit Δt_s, indicating the short-/long-term memories in human interest. Utilizing the interval between two consecutive records Δt, Karsai et al.[7] presented that the number of events in bursty period could be a proper indicator of the dependencies of temporal processes. The dwelling time dur at individual level,[8] however, is widely used in characterizing the scaling law. Zhou et al.[17] utilized time elapse e to depict the popularity aging, and showed that temporal-based predictor was superior to popularity-based predictor in predicting future nodes popularity. Generally, those typical temporal features are used in fitting distribution characteristics at data collective level. Although EW[4] can be applied in predicting user preference, its derivation relied heavily on the assumption of mobility and temporal models.

Spatial features based

The location of cellular network $long, lat$ served as a reasonable proxy for modeling human mobility.[5] In Gonzalez et al.,[9] the radius of gyration g was introduced to explain the truncated power-law distribution in human mobility. The distance between two consecutive records Δl,[5,9] however, was used to model the statistical characteristics in individual human trajectories. Moreover, Trestian et al.[18] investigated the usage of some applications at particular locations, such as home and work regions. Based on association rules mining, Tang et al.[19] proposed a three-stage strategy to predict custom's preference from contextual data. In Zhong et al.,[20] the location of check-ins was used in inferring users' demographics. Although the typical spatial features mentioned above contributed to modeling the characters of human mobility, their effects to user online behavior are not fully investigated. Thus, this article takes those spatial features into account in the task of predicting mobile user interests and analyzes the performance of spatial and temporal features aforementioned.

Preference based

Song et al.[5] adopted the preference selection scheme. Specifically, when a user chooses to return to a historically visited location with the probability P_{ret}, the next location i is chosen according to the probability $\Pi_i = f_i$, in which f_i is the probability that how often location i is visited before. Similarity, the model proposed by Zhao et al.[6] described the human interest dynamics with three states, namely exploration, inertia, and preferential return. When a user prefers to return to an interest category with the probability P_{ret}, the next interest category i is selected according to the probability that how often interest category i is visited before. Both[5,6] are typical preference-based prediction algorithms and the probability that user selects a state only depends on the number of historical locations (activities). To compare the performance of preference and spatial-temporal features in the prediction of mobile online activity, this article selects the model proposed in[6] due to its wide popularity.

Characteristics of mobile Internet usage

In this section, we study the characteristics of mobile network usage on a large-scale Usage Detail Records (UDRs), which is described in details in section "Dataset and preliminary." First, we investigate the diversity of mobile Internet users on spatiality and interest, respectively. Then, we measure how predictable are mobile users' interests and the effect of spatial-temporal information on predicting them. Finally, this article introduces centrality phenomenon both on spatiality and temporality which inspires our feature transformation method.

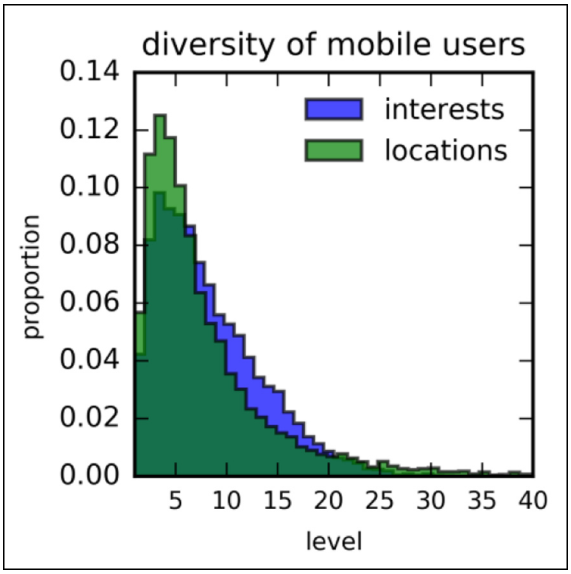

Figure 2. Diversity of mobile users' interests/locations. The horizontal axis represents the level of corresponding diversity, while the vertical axis denotes the proportion of corresponding users.

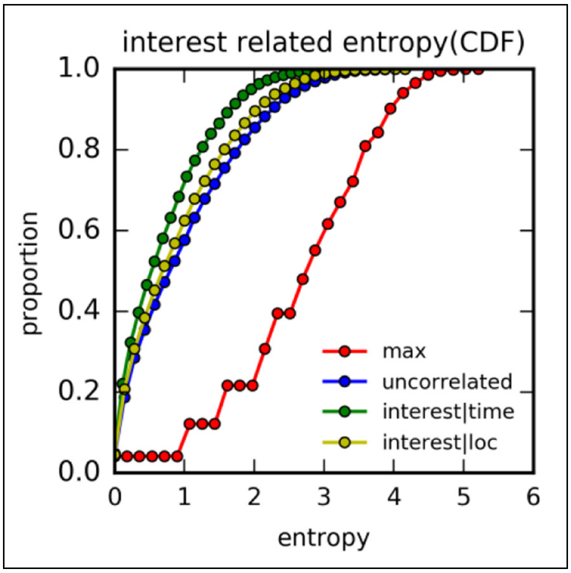

Figure 3. Cumulative distribution of users with max, uncorrelated, and conditional entropy. The form A|B means the conditional entropy of A under the condition B. To obtain the conditional entropy of activity under the condition time, we extract the hour in timestamp from each record.

Diversity

First of all, we take an overview of the diversity of mobile users. The diversity of a user means the number of unique interests/locations visited by him. The statistical results are shown in Figure 2.

Both the diversity of location and interest show the form of lognormal distribution, which are widely found in the analysis of human behavior.[21,22] Users with extreme location/interest diversity are rare, and most of the users have a constrained scope both on spatiality and interest. Specifically, users are more limited on spatiality than interest. The distribution of interest diversity is wider and shorter than that of location diversity. Since it is much easier for users to explore a new website than implementing physical movement, users are more free and willing to explore new interests on Internet, which makes interest prediction much trickier and more valuable. We then measure the predictability of mobile Internet users.

Predictability

To evaluate the predictability of mobile users' interests, this article utilizes information entropy inspired by Song et al.[11] A larger entropy value means the larger uncertainty. First, the max (or random) entropy of user i is defined as follows

$$H_i^{max} \equiv log_2 k_i \tag{1}$$

where k_i is the number of unique interests in the whole observation period. H_i^{max} indicates the maximum randomness of user i. Then, the uncorrelated entropy is

$$H_i^{un} \equiv -\sum_{j \in N_i} P_i(j) log_2 P_i(j) \tag{2}$$

where N_i is the set of interests containing k_i elements, and $P_i(j)$ is the probability of activity j. H_i^{un} indicates the predictability of user i. And finally, the conditional entropy of user i is defined as follows

$$H_i^{con} \equiv -\sum_{j \in M_i} P_i(j) \sum_{l \in N_i} P_i(l|j) log_2 P_i(l|j) \tag{3}$$

where M_i is the extra information set with m_i elements. H_i^{con} indicates the predictability of user interest when an auxiliary feature is specified. We collect the three kinds of entropy mentioned above for each user and give the cumulative distribution function (CDF) in Figure 3. Results show that mobile users' interests are far from random; on the contrary, they can be predicted to a certain degree. Moreover, both temporal (hour) and spatial (location) information contribute to improving predictability, which inspires us that both spatial and temporal information are meaningful and helpful. A natural question is why spatial and temporal information make mobile users' interests more predictable, and we will discuss it on spatiality and temporality, respectively, in the following sections.

Spatiality

To investigate the relation of mobile network usage on spatiality, we refer to tie strength theory proposed in

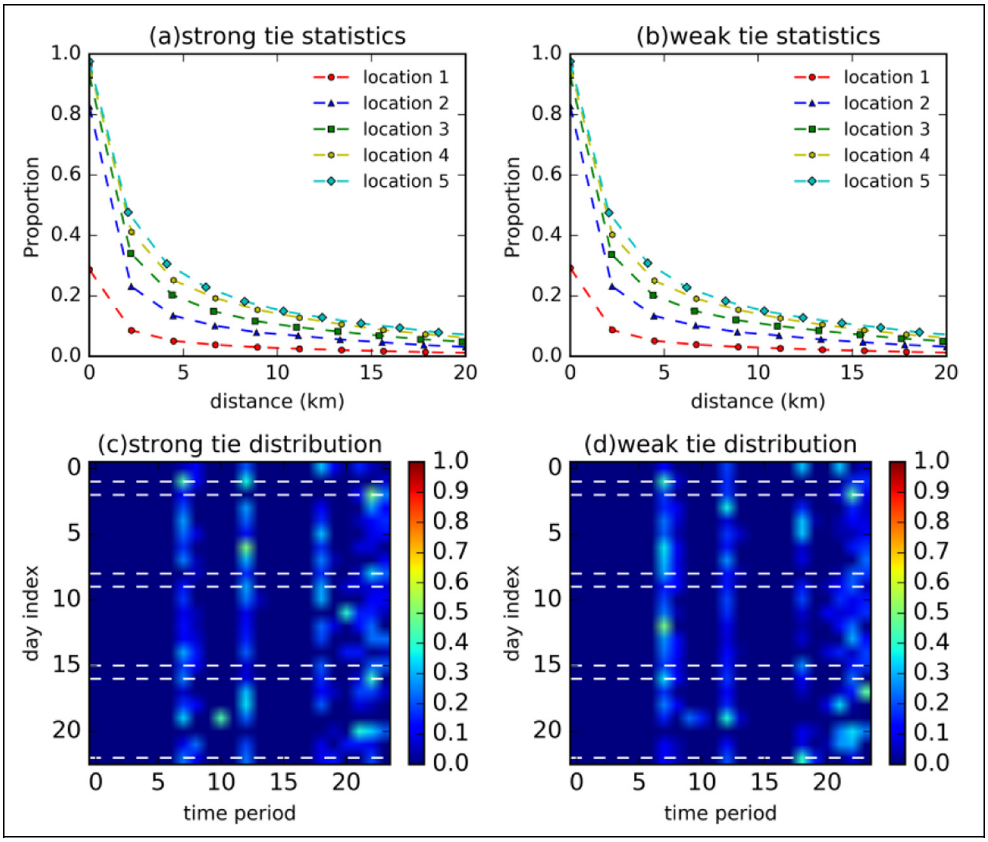

Figure 4. Network usage characteristics on spatiality and temporality: (a) and (b) The complementary cumulative distribution of strong/weak tie from five most favorable locations, respectively. (c) and (d) Network usage of a typical user in observation period (23 days). Color means the proportion of network usage in strong/weak ties. The white-dashed lines indicate weekends. Horizontal axis represents the timeslot, and the vertical axis represents the index of days.

Granovetter.[23] Let $S_i = \{s_i^1, s_i^2, \ldots, s_i^n\}$ denote the locations that user i visited. Then, the strength of tie between user i and base station j is $t_{ij} = (nT_{ij})/(\sum_{k=1}^{n} T_{ik})$, in which T_{ik} is the total time that user i contributes to base station k. We then define the strong tie when $t_{ij} \geq 1$, and weak tie when $0 < t_{ij} < 1$. For every user, this article collects his network usage duration at each location and resort them in descending order. We then compute the distances of current record to the top five locations and collect strong and weak ties, respectively. The complementary cumulative distribution function (CCDF) is shown in Figure 4(a) and (b).

Both strong and weak ties are statistically concentrated in the scope of several most favorable locations. Given a certain frequent visited location, the further a place is apart, the less likely it is to attract mobile network usages. Given certain distance threshold, the probability that network usages occur at a location tend to be in proportion to its popularity. Besides, over 70% of network usages (strong/weak ties) are contributed to the most frequent popular location (location 1), and the second popular location (location 2) attracts about 18% strong and weak ties.

Figure 4(a) and (b) indicate that users tend to access mobile Internet at several locations and they may be more informative compared to traditional spatial features, such as gyration and distance between consecutive records on spatial. Similar to the philosophy of principle components analysis (PCA), we select hotspots defined in section "TCB algorithm" as reference anchors to specify their effects on current online activity on spatiality. The details about this process are presented in section "TCB algorithm."

Temporality

To better illustrate how regular mobile Internet usage is on temporality, we randomly select a user from dataset and plot his network usages in Figure 4(c) and (d). Without losing of generality, we split timeline into 24 timeslots by hour. And weekends are represented by white-dashed lines.

Several clear vertical lines are observed in the whole observation period, indicating that user tends to access mobile Internet at several particular timeslots and the temporal pattern is stable. This phenomenon also coincides with human behavior characteristics such as memory effect[14,15] and period patterns.[16] So similar to spatial features, we utilize hot-times defined in section "TCB algorithm" as reference anchors to specify their effects on current online activity on temporality. We will provide the detailed description for this process in section "TCB algorithm."

Feature transformation based central behavior

According to the analysis provided in section "Characteristics of mobile Internet usage," mobile users reflect centralities both on spatiality and temporality. Specially, users tend to access mobile Internet at several most frequently visited locations (hotspots) on spatiality and contribute relatively more network usages at particular timeslots (hot-times). We then integrate these centralities into the designing of TCB.

TCB algorithm

The intuition behind TCB is that centralities (hotspots and hot-times) can affect user online activity. We assume that these centralities are stationary on time series, and the more closer a centrality is, the influential it can be. For the simplicity of illustration, we make the definitions below:

> ***Definition 1.*** Let hs_m^i be the mth hotspot of user i, the influence that hs_m^i affects online activity occurred at location j is defined as $HS(i,j,m) = I(hs_m^i)\, exp(-dist(j, hs_m^i))$. $I(hs_m^i)$ is the total influence of hs_m^i, and $dist(j, hs_m^i)$ is the Euclidean distance between hotspot hs_m^i and location j;
>
> ***Definition 2.*** Let ht_n^i be the nth hot-time of user i, the influence that ht_n^i affects online activity occurred at timeslot k is defined as $HT(i,k,n) = I(ht_n^i)\, exp(-inter(k, ht_n^i))$. $I(ht_n^i)$ is the total influence of ht_n^i, and $inter(k, ht_n^i)$ is the time interval between hot-time ht_n^i and timeslot k;
>
> ***Definition 3.*** For user i, the total influence of the kth centrality is $I(c_k^i) = (NS(c_k^i))/(NW(c_k^i) + 1)$, where $NS(c_k^i)$ and $NW(c_k^i)$ are the numbers of strong and weak ties at centrality c_k^i, respectively. The definitions of strong/weak ties are presented in section "Spatiality."

Algorithm *feature transformation method based central behavior (TCB)* is presented in the algorithm 1:

Algorithm 1: *TCB(records, candidates, win, m, n).*

```
 #obtain hotspots for every user in win days
 1: hotspots ← CentralityDetection(records, candidates, win,
      "spatial", m)
 #obtain hot-times for every user in win days
 2: hot−times ← CentralityDetection(records, candidates, win,
      "time", n)
 #obtain statistical information at hotspots and hot-time
 3: SI ← Statis_Info(records, hotspots, hot-times, win)
 4: result = [ ]
 5: for record in records:
 6:    u, l, t ← obtain basal information for current record
       #obtain the effect of each hotspot to current location l
 7:    $d_1^u, \ldots, d_m^u$ ← Effect_by_HS(l, hotspots[u])
       #obtain the effect of each hot-time to current time t
 8:    $e_1^u, \ldots, e_n^u$ ← Effect_by_HT(t, hot−times[u])
 9:    temp_result ← concate$\left([d_1^u, \ldots, d_m^u], [e_1^u, \ldots, e_n^u], SI[u]\right)$
10:    result.append(temp_result)
11: return result
```

To begin with, TCB collects m hotspots and n hot-times for each user, and the results are descending ordered according to their influences in *win* days. The details of *CentralityDetection* are discussed in the next section. For better behavior description, TCB obtains the statistical summaries at each hotspot and hot-time to describe user behavior. To this end, we refer to Wang et al.,[24] Zhao and Zhou,[25] and Palmisano et al.[26] and select average displacement to hotspot, average record duration on hotspot, average time interval to hot-time, and average record dwelling time on hot-time to characterize the user behavior. In section "Evaluation," we will demonstrate how these statistical summaries affect the performance of user interest prediction.

Then, in the loop L5-L10, TCB first fetches original information such as user u, time t, and location l from current record. Then, TCB computes the effects that current record received from each hotspot and hot-time, respectively, according to its influence, namely $d_1(e_1)$ is the effect received from most influential hotspot (hot-time), and $d_m(e_n)$ is the result of least influential one. The function *Effect_by_HS*(*Effect_by_HT*) is designed according to the definitions mentioned above. Note that the number of real hotspots (hot-times) can be less than $m(n)$, we complement the values related to missing hotspots or hot-times by zeros since these hotspots/hot-times have no effect on current record. In this way, original spatial-temporal information in raw record is projected into a new vector space, of which coordinates represent the effects received from the centralities (hotspots or hot-times) ranking in certain order based on their influences. Finally, the effects suffered from hotspots/hot-times, together with the statistical information at hotspots and hot-times, are concatenated into one record. The results returned by TCB are used for model training and validation.

Based on the above description, all features used in this article are shown in Table 1. They are classified into five groups according to their generation and background. In particular, HS and HT are feature sets produced by TCB. Both HS and HT consist of effects received from centralities and statistical summaries at centralities.

Table 1. Symbols and corresponding illustration.

Original features (O)	*dur*	Dwelling time of the record
	long	Longitude of the location of the record
	lat	Latitude of the location of the record
	t	Timeslot that record occurred
Temporal features (T)	Δt	Time interval between two consecutive records
	Δt_s	Time interval between two consecutive records in same interest
Spatial features (S)	Δl	Distance interval between two consecutive records
	g	Gyration
Hotspots-related features (HS)	d_i^u	Effect of *hotspot i* to current location *l* for user *u*
	$dis_{i, aver}^{u, win}$	Average displacement to *hotspot i* in *win* days for user *u*
	$sd_{i, aver}^{u, win}$	Average record duration on *hotspot i* in *win* days for user *u*
Hot-times related features (HT)	e_i^u	Effect of *hot-time i* to current timeslot *t* for user *u*
	$int_{i, aver}^{u, win}$	Average interval to *hot-time i* in *win* days for user *u*
	$td_{i, aver}^{u, win}$	Average record duration on *hot-time i* in *win* days for user *u*

Centrality detection

As it mentioned in sections "Spatiality" and "Temporality," hotspots and hot-times are behavior centralities on spatiality and temporality, respectively. Therefore, the philosophy behind *CentralityDetection* is similar to "hot points" detection in time series. Different from Isaacman et al.,[10]*CentralityDetection* is designed to find the most influential *k* centralities specified by *metric* (spatial or temporal) according to the influence values defined in section "TCB algorithm." When *metric* is set to "spatial," the historical locations of each user are processed, otherwise the timeslots. For the simplicity of illustration, we refer location and timeslot as "point."*CentralityDetection* collects the numbers of strong ties and weak ties at each point. In this case, *NS* and *NW* are the vectors with same length. Then, the influences of all points are computed according to Definition 3 and stored in dictionary *influence* by descending order based on the influence values. The number of centralities is the minimum of *k* and the length of points list. Therefore, for some users, the number of their centralities can be less than *k* if they use mobile Internet at less than *k* locations or timeslots. The pseudocode of *CentralityDetection* is presented as follows:

Algorithm 2: *CentralityDetection*(*records*, *candidates*, *win*, *metric*, *k*).

```
r_win $\leftarrow$ obtain records occurred in win
result_dic $\leftarrow \emptyset$      #initialize result dictionary
for user in candidates:
    NS $\leftarrow$ collect the number of strong_tie in r_win at each point on metric
    NW $\leftarrow$ collect the number of weak_tie in r_win at each point on metric
    influence $\leftarrow$ sort (NS/(NW + 1)) in descending order
    c_list = [ ]
    for i in range(min(len(influence), k)):
        c_list.append(influence[i].point)
    add user, c_list into result_dic
return result_dic
```

Let *N* be the number of *candidates*, and *k* the number of centralities. Then, the complexity of Algorithm 2 is O(*Nk*). Hotspots or hot-times obtained in this process are further used to collect statistical information and compute the effects they bring to mobile network usages.

Correlation analysis

In this section, this article investigates the correlation between different feature sets with users' interests. Distance correlation (DC) $\mathcal{R} \in [0, 1]$ is a new metric to measure the dependence between two random variables. It equals to 0 if and only if the two random variables are independent.[27] DC is effective both in linear and non-linear situations. Besides, it can be applied between the variables with different dimensions, and regardless of whether it is categorical, continuous, or discrete.[28] Thus, DC is quite suitable for measuring the correlation between feature sets and the corresponding interests.

To make the distance between different interests computable, this article utilizes dummy variables to represent each interest. For the details about DC, we refer the readers to Székely et al.,[27] and we apply the package "energy"[29] in the process of computing DC values. Since the complexity of DC is O(n^2) (n denotes the number of samples), it is impractical to compute the DC values from an overall perspective. Thus, this article utilizes different steps to split whole records into several blocks and collects the DC values of different feature sets at each block.

As shown in Figure 5, in general, both HS and HT have similar performance, and show great advantages over O/S/T, indicating that feature sets generated by TCB are much more informative in predicting mobile users' interests. The performance of O and S is similar, while T ranks the worst. It also suggests that classical spatial-temporal features (T and S) are limited in

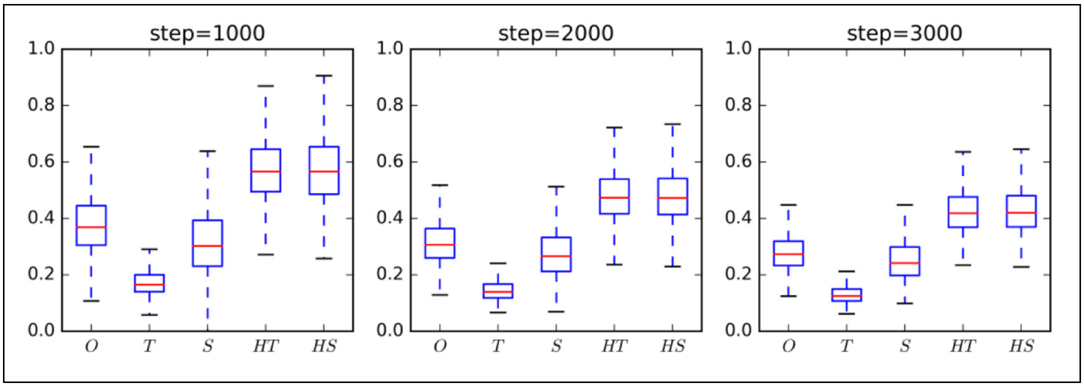

Figure 5. Correlation between different feature sets with users' interest with various steps. The red line in each box indicates the median. Only three steps are taken due to the huge time complexity. The numbers of hotspots and hot-times are set to 3, and statistical window is 7.

Table 2. Key fields and examples in UDRs.

phoneNUM	Start	End	Location	URL
68960814031	2014-11-21 12:02:19	2014-11-21 12:02:26	689B_83A1	m.baidu.com
69061452339	2014-11-21 06:23:03	2014-11-21 06:23:20	67BD_3345	m.sohu.com
69061454535	2014-11-21 13:40:52	2014-11-21 13:41:09	67BD_5BB9	wap.cmread.com

UDRs: Usage Detail Records; URL: Uniform Resource Identifier.

predicting mobile users' interests. In section "Evaluation," we will compare the performance of different feature sets in detail.

Evaluation

In this section, we compare the performances of various feature sets under the state-of-the-art classification algorithms. In particular, we investigate the performance of $X_{/STA}$ to analyze the importance of statistical summaries related to feature set X. Moreover, we also investigate the effects of the number of hotspots and hot-times, and the effect of length of time window used to obtain them. Although the framework of our experiment can be regarded as an interest prediction method, we lay our emphasis on the performance of various feature sets under standard classification algorithms.

Dataset and preliminary

UDRs used in this article span over 23 days, covering nine municipalities in the south of China. Each piece of UDR is generated when user accesses to mobile Internet via applications on his or her smart phone. The key fields and examples in our UDRs are provided in Table 2. Note that *Location* consists of location area code (LAC) and CELL ID of a cellular tower where *phoneNUM* is served, and the corresponding longitude and latitude can be referred by a known translation table. For privacy, all phone numbers are translated into hash codes before we can reach them.

Each record is a four-elements tuple $who, when, where, what$, indicating that user *phoneNUM* contributed $(end - start)$ seconds at *location* for mobile online activity in *URL*. Those four elements are typical and can be found in various kinds of datasets, such as news,[1] Twitter,[2] and the posts in discussion forums.[3] Besides, an increasing number of evidences, such as Serrano-Sanchez et al.[30] and Archer et al.,[31] have shown that users tend to spend more time and energy on screen-based activities than physical activities. Thus, UDRs contain meaningful messages on human behavior and provides fine-grained description for users' mobility and interest behavior.

To obtain ground truth data used for future training and validation, several challenges need to be considered. On one hand, due to the screen limitation of mobile devices, it is common to have mistaken operations in mobile Internet usages. Thus, not only the meaningful online behavior, massive noises are also captured. On the other hand, the bipartite matrix $user \sim URL$ is dramatically sparse due to the uniqueness of *URL*. Therefore, this article extracts the main website in *URL* to represent user interest, such as extracting "cmread.com" from "wap.cmread.com" in Table 2.

To obtain reliable and representative candidate users, we discard the individuals with less than 15 records everyday on average. When filtering candidate

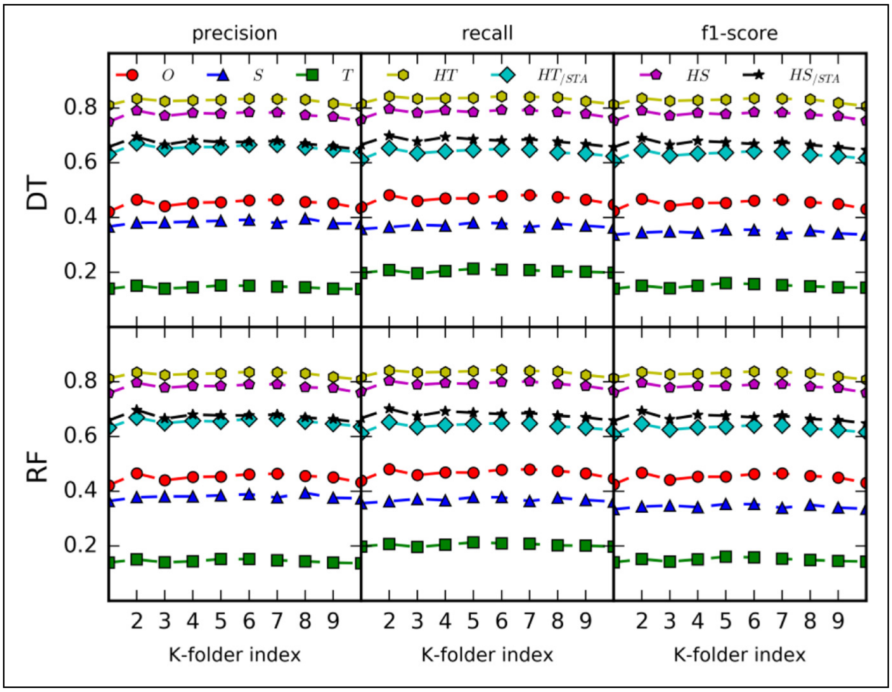

Figure 6. Performance (precision/recall/f1-score) of DT and RF, respectively, when different feature sets used.

websites, duration time, frequency, and the number of coverage users are referred. Specifically, in every aspect, websites are ranked in descending order according to their values, and we select the subsets when the energy exceeds defined threshold $k \in [0, 1]$. And then we choose the websites if they exist in all the subsets generating from each aspect. Finally, we obtain more than 7 million records covering 179 candidate websites and 9720 candidate users in our valid dataset when we set $k = 0.9$.

Euclidean distance is not sensitive if variates vary in small intervals. Moreover, scale transformation methods seem hard to guarantee the fairness among all variates. To avoid these defects, we choose classification algorithms in Pedregosa et al.[32] using entropy as an index in the process of modeling, namely DT and RF. In our cases, Gini impurity is used for both DT and RF in measuring the quality of a split. RF has 10 trees, and each of it with $sqrt(nf)$ features (nf is the number of features that RF receives).

This article then utilizes precision, recall, and f1-score to measure the performance of different feature sets, which are defined in equations (4)–(6).

$$precision = \sum_{l \in L} \omega_l p_l \tag{4}$$

$$recall = \sum_{l \in L} \omega_l r_l \tag{5}$$

$$f1\text{-}score = \sum_{l \in L} \frac{2\omega_l p_l r_l}{p_l + r_l} \tag{6}$$

Specifically, $p_l = TP_l/(TP_l + FP_l)$ is the precision of activity l, and $r_l = TP_l/(TP_l + FN_l)$ is the recall of activity l. L is the total candidate websites set; TP_l is the number of true positives of activity l; FP_l is the number of false positives of activity l; and FN_l is the number of false negatives of activity l. Finally, ω_l is the weight of activity l in the test dataset.

This article utilizes cross validation in our experiment to avoid overfitting. In particular, for each feature set, we split total data into K folders, in which $K - 1$ folders used for model training and the remaining used for validation. This process is repeated for K times so that each sample in the dataset is used both in model training and validation. We set $K = 10$ in our experiment.

Performance comparison

In this section, we compare the performances of TCB among different feature sets O/T/S. Since TCB generates HS and HT, and both HS and HT contain corresponding statistical summaries, we also investigate the performance of sub-feature sets (namely HS, HT) generated by TCB and the effect of statistical summaries related to them. Without losing of generality, the number of hotspots m and hot-times n are set to 3, respectively, and the statistical window win is set to 7.

Single feature set. First of all, we investigate the performance of independent feature set, respectively, and the results are given in Figure 6. Note that $X_{/STA}$ indicates

Table 3. Performance (precision/recall/f1-score) of dual combination of different feature sets.

		S	T	HT	HT$_{/STA}$	HS	HS$_{/STA}$
DT	O	0.436/0.431/0.432	0.45/0.45/0.449	0.809/0.814/0.81	0.756/0.763/0.758	0.796/0.801/0.797	0.758/0.763/0.759
	S		0.412/0.417/0.41	0.823/0.831/0.825	0.774/0.783/0.776	0.778/0.787/0.779	0.753/0.763/0.754
	T			0.818/0.815/0.816	0.549/0.544/0.542	0.771/0.768/0.768	0.644/0.645/0.643
	HT				0.825/0.833/0.826	0.808/0.814/0.809	0.805/0.811/0.806
	HT$_{/STA}$					0.8/0.805/0.801	0.764/0.77/0.765
	HS						0.773/0.783/0.774
RF	O	0.448/0.456/0.442	0.434/0.444/0.435	0.817/0.823/0.819	0.77/0.778/0.772	0.809/0.814/0.81	0.763/0.77/0.765
	S		0.373/0.383/0.373	0.825/0.832/0.826	0.773/0.782/0.774	0.783/0.793/0.784	0.751/0.762/0.753
	T			0.831/0.834/0.831	0.54/0.537/0.528	0.785/0.789/0.785	0.631/0.635/0.629
	HT				0.826/0.833/0.827	0.815/0.819/0.815	0.813/0.818/0.814
	HT$_{/STA}$					0.809/0.815/0.811	0.77/0.777/0.771
	HS						0.776/0.786/0.777

HT: hot-time; HS: hotspot.

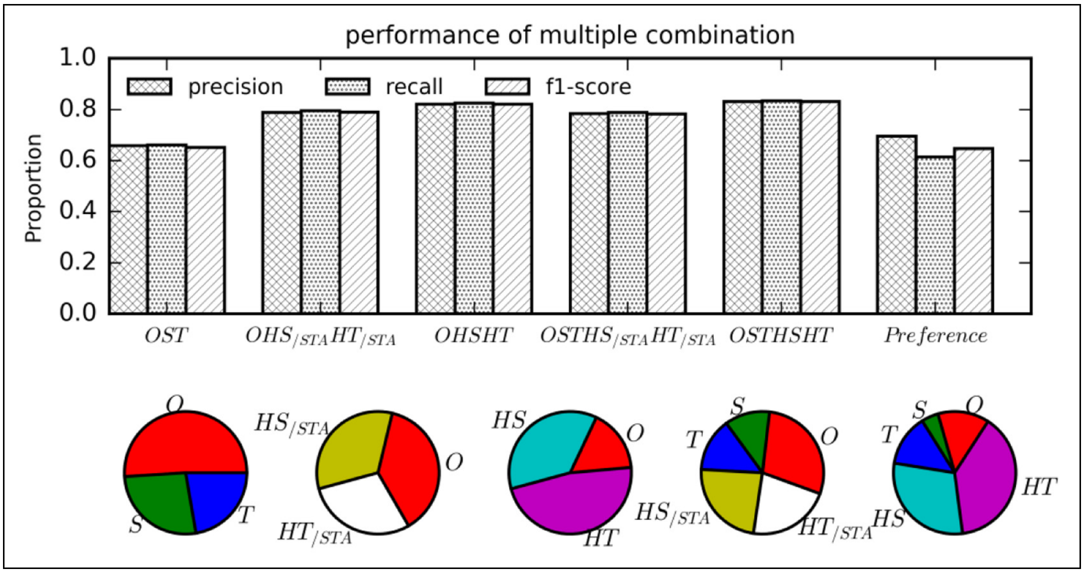

Figure 7. Prediction performance comparison. The upper panel shows the performance of different feature sets. The lower panel shows the importance of different feature sets in model fitting in corresponding classification task (from left to right).

the feature set related X without corresponding statistical summaries.

Regardless of metrics and classification algorithms, the tendencies of different feature sets are nearly identical in the process of cross validation, indicating that the feature sets used in DT and RF are stable and reliable. Compared to other feature sets, HS and HT rank the best, followed by HS$_{/STA}$ and HT$_{/STA}$, indicating that (1) feature sets generated by TCB are much informative and suitable for predicting mobile user interest, and (2) statistical information related to hotspots (hot-times) in HS (HT) is important for HS (HT) to achieve a better performance. On the contrary, the performances of traditional spatial-temporal feature sets S and T are even worse than that of original features O, implying that spatial-temporal features in single dimensionality are insufficient in predicting mobile user interest. Both spatial and temporal information should be taken into account. Moreover, the performances of O, S, T, HS, and HT are mainly consistent with the relation depicting in Figure 5, indicating that the DCs between different feature sets and user interest are reliable.

Dual combination. Second, we examine the performance of dual combination of different feature sets. Note that values in Table 3 are the average after 10 runs in cross validation.

Results show that the performance of DT and RF is very similar. Compared to original feature set O (Figure 6), both S and T provide additional meaningful information for classification. However, the improvement seems to be limited. Besides, both of the performance of OS and OT surpass the performance of ST, which means traditional spatial-temporal feature sets are redundant to each other, even less informative compared to original information recorded in O. On the contrary, feature sets generated by TCB are much more abundant, compared to the original and traditional spatial-temporal feature sets, and bring universal and remarkable improvement in dual combination cases. In particular, on temporality, integrating original information O, the precision improvement brought by HT is 38.3% compared to feature set T in the best performance (when RF is executed). While on spatiality, HS improves the precision by 36.1% compared to feature set S in the best performance. Moreover, although statistical summaries about hotspots and hot-time are meaningful in the prediction of user interest, original feature set O and traditional spatial feature set S can still make considerable compensation when statistical summaries are missing while the traditional temporal feature set T is helpless. Finally, despite their impressive performance of HS and HT when integrating O, S, and T, respectively, the combination of HS and HT does not show great improvement. An intuitive interpretation is that HS and HT are highly coupling in space–time. Since user mobility is constrained in a small area, mobile network usages occurred in hot-time are of great probabilities in hotspots.

Multiple combination. Next, we investigate the performance of multi-combined feature sets, and how useful each feature set is in different combination cases. RF is used in modeling for simplicity. Original feature set O is used as starting line since it is the basic information in raw data. We also compare the performance of preference select proposed in Zhao et al.[6] For the simplicity, the probability of inertia is set to 0. User prefers to return to a historically interest category by the probability P_{ret}, and explore a new interest category with the probability $1 - P_{ret}$. The historical interest category i is chosen according to the probability $\Pi_i = f_i$, in which f_i is the probability that how often interest category i is visited before. Then, we vary P_{ret} in $[0, 1]$ with interval 0.1 and present the best performance of preference selection. Results are shown in Figure 7. All values produced by different feature sets are the average after 10 runs.

As it shown in Figure 7, HS- and HT-related feature sets bring universal improvement to user interest prediction. Compared to OST, OHSHT promotes the performance by 16.2%, and the figure is 13% when $\mathrm{OHS}_{/\mathrm{STA}}\mathrm{HT}_{/\mathrm{STA}}$ is used. By integrating HS and HT, the final precision of OSTHSHT can even reach 83%, generating 17.2% improvement compared to using OST alone. Without statistical summaries related to HS/HT, the improvement decreases to 12.6%; however, it is still remarkable. Note that the performance of $\mathrm{OHS}_{/\mathrm{STA}}\mathrm{HT}_{/\mathrm{STA}}$ and $\mathrm{OSTHS}_{/\mathrm{STA}}\mathrm{HT}_{/\mathrm{STA}}$ is very close, which indicates traditional spatial-temporal features in S and T are redundant when taking $\mathrm{HS}_{/\mathrm{STA}}$ and $\mathrm{HT}_{/\mathrm{STA}}$ into account. Preference selection is only superior to OST, far less impressive than that combines HS and HT.

We then go further and investigate how much different feature sets matter when multiple sets are used in model training from the perspective of feature importance.[32] The importance of each feature set is the sum of corresponding features. Compared to traditional spatial-temporal feature sets S and T, HS and HT are more valuable in the process of modeling, which is obvious in the cases of OST and OHSHT. In the case of $\mathrm{OSTHS}_{/\mathrm{STA}}\mathrm{HT}_{/\mathrm{STA}}$, although the most important feature set is O, the performance improvement brought by the combination of O, S, T, $\mathrm{HS}_{/\mathrm{STA}}$, and $\mathrm{HT}_{/\mathrm{STA}}$ is about 30% compared to using O alone. Finally, in the case of OSTHSHT, HS and HT are the most significant feature sets in the process of classification modeling. The different importance distribution in $\mathrm{OSTHS}_{/\mathrm{STA}}\mathrm{HT}_{/\mathrm{STA}}$ and OSTHSHT also indicates that statistical information at hotspots and hot-times can make impressive contribution to users' interests prediction.

Effect of the number of hotspots/hot-times

Since the number of hotspots/hot-times affect the collection of the effects received from hotspots/hot-times ($\mathrm{HS}_{/\mathrm{STA}}$/ $\mathrm{HT}_{/\mathrm{STA}}$), and the corresponding statistical summaries at hotspots/hot-times, in this section, we investigate the performance of $\mathrm{HT}/\mathrm{HT}_{/\mathrm{STA}}/\mathrm{HS}/\mathrm{HS}_{/\mathrm{STA}}$ as a function of the number of hotspots/hot-times. For the simplicity of illustration, we refer hotspot and hot-time as centrality, and set *win* to 7. Results are shown in Figure 8.

At the very beginning, the performances of $\mathrm{HS}_{/\mathrm{STA}}$ and $\mathrm{HT}_{/\mathrm{STA}}$ show significant improvement when the number of centralities increases, and they reach saturation when the number of centrality gets larger. The saturation states for both feature sets show great advantages over O/S/T (Figure 6). However, the variation of the number of centralities brings little effect when corresponding statistical summaries are taken into account (see the performance of HS and HT), indicating that statistical summaries at centralities can make effective compensation when the number of centralities is limited. Moreover, the performance of HS is always superior to that of $\mathrm{HS}_{/\mathrm{STA}}$, which implies statistical summaries at hotspots cannot be replaced by a larger number of hotspots in $\mathrm{HS}_{/\mathrm{STA}}$. This phenomenon also verifies that statistical summaries related to hotspots and hot-times can make significant contribution to the prediction of mobile user interests.

To sum up, although a larger number of centralities can lead to a better prediction performance, it also

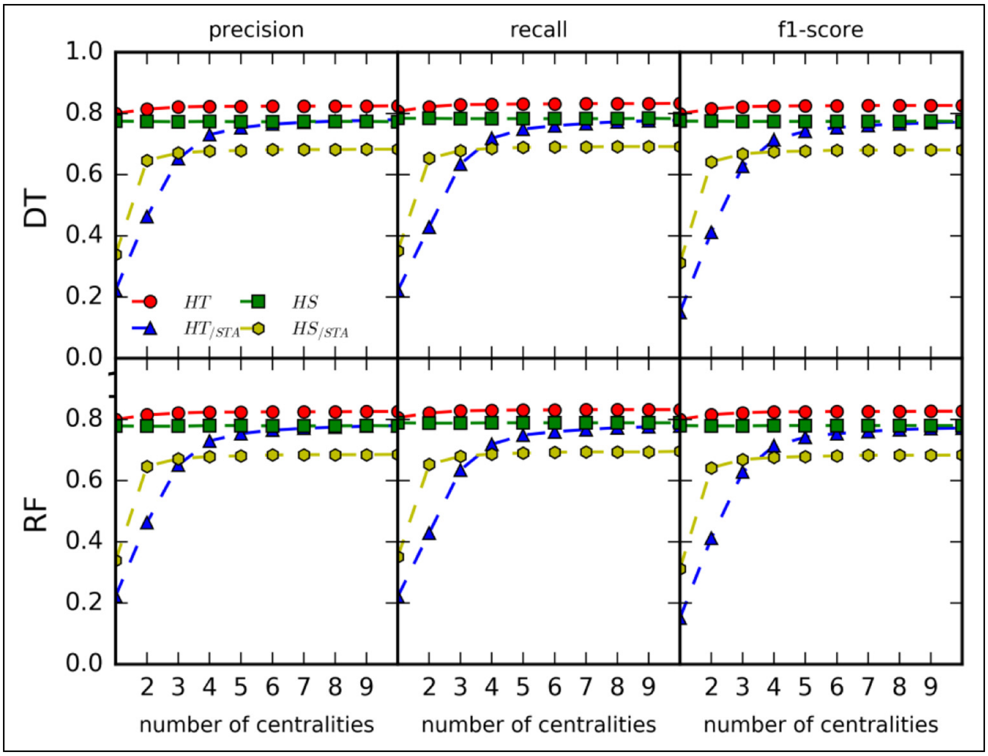

Figure 8. Performance of HS/HS$_{/STA}$/HT/HT$_{/STA}$ when the number of hotspots/hot-times change.

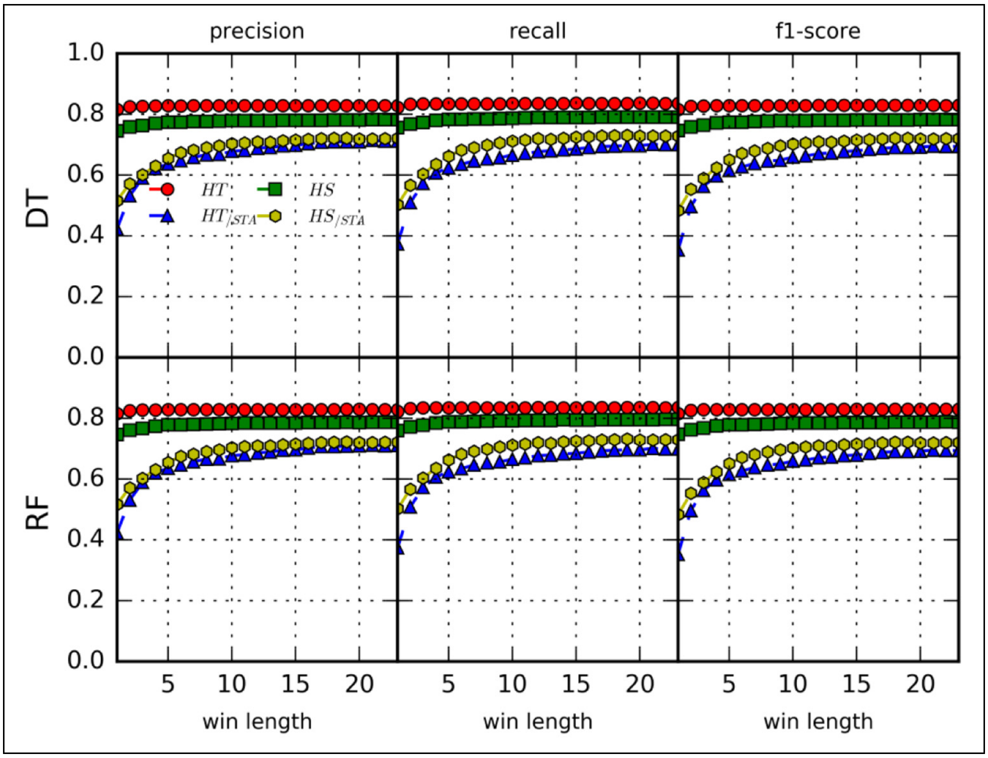

Figure 9. Performance variation when win length changes.

means a higher dimensionality for data processing. With the help of statistical summaries at centralities, TCB can achieve relatively high performance even in a limited dimensionality, which makes it favorable in the era of mobile big data.

Effect of statistical window

Since the hotspots/hot-times information and corresponding statistical summaries are extracted in a time window *win*, we then investigate the performance of $HT/HT_{/STA}/HS/HS_{/STA}$ when time window *win* changes. A higher *win* means a larger amount of data needed for the process of extracting hotspots/hot-time-related information. Without loss of generality, the granularity is set to day, and the number of hotspots *m* and hot-times *n* are set to 3, respectively. Results are shown in Figure 9.

Results show that time widow has little effect on the performance of HT, indicating that few records are required to obtain hot-time information that is used for projecting original temporal features into new vector space. On the other hand, the performance of HS, $HS_{/STA}$ and $HT_{/STA}$ increases along with *win*. Although a larger *win* indicates a detailed description for mobile network usage, the improvements are not significant when *win* is larger than 7. In conclusion, TCB is effective and only requires a small amount of data for building feature transformation space, which makes it quite applicable in the scenarios of mobile big data.

Conclusion

Based on the ground truth that human behaviors are highly predictable, this article proposes a novel features construction method TCB by utilizing hotspots and hot-times information. Specifically, TCB utilizes several hotspots and hot-times as reference anchors on spatiality and temporality, respectively. Then, the effects that current record received from each hotspot and hot-time are collected according to its influence and distance. Besides, statistical summaries, such as average displacement to hotspot, average record duration on hotspot, average time interval to hot-time, and average record dwelling time on hot-time, are also meaningful and integrated in mobile user interest prediction. Based on classical classification algorithms, such as DT and RF, the proposed TCB is validated on a large UDRs dataset generated in real physical world. Results show that features generated by TCB have an advantage over traditional spatial-temporal and preference features on precision, recall, and f1-score. With the help of TCB, the final precision can reach 83%, more than 17.2% improvement compared to using original and traditional spatial-temporal features when RF is conducted. TCB only requires $who, when, where$ in a short observation to produce more informative feature sets, it can be easily applied in multiple fields concerning user preference, such as customized recommender system, intelligent city, and demographic analysis.

Acknowledgements

The authors would like to thank anonymous reviewers.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was part of the National Natural Science Foundation of China (NSFC) under Grant No. 61371126, the National High Technology Research and Development Program of China (863 Program) under Grant No. 2014AA01A707, and the Applied Basic Research Programs of Wuhan under Grant No. 2014010101010026.

References

1. Szabo G and Huberman BA. Predicting the popularity of online content. *Commun ACM* 2010; 53(8): 80–88.
2. Asur S, Huberman BA, Szabo G, et al. Trends in social media: persistence and decay. In: *Proceedings of the 5th international conference on weblogs and social media (ICWSM)*, Barcelona, Catalonia, Spain, 17–21 July 2011, p. 434. Menlo Park, CA: The AAAI Press.
3. Aperjis C, Huberman BA and Wu F. Harvesting collective intelligence: temporal behavior in yahoo answers (arXiv preprint arXiv:1001.2320), 2010, https://arxiv.org/abs/ 1001.2320
4. Yuan Q, Cong G, Zhao K, et al. Who, where, when, and what: a nonparametric Bayesian approach to context-aware recommendation and search for Twitter users. *ACM T Inform Syst* 2015; 33(1): 2.
5. Song C, Koren T, Wang P, et al. Modelling the scaling properties of human mobility. *Nat Phys* 2010; 6(10): 818–823.
6. Zhao ZD, Yang Z, Zhang Z, et al. Emergence of scaling in human-interest dynamics. *Sci Rep* 2013; 3: 3472.
7. Karsai M, Kaski K, Barabási A-L, et al. Universal features of correlated bursty behaviour. *Sci Rep* 2012; 2: 397.
8. Zhao Z-D, Cai S-M and Lu Y. Non-Markovian character in human mobility: online and offline. *Chaos* 2015; 25(6): 063106.
9. Gonzalez MC, Hidalgo CA and Barabasi A-L. Understanding individual human mobility patterns. *Nature* 2008; 453(7196): 779–782.
10. Isaacman S, Becker R, Cáceres R, et al. Identifying important places in people's lives from cellular network data. In: Lyons K, Hightower J, Huang EM (eds) *Pervasive computing*. Berlin: Springer, 2011, pp.133–151.
11. Song C, Qu Z, Blumm N, et al. Limits of predictability in human mobility. *Science* 2010; 327(5968): 1018–1021.

12. Xiao-Yong Y, Xiao-Pu H, Tao Z, et al. Exact solution of the gyration radius of an individual's trajectory for a simplified human regular mobility model. *Chinese Phys Lett* 2011; 28(12): 120506.
13. Barabasi A-L. The origin of bursts and heavy tails in human dynamics. *Nature* 2005; 435(7039): 207–211.
14. Hou L, Pan X, Guo Q, et al. Memory effect of the online user preference. *Sci Rep* 2014; 4: 6560.
15. Pan X, Hou L, Stephen M, et al. Long-term memories in online users' selecting activities. *Phys Lett A* 2014; 378(35): 2591–2596.
16. Malmgren RD, Stouffer DB, Motter AE, et al. A Poissonian explanation for heavy tails in e-mail communication. *P Natl Acad Sci USA* 2008; 105(47): 18153–18158.
17. Zhou Y, Zeng A and Wang W-H. Temporal effects in trend prediction: identifying the most popular nodes in the future. *PLoS ONE* 2015; 10(3): e0120735.
18. Trestian I, Ranjan S, Kuzmanovic A, et al. Measuring serendipity: connecting people, locations and interests in a mobile 3G network. In: *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, Chicago, IL, USA, 4–6 November 2009. New York, NY: ACM Press.
19. Tang H, Liao SS and Sun SX. A prediction framework based on contextual data to support mobile personalized marketing. *Decis Support Syst* 2013; 56: 234–246.
20. Zhong Y, Yuan NJ and Zhong W. You are where you go: inferring demographic attributes from location check-ins. In: *Proceedings of the eighth ACM international conference on web search and data mining*, Shanghai, China, 2–6 February 2015. New York, NY: ACM Press.
21. Van Mieghem P, Blenn N and Doerr C. Lognormal distribution in the digg online social network. *Eur Phys J B* 2011; 83(2): 251–261.
22. Mahanti A, Carlsson N, Arlitt M, et al. A tale of the tails: power-laws in Internet measurements. *IEEE Network* 2013; 27(1): 59–64.
23. Granovetter MS. The strength of weak ties. *Am J Sociol* 1973; 78: 1360–1380.
24. Wang W, Pan L, Yuan N, et al. A comparative analysis of intra-city human mobility by taxi. *Physica A* 2015; 420: 134–147.
25. Zhao Z-D and Zhou T. Empirical analysis of online human dynamics. *Physica A* 2012; 391(11): 3308–3315.
26. Palmisano C, Tuzhilin A and Gorgoglione M. Using context to improve predictive modeling of customers in personalization applications. *IEEE T Knowl Data En* 2008; 20(11): 1535–1549.
27. Székely GJ, Rizzo ML and Bakirov NK. Measuring and testing dependence by correlation of distances. *Ann Stat* 2007; 35(6): 2769–2794.
28. Li R, Zhong W and Zhu L. Feature screening via distance correlation learning. *J Am Stat* A*ssoc* 2012; 107(499): 1129–1139.
29. Rizzo ML and Szekely GJ. *Energy: E-statistics (energy statistics)*. R package version 1.6.2. http://CRAN.R-project.org/package = energy (2014, accessed 12 September 2016).
30. Serrano-Sanchez JA, Martí-Trujillo S, Lera-Navarro A, et al. Associations between screen time and physical activity among Spanish adolescents. *PLoS ONE* 2011; 6(9): e24453.
31. Archer E, Shook RP, Thomas DM, et al. 45-year trends in women's use of time and household management energy expenditure. *PLoS ONE* 2013; 8(2): e56620.
32. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12: 2825–2830.