Computers, Environment and Urban Systems 58 (2016) 97–106

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

# A contextualized and personalized model to predict user interest using location-based social networks

Ming Li [a,*], Günther Sagl [b], Lucy Mburu [a], Hongchao Fan [a]

[a] GIScience Research Group, Institute of Geography, Heidelberg University, Germany
[b] Geoinformation and Environmental Monitoring, Engineering & IT, Carinthia University of Applied Sciences, Austria

ARTICLE INFO

*Article history:*
Received 21 December 2014
Received in revised form 29 March 2016
Accepted 30 March 2016
Available online 23 April 2016

*Keywords:*
User interest
Context-awareness
Personalization
Prediction
Location-based social networks

ABSTRACT

The accurate determination of user interest in terms of geographic information is essential to numerous mobile applications, such as recommender systems and mobile advertising. User interest is greatly influenced by the usage context and varies across individuals; therefore, a user interest model should incorporate these individual needs and propensities. In this paper, we present an approach to model user interest in a contextualized and personalized manner based on location-based social networks. Multinomial logistic regression is employed to quantify the relationship between user interest and usage context at both the aggregate and individual levels. The proposed approach is tested in a real-world application using Foursquare check-ins issued between February and June 2014 in the three major cities of Chicago, Los Angeles and New York. Results demonstrate the capability of the contextualization process for capturing contextual influences on user interest, and that such influences can be observed at a fine-grained scale at the individual level through the personalization process. The proposed approach therefore enables contextualized and personalized estimation of user interest, thereby contributing useful information to follow-up mobile applications.



## 1. Introduction

Thanks to the rapid technological advancements in portable devices and telecommunication, citizens are increasingly exploring the world with their mobile devices at hand every day. The trend of computations shifting from the desktop to the mobile environment has enabled geographic information to emerge as a centerpiece to enhance human–computer interaction (HCI). Mobile users unwittingly create and utilize geographic information when they interact with the real world through their mobile devices, and such interaction often involves making decisions depending on the immediate surroundings, nearby venues, facilities, services, etc. Among the great variety of available geographic information, only a small portion is usually of interest to an individual. Presenting all available information regardless of relevance would inevitably undermine the efficiency of decision making (Reichenbacher, 2005; Li, Sun, & Fan, 2015). This makes it crucial for mobile application developers to acquire precise and personalized knowledge of user interest taking into account their contextual conditions.

Two complexities arise in determining what information will be of interest to which users. The first stems primarily from the dynamic nature of mobile user interests. The interests vary significantly and are influenced by the usage context, which also varies from one person to another. The model of user interest should, thus, take this dynamism into account. The second obstacle regards the identification of the data that will represent user interest in a reliable manner. It is very difficult to determine the exact contexts and interests of mobile users, because both aspects are vague and heterogeneous. Fortunately, nowadays' mobile communication behavior, specifically the technical possibility in combination with users' desire to instantly share information with others via internet platforms, provides potential solutions to aforementioned problems. With the emergence of location-based social networks (LBSNs) such as Twitter and Foursquare, individuals are continuously sharing their current location, engagements, opinions about events or objects, etc. Consequently, a plethora of publicly available data is generated from these platforms, and can potentially reflect the behavior of millions of citizens at a remarkable level of detail (Roick & Heuser, 2013).

The Foursquare platform, as an example, is one of the most popular LBSNs worldwide, with more than 55 million users and over 6 billion check-ins (BrandonGaille, 2015). Every time a user checks-in at a place, a unique data entry, i.e., a check-in record, is created in the Foursquare database with attached contextual information such as the time and location. The location information in Foursquare is not only recorded with geographic coordinates, but also characterized with predefined hierarchical categories. This categorization is one of the outstanding features that distinguish Foursquare from other LBSNs (Cramer, Rost, & Holmquist, 2011), and it provides a special insight into the semantics of the locations and, in turn, the user interest. Hence, the check-in data provides a convenient way to learn how user interest relates to

* Corresponding author.
*E-mail address:* ming.li@geog.uni-heidelberg.de (M. Li).

http://dx.doi.org/10.1016/j.compenvurbsys.2016.03.006
0198-9715/

usage context. The knowledge derived from such data can thereafter be applied in a variety of follow-up applications, such as information retrieval, recommender systems, mobile advertising, etc.

To this end, this paper proposes an approach to uncover the influence of context on user interest at both the aggregate and the individual levels. Two processes are at the core of this approach: the contextualization process applies multinomial logistic regression to all users and quantifies the relationship between user interest and usage context at the aggregate level. The personalization process distinguishes personal differences and quantifies the relationship at the individual level. Together, these two processes have the potential to reveal new insights into user interest and usage context, as well as into the relationships between both.

The rest of the paper is organized as follows. Section 2 presents a brief review of existing studies on context-awareness, prediction and follow-up applications. Section 3 presents the proposed approach with its conceptual framework and main modules. Section 4 tests the approach using Foursquare datasets from three US cities. Finally, Section 5 concludes the work and provides directions for future work.

## 2. Related work

With computation becoming increasingly pervasive, plenty of research has been conducted for the mobile environment in the recent years. Related research directions include, for example, the mobile usage context and context-aware applications (Section 2.1), various kinds of prediction applications (Section 2.2) and user-interest-based applications (Section 2.3).

### 2.1. Context, context-awareness and contextualization: definitions and applications

The notion of context originates from linguistics (Garcia, Duranti, & Goodwin, 1993), and has in the past been considered and defined from many perspectives, depending on the application domains. Among the most frequently cited definitions, we can mention for example the one by Abowd et al. (1999: 306), who define context as “any information that can be used to characterize the situation of an entity”. In the domain of mobile computing, usage context typically includes location, time, user characteristics (activity, mood, etc.), device characteristics (ability, connection, etc.) and information about surroundings (weather, services, etc.) (Nivala & Sarjakoski, 2003; Reichenbacher, 2004). Other than the enumeration of context elements, context is also represented by ontologies (e.g., Wang, Zhang, Gu, & Pung, 2004) or with visual interfaces (Tomaszewski & MacEachren, 2012) to adapt to different application scenarios.

Mobile applications usually need to understand and react to usage contexts in order to better meet the needs of mobile users. This ability is called context-awareness and the process to achieve this ability is contextualization. Many context-aware mobile applications have in the past years been designed for different purposes, contextualized to different types of contexts, and implemented in different ways. Since the creation of mobile context-aware computing by Abowd et al. (1997) with the proposal of Cyberguide, mobile guides have been among the most common context-aware applications. Based on some pre-defined rules, these guides deliver contextualized information in the form of texts (Cheverst, Davies, Mitchell, Friday, & Efstratiou, 2000), images (Lim, 2012), audio data (Chittaro & Burigat, 2005) or maps (Zipf, 2003), and serve in various settings, such as cities (Carlsson, Walden, & Yang, 2008), exhibitions (Oppermann & Specht, 2000) or museums (Ghiani, Paternò, Santoro, & Spano, 2009). Some researchers proposed to adapt not only the information but also the visualization of this information to the context, which led to the emergence of the mobile cartography research field (Reichenbacher, 2004).

The contextualization process used in these applications is typically based on context information collected with sensors (Sagl, Resch, & Blaschke, 2015) and a set of predefined rules in the form of key-value pairs, tagged encoding, ontologies, etc. (see the literature reviews conducted by Chen & Kotz, 2000; Bettini et al., 2010). The rules often involve a complex procedure, including preliminary design, survey activity and follow-up maintenance. In this work, we use a regression model instead of predefined rules for the contextualization process. In this manner, the additional work requiring manual intervention can be replaced by statistical approaches. Moreover, the regression model can provide additional and quantified knowledge regarding how context can influence mobile users, which can benefit many related studies on user behavior.

### 2.2. Predictions for mobile users: targets and algorithms

As noted by Rudin (2012), prediction lies at the heart of almost every scientific discipline, and it is a key topic in machine learning and statistics. From a purely academic point of view, predictions can assist in gaining knowledge since they allow constructing dynamic models that can directly be tested against the set of previous states. Beyond research, predicting is highly useful in a variety of practical situations. In the literature, one can find a large body of prediction research, with different targets and algorithms, and based on different data sources. For example, LBSN datasets have frequently been used in the recent years to predict social ties and links (Cho, Myers, & Leskovec, 2011; Wang, Pedreschi, Song, Giannotti, & Barabasi, 2011), user behavior and mobility (Do & Gatica-Perez, 2012; Preoţiuc-Pietro & Cohn, 2013), user activity (Bart, Zhang, & Hussain, 2013), users' whereabouts (Steiger, Westerholt, Resch, & Zipf, 2015), users' next visiting places (Gambs, Killijian, & del Prado Cortez, 2012; Noulas & Scellato, 2012), etc. Various algorithms have been used for these applications, such as conditional models (Do & Gatica-Perez, 2012), random walk (Noulas, Scellato, Lathia, & Mascolo, 2012), decision tree (Noulas & Scellato, 2012) and Markov chains (Gambs et al., 2012).

In contrast with previous research, we herein intend to predict user interest in terms of geographic information given their usage context using multinomial logistic regression. We use LBSN datasets, and more specifically the Foursquare check-in datasets, to derive user interest, because this knowledge is essential for many follow-up applications that are introduced in the next subsection.

### 2.3. Follow-up applications based on user interest

User interest has been extensively studied for a long time within the context of web search and information retrieval, as far back as the era of the traditional web (Claypool, Brown, Le, & Waseda, 2001; Qiu & Cho, 2006). In the era of mobile web, the demand to fulfill user requirements has become even more urgent. In order to support the decision-making process while avoiding unnecessary distractions, many studies (e.g., Da Costa Pereira, Dragoni, & Pasi, 2012; Leung, Lee, & Lee, 2013) have illustrated the need to deliver the information according to the level of priority of user interest.

Knowledge about user interest is useful in a number of ways. The most common way is to rely on user interests to measure user similarity (Liu, Chen, Xiong, Ding, & Chen, 2012; Gao, Dong, & Fu, 2015), whereby similarity between users is a key component of *collaborative filtering*, a powerful tool to enhance information retrieval. User interest is also essential to recommender systems. Based on the understanding of user interest, researchers have designed and implemented various kinds of personalized systems to recommend venues (Bao, Zheng, & Mokbel, 2012; Noulas et al., 2012; Liu, Liu, Aberer, & Miao, 2013), routes (Kurashima, Iwata, Irie, & Fujimura, 2013) and friends (Chu, Wu, Wang, Chen, & Chen, 2013) to mobile users. Mobile advertising is another application field where knowledge about user interest is critical. Researchers have attempted to design personalized mobile advertising to meet the needs of potential customers (Chen & Hsieh, 2012) and

merchants (Li & Du, 2012), for whom user interest is always taken into serious consideration.

Based on the obvious utility of user interest, we propose an approach to model this information based on LBSN datasets, hoping to provide useful knowledge of contextualized user interest for follow-up applications.

## 3. Modeling user interest with multinomial logistic regression

This section presents the approach to model user interest. The conceptual framework of the proposed approach is first presented, and the main modules of model training, prediction and evaluation, are explained in details throughout the rest of this section.

### 3.1. The conceptual framework

The conceptual framework of the proposed approach is composed of four modules: A) data preparation, B) model training, C) model prediction and D) model evaluation (Fig. 1). Data preparation handles the preprocessing of datasets. Model training is the core module of the proposed approach and includes two steps, namely contextualization and personalization. Model prediction applies the trained model to the predicting dataset, and model evaluation deals with appraisal using predefined statistical indicators.

Modules B to D are explained in the rest of this section, while the data preparation module is explained later in Section 4.1 along with the description of the experimental datasets.

### 3.2. Contextualization

As the first step of the model training phase, contextualization aims to quantify the overall relationship between user interest and usage context at the aggregate level. In this subsection, we introduce the context factors that are used as predictors in the model and explain the steps of the contextualization process.

#### 3.2.1. Context factors as model predictors

As mentioned in the related work section, Foursquare check-in data contain crucial contextual information. Central to the interest of this study are temporal and spatial details that can be extracted respectively from the timestamp and location of each check-in record. More advanced context elements can be obtained using preprocessing approaches. For example, a check-in sequence can be generated for each individual by tracking the check-ins, and then be used to estimate the engagement activities. Other types of context elements, such as mood or accompaniment, are also potentially contained in the datasets in the form of text messages. With some text mining techniques, they could offer additional information about the user's immediate situation.

Within the scope of this paper, we consider three types of context factors. First, the temporal context is represented by the hour and weekday of the check-in. It is implicitly assumed that the interests of mobile users are related to the time of the day or the day of the week. Second, the spatial context is represented by the zip code region where a check-in takes place. It is assumed in this case that zip codes can be linked to the functional configurations of the given region, such as residential, educational, business, etc., that may influence the user interest (Batty,

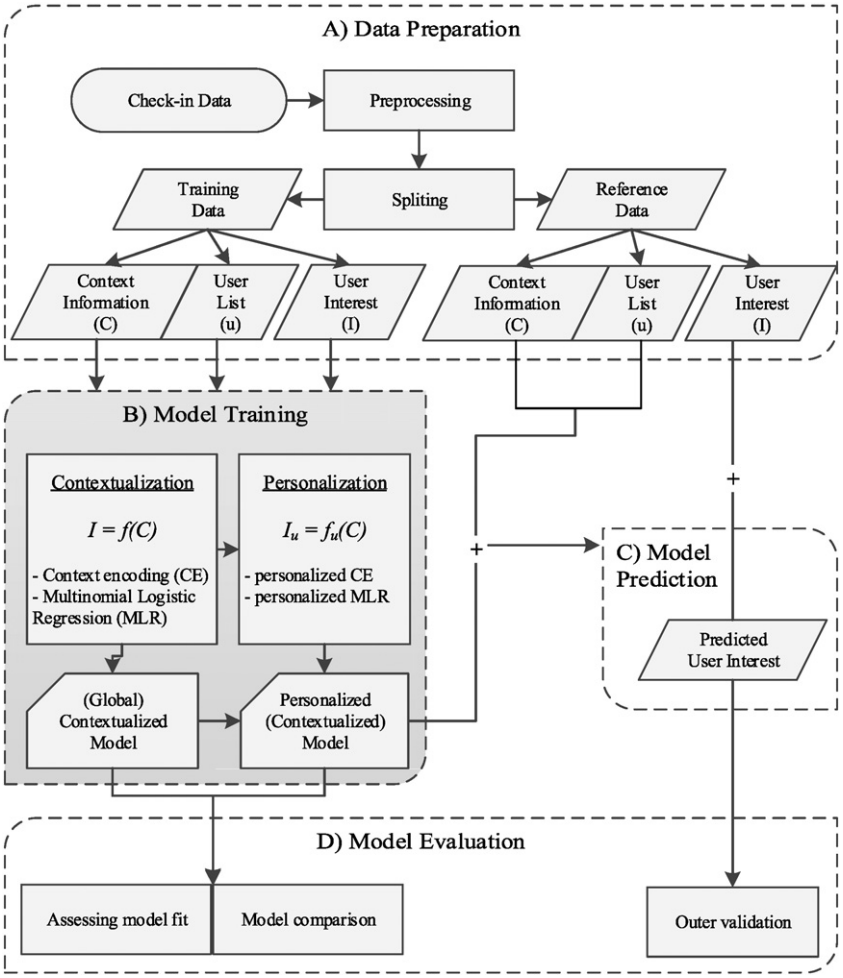


**Fig. 1.** The conceptual framework of the proposed approach.

2008). Then, the sequential context is represented by the previous check-in places extracted from the personal sequences as mentioned above. Here, we adopt the previously proposed assumption (Liu et al., 2013) whereby consecutive check-ins correlate with each other due to the potential involvement of users in activities that are semantically meaningful.

More advanced context elements can be extracted from the dataset, extending beyond the scope of this paper. However, the proposed approach can be extended accordingly to integrate other context elements that can be mined from this dataset.

*3.2.2. Context encoding: dealing with high-cardinal categorical predictors*

One shortcoming of regression models is their inability to deal with high-cardinal categorical predictors, i.e., categorical predictors with a large number of possible values (Zumel, 2012), which are typical to the type of data used in this study. For example, the temporal context might contain up to 168 possible values (24 h times 7 weekdays), while the spatial context might include hundreds of unique zip code values. One solution is to aggregate similar values. For example, because all noontime periods of the five workdays could presumably influence user interest in an identical manner, these temporal context values might be aggregated into a single value. While such solution seems convenient, it can result in significant information loss and sometimes conflicting outcomes, depending on the level at which the aggregation is carried out.

Alternatively, this study adopts the strategy proposed by Micci-Barreca (2001), and transforms all possible values of categorical predictors into continuous numeric values, such that the resulting numeric value corresponds to the impact that the categorical value exerts on the response. In this way, the high-cardinality problem is solved with negligible loss of information. This is the approach that the current study adopts.

In this work, the process of transforming categorical variables is referred to as context encoding (CE). Suppose there exist a context factor, $X$, with all its possible values $\{x_i\}$, and a response outcome denoting the user interest, $Y$, with $J$ possible types of interest $\{y_j\}$. Each $i$th value, $x_i$, is transformed into a vector of numeric values $\{s_{ij}\}$, such that $s_{ij}$ denotes the influence of context type $x_i$ on $y_j$. This value is usually estimated with the posteriori probability of observing $y_j$ given $x_i$, which is formalized as follows:

$$\begin{aligned} &x_i \rightarrow \{s_{ij}\}, j \in [1, J] \\ &s_{ij} \cong \Pr\left(Y = y_j | x_i\right) = \frac{n_{i,Y=y_j}}{n_i}. \end{aligned} \tag{1}$$

As is typical with social media data, the LBSN data has a heavy tailed distribution. This is because the majority of the data is generated by a small percentage of active users, and under certain conditions or contexts. Thus, despite the acquisition of a large dataset, the sample size ($n_i$ in Eq. (1)) could be too small to draw a plausible estimation when a condition ($x_i$) is exerted. To overcome this limitation, the estimation of $s_{ij}$ is slightly modified by introducing the priori probability, $\Pr(Y = y_j)$, with a weighting factor, $\lambda_{n_i}$:

$$\begin{aligned} s_{ij} &= \lambda_{n_i} \Pr\left(Y = y_j | x_i\right) + \left(1 - \lambda_{n_i}\right) \Pr\left(Y = y_j\right) \\ &= \lambda_{n_i} \frac{n_{i,Y=y_j}}{n_i} + \left(1 - \lambda_{n_i}\right) \frac{n_{Y=y_j}}{n_{TR}}. \end{aligned} \tag{2}$$

The weighting factor is a function of the sample size $n_i$, and is bounded between 0 and 1. The rationale of Eq. (2) is that when the sample size is big enough, more credit can be assigned to the posteriori estimation, while in the case of too small sample size, we place more trust on the priori estimation and assume that the conditional knowledge will provide no additional information. The weighting factor can be determined with Eq. (3):

$$\lambda_{n_i} = \frac{1}{1 + e^{-\frac{n_i - k}{f}}}. \tag{3}$$

Eq. (3) represents an S-shaped function that is bounded between 0 and 1. The $k$ parameter determines the number of sample sizes when both the posteriori and priori probabilities take half credits (0.5), and the $f$ parameter controls the slope of the function around the inflection point.

In this way, the context encoding process encodes all possible values of the categorical predictors into numeric values and solves the problem of high-cardinal categorical predictors to enable further modeling. Since the encoding process is conducted at the aggregate level, in the following, we refer to this process as *global context encoding*.

*3.2.3. Multinomial logistic regression: quantifying the relationship*

Multinomial logistic regression has several advantages, such as providing explicit estimates of the variables' explanatory value (Zumel, 2012). In this work, we attempt to quantify the relationship between user interest and contextual conditions. Formally, this relationship can be expressed as follows:

$$I = f(C) \rightarrow Y \sim s_{ij}. \tag{4}$$

Eq. (4) defines the user interest, $I$, as a function of a contextual condition, $C$. This function can be determined through a multinomial logistic regression model by treating user interest as the dependent variable ($Y$) and the encoded contexts ($s_{ij}$) as the explanatory variables. Thus, the user interest can be modeled from the context at the aggregate level. In the following, we refer to this model as the global model.

*3.3. Personalization*

Since user interest does not only vary according to contextual conditions, but also across individuals, the second step is to personalize the user interest.

The globally encoded contexts, $s_{ij}$, which make up the explanatory variables introduced in Eq. (2), denote the average influence that a specific contextual condition $x_i$ exerts on a user interest $y_j$, without considerations of the influence of personal dynamics. Considering that context influences users in different ways, the first step should be to personalize the context encoding. However, the heavy tailed distribution causes many individual users to have insufficient training data for reliable context encoding. Thus, for users with personal training data less than a certain threshold $Z$, the globally encoded value is applied. Otherwise, a personal encoded value is determined using the weighted average of personal posteriori and priori probabilities, as described in Eq. (5):

$$\begin{aligned} &x_{u,i} \rightarrow \{s_{u,ij}\}, j \in [1, J] \\ &s_{u,ij} = \begin{cases} \lambda_{n_{u,i}} \Pr\left(Y = y_j | x_i, u\right) + \left(1 - \lambda_{n_{u,i}}\right) \Pr\left(Y = y_j | u\right), n_u > Z \\ s_{ij}, n_u \le Z \end{cases} \\ &\quad = \begin{cases} \lambda_{n_{u,i}} \dfrac{n_{u,i,Y=y_j}}{n_{u,i}} + \left(1 - \lambda_{n_{u,i}}\right) \dfrac{n_{u,Y=y_j}}{n_u}, n_u > Z \\ \lambda_{n_i} \dfrac{n_{i,Y=y_j}}{n_i} + \left(1 - \lambda_{n_i}\right) \dfrac{n_{Y=y_j}}{n_{TR}}, n_u \le Z \end{cases} \end{aligned} \tag{5}$$

Another common problem with personalized context encoding relates to the missing context, when a certain value of context $x_i^*$ that appears in the predicting data is not present in the training data. In such situations, there is a lack of information to apply Eq. (5) when it comes to the emerging context. As a solution, Eq. (6) represents a weighted average of the prior personal probability and the globally encoded value of the missing context:

$$\begin{aligned} s_{u,i^*j} &= \lambda_{n_u} \Pr\left(Y = y_j | u\right) + (1 - \lambda_{n_u}) s_{i^*j} \\ &= \lambda_{n_u} \frac{n_{u,Y=y_j}}{n_u} + (1 - \lambda_{n_u}) s_{i^*j}, n_u > Z. \end{aligned} \tag{6}$$

The final result of personalization is a hybrid result combining the global and personal models, as described in Eq. (7). In the case of

insufficient personal training data (i.e., $Z$), the personalization step does not change the global model, i.e., $f_u(C) = f(C)$. Otherwise, it results into a personal model.

$$I_u = f_u(C)\begin{cases} \rightarrow Y_u \sim s_{u,ij}, n_u > Z \\ = f(C), \quad n_u \leq Z \end{cases} \tag{7}$$

### 3.4. Model prediction

The model prediction module puts the trained model into practice. The workflow is depicted in Fig. 2.

The predicting dataset can be either with or without a reference. If the predicting dataset corresponds to a true result, the prediction result is evaluated by comparing it to the reference data. In this way, the prediction module can be integrated into the model evaluation module (see Section 3.5). Otherwise, the prediction results only serve as the input for follow-up applications that require knowledge about user interest.

### 3.5. Model evaluation

For model evaluation, we apply a list of statistical indicators that evaluate the model from different perspectives:

- Assessing model fitness: model fitness is perhaps the most important way to evaluate a regression model in the absence of external reference data. Model fitness describes how well a model explains the training data. In this study, we apply four statistics to assess model fitness, namely *correct classification rate* (CCR), deviance, *Akaike information criterion* (AIC) and pseudo $R^2$. CCR assesses the model fitness by evaluating how often the classification result agrees with the true observation of user interest. Deviance assesses the model in absolute scales, while AIC is a relative evaluation of models that provides a means for model selection. As generalizations of $R^2$ in logistic regression, several pseudo $R^2$ have been developed and this paper applies two most common ones: McFadden's $R^2$ and Cox and Snell's $R^2$.
- Model comparison: the likelihood ratio test is used to compare the goodness-of-fit of two models, one of which (the null model) is a special case of the other (the alternative model). It tests how significantly the alternative model improves the null model.
- Outer validation: outer validation is only used when the predicting data has a real reference, and it is validated with the correct prediction rate (CPR). After the trained model is applied to the predicting dataset, a prediction result can be generated. If a reference dataset is at hand for the predicting data, it is possible to compare the two lists and to check how often they agree with each other.

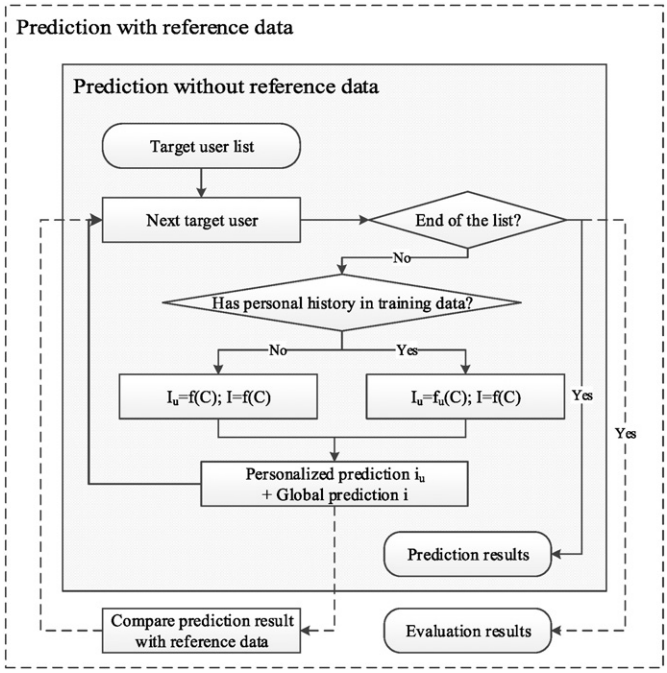


**Fig. 2.** The workflow of prediction module.

## 4. Experimental results and analysis

In this section, we test our approach with three Foursquare check-in datasets. The datasets and their preparation are described in Section 4.1. The subsequent sections present the results of the contextualization and personalization respectively, and Section 4.4 presents the results of prediction module.

### 4.1. The datasets and data preparation

For this study, three datasets were extracted from the Foursquare database. The datasets comprise 167,542, 119,859 and 532,248 check-ins generated respectively by 10,953, 11,893 and 32,184 users from the US cities of Chicago, Los Angeles and New York. The time period spans 150 days, between February 1st and June 30th, 2014.

#### 4.1.1. Data preprocessing

In order to apply the proposed framework to the study data, four preprocessing steps are required:

- Prepare the user interest information: involves matching check-in records to the category hierarchy tree[1] provided by Foursquare and further matching of check-in venue categories to corresponding user interests.
- Prepare the temporal context: involves extracting the hour and weekday from the timestamps of each check-in record. The two elements together constitute the temporal context.
- Prepare the spatial context: involves matching the locations of check-in records with the corresponding administrative zip code regions of the three cities. The zip codes represent the spatial context.
- Prepare the sequential context: involves a more complex extraction of information from the dataset through the following steps:

1) Generate a unique visiting sequence for each user, which order is determined by the check-in timestamps;
2) Attach each check-in record with its previous check-in from the sequence;
3) Concurrently with step 2, keep a record of the time interval between two consecutive check-ins, and compute a weight value for the previous check-in based on the identified time interval. In this study, we apply the equation weight $= 2^{1\text{-interval}/3600}$ so that a shorter time interval is assigned with more weight;
4) Clean the data by a) removing the consecutive check-ins at the same places; and b) resetting the previous check-in to *unknown* if the time interval exceeds a given threshold (e.g., 12 h).

#### 4.1.2. Data splitting

To test our approach, each of the three datasets is split into two parts. The first part (80% of the original dataset) composes the training data. The second part (20%) is the referencing (or predicting) data. We distinguish three types of users, based on the placement in training and referencing data:

[1] https://developer.foursquare.com/categorytree.

- Type A: only appears in the training data (red blocks in the lower User chart of Fig. 3);
- Type B: only appears in the referencing data (purple blocks in the User chart of Fig. 3); and
- Type C: appears in both training and referencing data (green or blue blocks in the User chart of Fig. 3).

Depending on the user types, the check-ins are used in different ways:

- **Case I.** The training check-ins created by users of Type A are only used to build the global model (red blocks in the upper Check-in chart of Fig. 3);
- **Case II.** The training check-ins created by users of Type C are used to build both global and personal models (green blocks in the Check-in chart of Fig. 3);
- **Case III.** The referencing check-ins created by users of Type B are only used for the global model (purple blocks in the Check-in chart of Fig. 3); and
- **Case IV.** The referencing check-ins created by users of Type C are used for both the global and personal models (blue blocks in the Check-in chart of Fig. 3).

The composition of users and check-ins is depicted in Fig. 3.

This organization of the data yields different sub-datasets for multiple uses and enables a representative evaluation of the proposed approach.

### 4.2. Contextualization and evaluation of the global model

As explained in Section 4.1.2, contextualization essentially involves leveraging the training data (Case I and Case II). The process of contextualization, as explained in detail in Section 3.2, generates global models of user interest. To observe the contributions of each explanatory variable (context factor), five models were generated for each dataset to depict the context factors:

- Null model, which uses no context;
- Temporal/Spatial/Sequential model, each of which uses one single context factor; and
- Full model, which uses all three factors.

Comparing the five models allows observing the contribution of each context factor to the user interest. The findings were consistent for the three datasets, and for the sake of conciseness, we present only the evaluation results for the city of Chicago (Table 1 and Table 2).

Table 1 shows the results of the model fitness evaluation with the statistics explained in Section 3.5. It can be gathered that:

- The four last contextualized models were better than the null model (not contextualized) with higher CCR, lower AIC and much higher pseudo $R^2$.
- Among the three investigated context factors, spatial context provided the greatest contribution, because statistics showed the spatial model to be better than the temporal and sequential models.
- The full model including all three predictors emerged as the best model of all with the highest CCR and pseudo $R^2$ as well as the lowest AIC value.

Table 2 compares each of the first four models with the full model. All the likelihood-ratio tests confirm that including all predictive information improves the other models.

Together the evaluation results prove the effectiveness of contextualization and suggest that integrating all three context factors enables a better understanding of user interest. Therefore, the subsequent personalization step is based on the full model with three predictors.

### 4.3. Personalization and evaluation of the personal model

The personalization is carried out by following the approach presented in Section 3.3. As explained in Section 4.1, this process is only applied to the check-in datasets for Case II. Table 3 compares the CCR of global models with that of personal models. According to these statistics, the goodness-of-fit of personal models was much better than that of their global counterparts.

Further analysis reveals the superiority of personal models over global ones in more details (Fig. 4). Again, for the sake of conciseness, we present only the results for the city of Chicago. Fig. 4 (a) through (c) presents the classification table, and the value of a cell with


**Fig. 3.** Data composition after preprocessing and splitting.

**Table 1**
The result of model fitness evaluation with check-in data from Chicago ($n = 167{,}542$).

| Model name | CCR | Deviance | AIC | McFadden.$R^2$ | CoxSnell.$R^2$ |
|---|---|---|---|---|---|
| Null | 0.2832672 | 728,525.0 | 728,543.0 | 0.00000000 | 0.0000000 |
| Temporal | 0.3249672 | 674,125.9 | 674,323.9 | 0.07467021 | 0.2561453 |
| Spatial | 0.3921409 | 608,978.4 | 609,176.4 | 0.16409405 | 0.4781035 |
| Sequential | 0.3045034 | 703,267.7 | 703,447.7 | 0.03466907 | 0.1283694 |
| Full | 0.4468578 | 556,453.5 | 556,993.5 | 0.23619159 | 0.6078066 |

coordinate ($I_x$,$I_y$) indicates the probability of estimating user interest as $I_y$ when the real interest is $I_x$. Thus, the values of all cells sum-up to unity and cells along the diagonal ($I_x = I_y$) denote a correct classification and value of these cells sum-up to CCR. In order to provide a reference, Fig. 4 (c) depicts the case when all classification results are correct. In this case, the values along the diagonal sum-up to unity (CCR = 1).

Returning to the global model (Fig. 4 (a)), cells off the diagonal line (denoting incorrect classification) contained very high probabilities. Consequently, the sum of the correct estimates on the diagonal line amounted to a low overall CCR. For example, there was about 10% probability of wrongly classifying user interest as "Travel and Transport" (Tr&T) when the real user interest was "Food". The probabilities of making wrong classifications drop quickly for the personal model in Fig. 4 (b). Comparing the precision ($|\{I_x = I_y\}|/|\{I_x\}|$) and recall ($|\{I_x = I_y\}|/|\{I_y\}|$) of each type of user interest for both models (Fig. 4 (d)) revealed further that the personal model outperformed the global model in terms of both precision and recall.

Results of Table 3 and Fig. 4 demonstrate that the personalization process can significantly improve the model's goodness-of-fit with regard to classification. The implication of this observation is that due to the individual-level variability of context, the aggregate level is insufficient for depicting finer grained details and must be supplemented by personalization. This is, however, not to disregard the efficiency of contextualization for capturing valuable patterns of user interest at the aggregate level.

### 4.4. Model prediction and outer validation

While the previous sections have focused on the results of evaluating the training data, this section presents the user interest prediction results for the reference data (Cases III and IV). Reference is made to the trained global and personal models (employing data of Case II). Because real users can be obtained from the reference data, further evaluation of the trained models was possible using correct prediction rate (CPR).

Fig. 5 depicts the result of CPR of global and personal models for the three datasets, with results of CCR for both models (see Section 4.3) plotted alongside the CPR estimates to enable comparison.

Estimates of CCR and CPR depicted similar behavior across the global models. Both estimates were averaged around 0.5, and the values began to fall increasingly for users with a personal learning history of beyond 30 check-in records. In contrast, estimates across the personal models behaved differently. CPR values here increased with increasing length of personal training data, while the CCP values decreased. Unlike the CCR values, which were relatively high for the personal models (>0.9), the CPR values were quite low, particularly when the length of personal training data was small. Upon this length being smaller than 12, the CPR value of the personal model became even smaller than that of the global model. This implies that for users with a length of personal learning data shorter than 12, personalization is not reliable.

**Table 3**
CCR comparison of global and personal models.

| Model | Chicago | Los Angeles | New York City |
|---|---|---|---|
| Global | 0.3161 | 0.4441 | 0.4096 |
| Personal | 0.9369 | 0.9645 | 0.9589 |

In this way, the evaluation process provides an empirical ground for setting the value of $Z$ when using the hybrid model: when the length of personal training data is shorter than 12, the global model is more reliable and is therefore recommended for prediction. Only when the length is longer than 12 can one safely apply the personal model.

The comparison of CPR estimates for the global and personal models in separate groups shows results that are similar to the previous observation (see Fig. 6). It explicitly shows that the global model performs better when the length of personal training data is short. With this insight, we adjusted the overall result of the hybrid model. The final CPR of the three datasets based on the proposed approach for Chicago, Los Angeles and New York City is 34.87%, 44.99% and 34.22% for the global model and 47.93%, 51.49% and 44.76% for the personal model.

## 5. Conclusion and future work

In this paper, we have proposed a multinomial logistic regression approach to understand and predict user interest in a contextualized and personalized way. The proposed approach has been tested using Foursquare check-in datasets for Chicago, Los Angeles and New York. The experimental results revealed that the contextualization process with the consideration of three types of context factors (temporal, spatial and sequential) can effectively capture overall information about user interest, but is insufficient for depicting the finer-grained knowledge at the individual scale. For this purpose, the personalization process becomes crucial for enhancing the user interest prediction. In addition, the study results show that even though personalization yields significant model goodness-of-fit, it does not always result in high prediction accuracy when the length of personal training data is too short. Our results further suggest that a personal history that is longer than the empirical threshold value of 12 is sufficient for personalizing user interest. Below this threshold, the use of personal history becomes unreliable. The strength of the proposed approach is the simplicity of its application within the framework of LBSN datasets, such as Foursquare, which are freely available.

In summary, in this paper, we have demonstrated the capability of LBSN datasets for modeling and predicting user interest with three large Foursquare datasets from three American cities. The contribution of this research is twofold. From an academic perspective, it introduces a new idea regarding the utilization of context and the implementation of context-awareness. From the practical perspective, this research can enhance follow-up applications with further knowledge of user interest. To conclude, this research enriches existing studies by adding novel insights into the prediction of user interest using location-based social networks.

**Table 2**
The result of model comparison in pairs.

| Likelihood-ratio test | Model name | Resid. Df | Resid. Dev | Df | LR stat. | Pr(Chi) | Sig. |
|---|---|---|---|---|---|---|---|
| | Full | 1,654,263 | 556,453.5 | | | | |
| Full over null | Null | 1,654,524 | 728,525.0 | 261 | 172,071.48 | 0 | *** |
| Full over temporal | Temporal | 1,654,434 | 674,125.9 | 171 | 117,672.36 | 0 | *** |
| Full over spatial | Spatial | 1,654,434 | 608,978.4 | 171 | 52,524.86 | 0 | *** |
| Full over sequential | Sequential | 1,654,443 | 703,267.7 | 180 | 146,814.19 | 0 | *** |

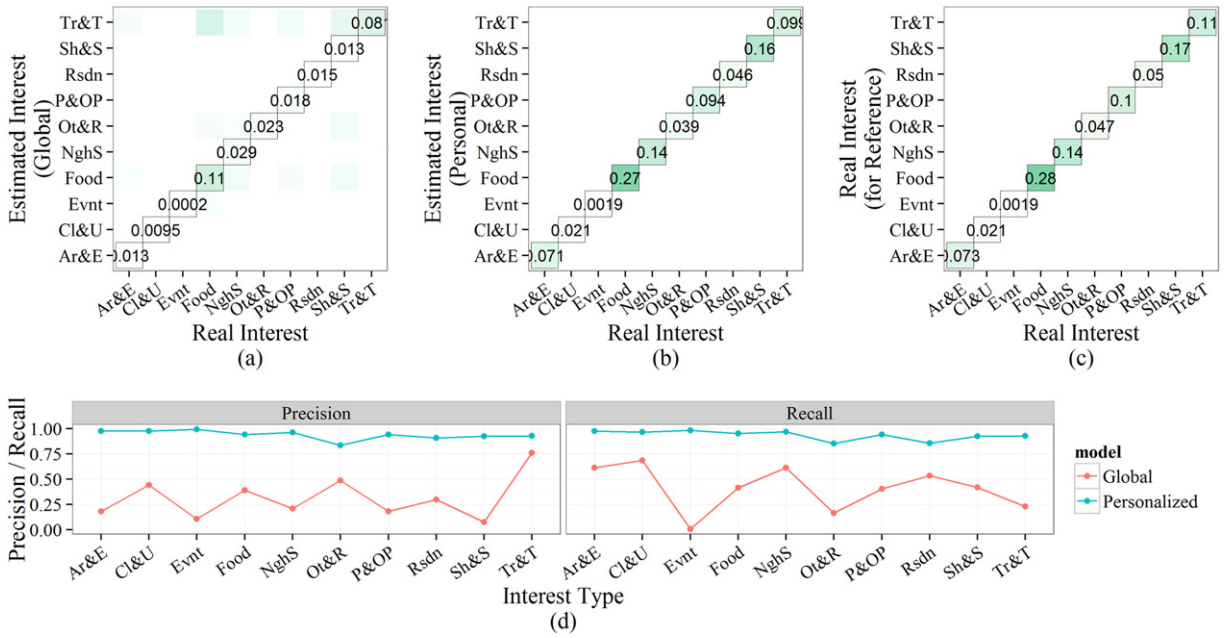


**Fig. 4.** Analysis of the classification results (Chicago).

However, our study has several limitations. The first is with regard to the ongoing debate about the relative representativeness (or lack thereof) of social media datasets (see, e.g., Ruths & Pfeffer, 2014; Tufekci, 2014). A social media platform typically attracts certain groups of users, and the statistical conclusions drawn from its dataset are biased to the respective users. For instance, the users of Foursquare tend to be highly active and mobile youths who are willing to broadcast their daily experiences online (BrandonGaille, 2015). Consequently, the knowledge of the contextualized user interest that is learned from this dataset is biased with respect to these users. Nevertheless, we assume that many follow-up applications tend to be designed for similar kind of users, and this limitation could be negligible from the perspective of the follow-up applications.

In addition, there exist fake or missing check-ins in Foursquare datasets (Cramer et al., 2011). Fake check-ins occur when users check-in at a certain place but do not actually arrive there. In contrast, missing check-ins occur when users do not check-in at the places where they arrived. A common example is “Residence”. Although individuals return home each day, they do not necessarily update this record with Foursquare. The fake and missing check-ins result in a gap between the users' spatiotemporal behavior, as it is observed from the LBSN datasets, and their real behavior, as it is studied by, e.g., *Time Geography* (Miller, 2005). Nevertheless, it is reasonable to assume that one is interested in a place when he/she creates a fake check-in or disinterested in a place when there is a missing check-in. Since this study and many follow-up applications are more concerned with user interest rather than their real trajectories, the knowledge learned from such flawed dataset is still valuable.

Like most spatial observations, the analysis of check-in data is complicated by the regression assumption of independence (Hilbe, 2014). This is because check-in observations are usually clustered in geographic space, and their residuals are often characterized by spatial autocorrelation. Future analysis can incorporate strategies to model the residual spatial autocorrelation, such as eigenvector spatial filtering (Griffith, 2000).

Finally, different representations of contexts could also influence the model and prediction results. For example, this study applies zip code to


**Fig. 5.** CCR and CPR of global and personal models for three datasets.


**Fig. 6.** CPR of the global and personal models for different lengths of personal training data.

represent spatial context. Although straightforward, such representation may fail to capture many local characteristics, since locations are equally treated as long as they fall into the same zip code region. Alternatively, the spatial context might be represented using spatial clusters. However, outcomes directly from spatial clusters are sensitive to the spatial scale and cluster size. In this regard, Andrienko, Andrienko, Mladenov, Mock, & Politz, 2010, for example, has introduced an approach to divide space so that some convex polygons of desired size enclose existing spatial clusters of points. In the future, we are also interested in examining how different representations of context could influence the model and prediction results.

## Acknowledgement

We would like to express our thankfulness to the anonymous reviewers. Their valuable comments contributed greatly towards the improvement of our work. Furthermore, we are grateful to the China Scholarship Council (CSC) for providing the funding for the doctoral studies in GIScience research group of University Heidelberg.

## References

Abowd, G. D., Atkeson, C. G., Hong, J., Long, S., Kooper, R., & Pinkerton, M. (1997). Cyberguide: A mobile context-aware tour guide. *Wireless Networks*, *3*(5), 421–433.

Abowd, G. D., Dey, A. K., Brown, P., Davies, N., Smith, M., & Steggles, P. (1999). Towards a better understanding of context and context-awareness. In H. -W. Gellersen (Ed.), *Handheld and ubiquitous computing* (pp. 304–307). Berlin/Heidelberg: Springer.

Andrienko, G., Andrienko, N., Mladenov, M., Mock, M., & Politz, C. (2010). Discovering bits of place histories from people's activity traces. *2010 IEEE symposium on visual analytics science and technology* (pp. 59–66). IEEE.

Bao, J., Zheng, Y., & Mokbel, M. (2012). Location-based and preference-aware recommendation using sparse geo-social networking data. *Proceedings of the 20th international conference on advances in geographic information systems* (pp. 199–208). ACM.

Bart, E., Zhang, R., & Hussain, M. (2013). Where would you go this weekend? *Time-dependent prediction of user activity using social network data*. Boston, USA: ICWSM.

Batty, M. (2008). The size, scale, and shape of cities. *Science (New York, N.Y.)*, *319*(5864), 769–771.

Bettini, C., Brdiczka, O., Henricksen, K., Indulska, J., Nicklas, D., Ranganathan, A., & Riboni, D. (2010). A survey of context modelling and reasoning techniques. *Pervasive and Mobile Computing*, *6*(2), 161–180.

BrandonGaille (2015). 26 great foursquare demographics [online]. Available from: http://brandongaille.com/26-great-foursquare-demographics/ (Accessed 1 Sep 2015)

Carlsson, C., Walden, P., & Yang, F. (2008). Travel MoCo — A mobile community service for tourists. *2008 7th international conference on mobile business* (pp. 49–58). Ieee.

Chen, G., & Kotz, D. (2000). A survey of context-aware mobile computing research. *Technical report TR2000-381*. Dept. of Computer Science, Dartmouth College.

Chen, P. -T., & Hsieh, H. -P. (2012). Personalized mobile advertising: Its key attributes, trends, and social impact. *Technological Forecasting and Social Change*, *79*(3), 543–557.

Cheverst, K., Davies, N., Mitchell, K., Friday, A., & Efstratiou, C. (2000). Developing a context-aware electronic tourist guide: Some issues and experiences. *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 17–24). The Hague, The Netherlands: ACM.

Chittaro, L., & Burigat, S. (2005). Augmenting audio messages with visual directions in mobile guides: An evaluation of three approaches. *Proceedings of the 7th international conference on human computer interaction with mobile devices & services — MobileHCI '05* (pp. 107). New York, New York, USA: ACM Press.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and mobility: User movement in location-based social networks. *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining — KDD '11* (pp. 1082). New York, New York, USA: ACM Press.

Chu, C. -H., Wu, W. -C., Wang, C. -C., Chen, T. -S., & Chen, J. -J. (2013). Friend recommendation for location-based mobile social networks. *2013 seventh international conference on innovative mobile and internet services in ubiquitous computing* (pp. 365–370). IEEE.

Claypool, M., Brown, D., Le, P., & Waseda, M. (2001). Inferring user interest. *IEEE Internet Computing*, *5*(6), 32–39.

Da Costa Pereira, C., Dragoni, M., & Pasi, G. (2012). Multidimensional relevance: Prioritized aggregation in a personalized information retrieval setting. *Information Processing and Management*, *48*(2), 340–357.

Cramer, H., Rost, M., & Holmquist, L. E. (2011). Performing a check-in: Emerging practices, norms and 'conflicts' in location-sharing using foursquare. *Proceedings of the 13th international conference on human computer interaction with mobile devices and services* (pp. 57–66). ACM.

Do, T. M. T., & Gatica-Perez, D. (2012). Contextual conditional models for smartphone-based human mobility prediction. *Proceedings of the 2012 ACM conference on ubiquitous computing — UbiComp '12* (pp. 163). New York, New York, USA: ACM Press.

Gambs, S., Killijian, M. -O., & del Prado Cortez, M. N. (2012). Next place prediction using mobility Markov chains. *Proceedings of the first workshop on measurement, privacy, and mobility — MPM '12* (pp. 1–6). New York, New York, USA: ACM Press.

Gao, Q., Dong, X., & Fu, D. (2015). A context-aware mobile user behavior based preference neighbor finding approach for personalized information retrieval. *Procedia Computer Science*, *56*, 471–476.

Garcia, A., Duranti, A., & Goodwin, C. (1993). Rethinking context: Language as an interactive phenomenon. *Contemporary Sociology*.

Ghiani, G., Paternò, F., Santoro, C., & Spano, L. D. (2009). UbiCicero: A location-aware, multi-device museum guide. *Interacting with Computers*, *21*(4), 288–303.

Griffith, D. A. (2000). Eigenfunction properties and approximations of selected incidence matrices employed in spatial analyses. *Linear Algebra and its Applications*, *321*(1–3), 95–112.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University Press.

Kurashima, T., Iwata, T., Irie, G., & Fujimura, K. (2013). Travel route recommendation using geotagged photos. *Knowledge and Information Systems*, *37*(1), 37–60.

Leung, K. W. T., Lee, D. L., & Lee, W. C. (2013). PMSE: A personalized mobile search engine. *IEEE Transactions on Knowledge and Data Engineering*, *25*(4), 820–834.

Li, K., & Du, T. C. (2012). Building a targeted mobile advertising system for location-based services. *Decision Support Systems*, *54*(1), 1–8.

Li, M., Sun, Y., & Fan, H. (2015). Contextualized relevance evaluation of geographic information for mobile users in location-based social networks. *ISPRS International Journal of Geo-Information*, *4*(2), 799–814.

Lim, T. Y. (2012). Designing the next generation of mobile tourism application based on situation awareness. *2012 Southeast Asian network of ergonomics societies conference (SEANES)* (pp. 1–7). Ieee.

Liu, Q., Chen, E., Xiong, H., Ding, C. H. Q., & Chen, J. (2012). Enhancing collaborative filtering by user interest expansion via personalized ranking. *IEEE Transactions on Systems Man and Cybernetics Part B, Cybernetics : a publication of the IEEE Systems, Man, and Cybernetics Society*, *42*(1), 218–233.

Liu, X., Liu, Y., Aberer, K., & Miao, C. (2013). Personalized point-of-interest recommendation by mining users' preference transition. *Proceedings of the 22nd ACM international conference on conference on information & knowledge management* (pp. 733–738). San Francisco, CA, USA: ACM.

Micci-Barreca, D. (2001). A preprocessing scheme for high-cardinality categorical attributes in classification and prediction problems. *ACM SIGKDD Explorations Newsletter*, *3*(1), 27.

Miller, H. J. (2005). A measurement theory for time geography. *Geographical Analysis*, *37*(1), 17–45.

Nivala, A. -M. M., & Sarjakoski, L. T. (2003). An approach to intelligent maps: Context awareness. *2nd workshop on'HCI in mobile guides*.

Noulas, A., & Scellato, S. (2012). Mining user mobility features for next place prediction in location-based services. *Data mining (ICDM), 2012 IEEE 12th international conference on* (pp. 1038–1043). Brussels, Belgium: IEEE.

Noulas, A., Scellato, S., Lathia, N., & Mascolo, C. (2012). A random walk around the city: New venue recommendation in location-based social networks. *2012 international conference on privacy, security, risk and trust and 2012 international conference on social computing* (pp. 144–153). Ieee.

Oppermann, R., & Specht, M. (2000). A context-sensitive nomadic exhibition guide. In P. Thomas, & H. -W. Gellersen (Eds.), *Handheld and ubiquitous computing* (pp. 31–54). Berlin Heidelberg: Springer.

Preoţiuc-Pietro, D., & Cohn, T. (2013). Mining user behaviours: A study of check-in patterns in location based social networks. *Proceedings of the 5th annual ACM web science conference on — WebSci '13* (pp. 306–315). New York, New York, USA: ACM Press.

Qiu, F., & Cho, J. (2006). Automatic identification of user interest for personalized search. *Proceedings of the 15th international conference on world wide web* (pp. 727–736). ACM.

Reichenbacher, T. (2004). *Mobile cartography — Adaptive visualization of geographic information on mobile devices*. Technischen Universität München.

Reichenbacher, T. (2005). The importance of being relevant. *Proceedings XXII international cartographic conference A Coruna Spain July* (pp. 1–15). Citeseer.

Roick, O., & Heuser, S. (2013). Location based social networks — Definition, current state of the art and research agenda. *Transactions in GIS*, *17*(5), 763–784.

Rudin, C. (2012). Prediction: Machine learning and statistics [online]. MIT open course. Available from: http://ocw.mit.edu/courses/sloan-school-of-management/15-097-prediction-machine-learning-and-statistics-spring-2012/ (Accessed 19 Aug 2015)

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063–1064.

Sagl, G., Resch, B., & Blaschke, T. (2015). Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities. *Sensors (Basel, Switzerland)*, *15*(7), 17013–17035.

Steiger, E., Westerholt, R., Resch, B., & Zipf, A. (2015). Twitter as an indicator for whereabouts of people? Correlating Twitter with UK census data. *Computers, Environment and Urban Systems*, *54*, 255–265.

Tomaszewski, B., & MacEachren, A. M. (2012). Geovisual analytics to support crisis management: Information foraging for geo-historical context. *Information Visualization*, *11*(4), 339–359.

Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *ICWSM '14: Proceedings of the 8th international AAAI conference on weblogs and social media*.

Wang, D., Pedreschi, D., Song, C., Giannotti, F., & Barabasi, A. -L. (2011). Human mobility, social ties, and link prediction. *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining — KDD '11* (pp. 1100). New York, New York, USA: ACM Press.

Wang, X. H., Zhang, D., Gu, T., & Pung, H. K. (2004). Ontology based context modeling and reasoning using OWL. *IEEE annual conference on pervasive computing and communications workshops, 2004. Proceedings of the second* (pp. 18–22). IEEE.

Zipf, A. (2003). Task oriented map-based mobile tourist guides. International workshop: 'HCI in mobile guides'. *Mobile HCI 2003. Fifth international symposium on human computer interaction with mobile devices and services*. Itlay: Undine.

Zumel, N. (2012). *Modeling trick: Impact coding of categorical variables with many levels | win-vector blog [online]. Win-vector blog.* Available from: http://www.win-vector.com/blog/2012/07/modeling-trick-impact-coding-of-categorical-variables-with-many-levels/ (Accessed 17 Aug 2015)