proceedings

MDPI

Proceedings

Discovering User's Trends and Routines from Location Based Social Networks †

Sergio Salomón [1,*], Rafael Duque [2] and José Luis Montaña [2]

[1] Axpe Consulting Cantabria S.L., 39600 Camargo, Spain
[2] Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, 39005 Santander, Spain; rafael.duque@unican.es (R.D.); joseluis.montana@unican.es (J.L.M.)
* Correspondence: ssalomong@axpecantabria.com
† Presented at the 12th International Conference on Ubiquitous Computing and Ambient Intelligence (UCAmI 2018), Punta Cana, Dominican Republic, 4–7 December 2018.

Published: 30 October 2018

Abstract: Location data is a powerful source of information to discover user's trends and routines. A suitable identification of the user context can be exploited to provide automatically services adapted to the user preferences. In this paper, we define a Dynamic Bayesian Network model and propose a method that processes location annotated data in order to train the model. Finally, our model enables us to predict future location contexts from the user patterns. A case study evaluates the proposal using real-world data of a location-based social network.

Keywords: location based social networks; user modeling; probabilistic graphical models; geolocation

1. Introduction

Context-aware systems identify automatically the user interaction context [1] to provide service adapted to user circumstances. For this purpose, the context-aware systems make use of wireless sensors and networks that monitor user's daily activities in order to have a better knowledge of the user context [2]. A suitable identification of the user context can be exploited to provide automatically services adapted to the user and to reduce the user effort to communicate their requirements [3]. Users also can participate in the task of characterizing their interaction context [4]. Thus, local-based social networks enable users to broadcast assessment of the services and places of their interaction contexts.

The big amount of data collected by sensors and applications like location-based social networks (geolocation, services used, schedules, etc.) allow software developers to build descriptive models of the users and the contexts in which they interact. The amount and variety of information collected make us consider the development of mechanisms that not only build descriptive models of the users but also recommend new services and contexts more adapted to the user preferences.

To approach the development of these mechanisms, this paper proposes a method that follows a process made up of four phases. First, information about the user context is collected by sensors and applications; the method relates raw data extracted by location sensors with semantic labels that describe more in-depth the context of the user. Second, this data is processed to identify multiple user contexts (spatial, temporal and semantic). Next, the annotated data is used to train a dynamic Bayesian network that models the user. Finally, the models allow us to infer and predict future context information of the user.

The remainder of this article is organized as follows. Section 2 reviews the main contributions related to generating user models to be exploited by context-aware systems. Section 3 describes our method for generating user models. Section 4 discusses a case study in which the method has been

applied to processed data of a location-based social network and to generate user models. Section 5 analyses the conclusions drawn from this work.

2. Related Work

One of the main purposes of the context-aware systems [5] is providing features that reduce the user effort to interact with technological devices. For this, the devices embedded in the context should identify automatically the user's requirements. A paradigmatic example can be found in the intelligent information displays that automatically select, filter and show contents adapted to each user (age, preferences, habits, etc.). This descriptive information of the user is known as user model [6]. Thus, the purpose of context-aware systems makes user modeling a relevant task. One of the first user modeling approaches uses a set of attribute-value pairs to describe the user [7]. These attributes are usually related to personal preferences (favorite sports, type of music, etc.). This kind of user models can be processed by interactive systems to provide services and information of interest. This approach can be enriched with case-based techniques in which the system collects information on user behavior [8]. Experts can also enrich the user model with a knowledge base with the main characterizes of specific users [9]. Moreover, an approach based on stereotypes can be used to label user. For this, an expert should define a set of rules that assign a stereotype to each user [10].

Location-based social networks (LBSN) are a recent collaborative phenomenon that combines information on the user activity in a specific context with an online social network [11]. Chorley et al. [12] analyze the behavior of the users of LBSNs to find personality traits like conscientiousness, openness, neuroticism or extroversion. Thus, information registered by LBSNs can be used to build user models made up of personality traits. Users behaviors of LBSNs have been also analyzed to discover their mobility patterns [13]. According to the context-aware system purpose, artificial intelligence techniques can be used not only to generate descriptive models of the user but also to infer new contexts adapted to the user behavior. For example, Eagle et al. [14] learn movement patterns and recognize abnormal behavior through Dynamic Bayesian networks. Hasan and Ukkusuri [15] use Probabilistic Topic Models and contextual information to infer life-style patterns and discover hidden regularity in the location choice behavior. Cho [16] applies k-Nearest Neighbors and Decision Trees to recognize the user location from context features (time and transportation mode) and then predicts future locations with Hidden Markov Models.

3. Methods

We address the problem of learning user models from the individual's routines and trends in LBSN data. In order to do so, we identify different kinds of contextual information in the user interaction. This data requires a preprocessing phase to allow its exploitation. After the preprocessing step, we train user models that represent the user behavior according to a target variable. The proposed models also can be used to generate descriptions according to the user routines. Finally, these models are evaluated by its ability to infer the target value.

3.1. Data Features

LBSNs data are made up of explicit interactions performed by the users. The users of this kind of platforms perform "check-in" actions which register their location. Therefore, these data are collected with a lack of continuity and regularity over time, which does not allow us to extract complete daily behaviors. However, these data enable us to know the context in which the user interacts. This knowledge can be used to predict the users' behaviors.

Usually, the LBSNs provide the next data information: user identifier, geographic coordinates of the visited place, the timestamp in which the user interaction happens, place identifier and some category or type of the place. Sometimes other semantic information is included, but the previously enumerated can be considered the basic setting in every LBSN. In this work, we consider user routines in terms of the place category, which will be our target value.

3.2. Data Preprocessing

In the preprocessing phase, we transform time and spatial information to allow its use in the user models. This will generate new features to characterize the user temporal and spatial context. We also consider semantic context to infer for the problem we address.

For the temporal context, we discretize continuous values of time to obtain different time intervals inside a day. We introduced in [17] different strategies to create time intervals: by fixed size values (e.g., each interval of one-hour size), by means of data distribution (e.g., computing percentiles to obtain intervals with an equally distributed timestamps) or by means of the data density (e.g., applying a clustering algorithm). In this case, we choose a fixed size strategy with 6 intervals of 4 h. We named those intervals, from 00:00 to 23:59, as follows: late night, early morning, morning, afternoon, evening and night. This gives us an intuitive representation of time with the same partition for all users. Along with this interval, we also extract features of day of the week and identification of non-working days or holidays. Thus, we describe the temporal context of the interaction by this three variables (e.g., <*Monday, Working day, Afternoon*>).

In the case of the spatial context, we identify the geographic regions that are frequent for each user. This can be achieved through multiple clustering algorithms (that do not require a predefined number of clusters) as well:

- Density-based clustering (like DBSCAN): the geographic points are grouped in dense groups and the isolated points can be considered as outliers (not belonging to a relevant region for the user).
- Grid-based clustering: the geographic space is partitioned with a grid of fixed size and the cells with few points can be ignored.
- Hierarchical clustering: the partition is built as a hierarchy of groups, which can allow using different levels of spatial granularity, and where the division/agglomeration is limited by a distance threshold.

Depending on the specific application, one strategy may be preferred over the others. Since the data size is usually significant in this problem, and we prefer an approach with homogeneous regions for all users, we choose the grid-based strategy. Then, the spatial context of each interaction is given by its cell in the grid.

Finally, we include in our method a semantic context that could be obtained from multiple sources. This type of context is mainly associated with the user activity or intention, but this information is rarely explicit. Moreover, we can easily assume that the user activity is strongly related to the category of the place the user visits. Therefore, we will infer this context label from the location type using an external service (e.g., as a more general category that groups together several location types: (restaurant, coffee shop, bakery) $\in$ "Food/drinks"). This approach depends on the specific LBSN and could be unavailable, where will be necessary resort to keyword processing in the location type label (for example, with regular expressions). Because of this, we propose as future work the inference of the semantic context as a hidden state or latent variable in the data.

3.3. User Model

We approach a problem in which multiple variables can provide information about the user context. Moreover, these variables are usually related between them. In other words, there are multiple probabilistic dependencies between contexts and attributes (the spatial context depends on the timespan; the semantic context depends on the timespan - and the location, etc.). For these reasons, we generate a Bayesian Network (BN) that models the users' interactions and learns their routines. Figure 1 shows the probabilistic dependencies modeled by the directed acyclic graph that defines the Bayesian network structure.

We are also interested in finding the relationship between the current location or state of the user and their previous locations or states, so we propose the use of Dynamic Bayesian Network (DBN) over the previous Bayesian structure. For this, we considered the type of visited place and the

semantic context as discrete stochastic processes (i.e., the type of visited place depends on the previous place) which fulfills the Markov property. The transition structure represented in Figure 1 models this property and the attributes of the user context.

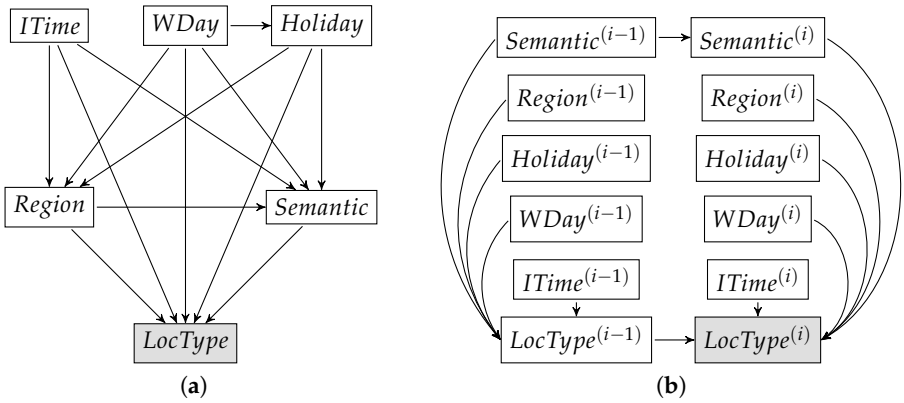

Figure 1. (**a**) Bayesian Network structure; (**b**) Dynamic Bayesian Network transition model.

For the parameter learning, the model distributions are adjusted through Maximum A Posteriori (MAP) and Maximum Likelihood Estimation (MLE) methods. These probabilistic structures have the ability to handle missing data and can be used for user description as well. In the case where some information is missing, the model is capable to infer the most probable values through the remaining variables; i.e., it is possible to give beliefs about unknown context given other context evidence. As description tools, these models provide the measurement of confidence (as used in association rule learning) about any interaction pattern. For example, we could check the likelihood of the user going to the cinema given the time evidence <*Friday, Working day, Evening*>, the current semantic context *Arts/Entertainment* or the past state *Train Station*.

4. Experimentation

We test the methods described in Section 3 on a case study with real-world data to check its performance. In order to do so, we use a *Foursquare* dataset [18] of "check-ins" from the city of Tokyo. This anonymized repository contains 573,703 visits records from over 10 months, and with 2293 different users. For each user interaction, the dataset includes latitude and longitude coordinates, user identifier, the current timestamp, an identifier of the visited location and its location category. In this data, there are 61,858 different locations and 247 different categories.

Through the dataset preprocessing, we infer information of the temporal, spatial and semantic context. First, the features of day of the week, time interval and holiday/working day flag (where only weekends are considered holidays in this case) are extracted to form the temporal context. Next, the geographic regions-of-interest for each user are identified by grid-based segmentation, with rectangular cells of 1 km width and height approximately. In addition, the cells with low density (containing less than 5 % of the user data) are discarded. Finally, it is obtained the semantic context through the Foursquare hierarchy in which all location categories are organized. In order to do so, for each location type is associated its broader parent category as context (e.g., *Afghan restaurant* $\rightarrow$ *Food*, *Mall* $\rightarrow$ *Shop/Service*). In total, there are 9 different context classes: *Arts/Entertainment*, *Travel/Transport*, *Shop/Service*, *Professional/Other*, *Nightlife*, *Food*, *Outdoors/Recreation*, *Academic*, *Residence*.

After the data preprocessing, we evaluate several user models at the task of location type prediction. As the complete DBN model (see Figure 1) is too computationally expensive, we use a more basic case with only semantic context and past state information, as shown in Figure 2. We also defined two more simple Bayesian networks along with the one with all features (BN-A): a Bayesian network with only the temporal context (BN-T) and another one with the spatial and temporal context (BN-ST). In addition of the Bayesian ones, we use other models for benchmarking: random choice of

the user's location types (referred to as RND), the most frequent location type choice (referred to as MODE), a first-order Markov chain model (referred to as MC) and a Probabilistic Finite Automaton (equivalent to a Weighted Finite Automaton) for the transitions between location types depending on the time interval (see [17] for more details about this model).

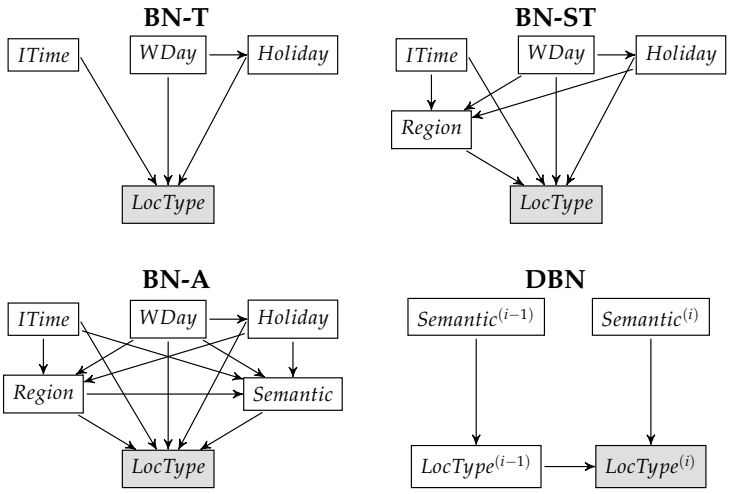

Figure 2. Bayesian models for experimentation.

The models' evaluation is based on the assessment of the efficiency for predicting the type of visited site, using information of the current context and previous states. We divide each user data in 80% as training and 20% as test. The results of this evaluation are shown in Figure 3 and the summary Table 1. Here, we show the distribution of the mean accuracy of all 2293 users for each model.

Table 1. Summary of model accuracies.

Model	Min	Median	Mean	Max
RND	0.0	0.0303	0.0372	0.3913
MODE	0.0	0.3571	0.3829	**1.0**
MC	0.0	0.3729	0.3911	**1.0**
PFA	0.0	0.3714	0.3895	**1.0**
BN-T	0.0	0.3333	0.3531	0.9583
BN-ST	0.0	0.2727	0.3017	0.9545
BN-A	0.0	0.4510	0.4560	0.9600
DBN	**0.0476**	**0.5806**	**0.5816**	**1.0**

From the obtained results, we draw several conclusions. First, it is noticeable a very high variance in the accuracy values (represented as points in Figure 3). This is caused by the heterogeneity between users, as the interactions analyzed are explicit and there are users much more regular or active than others. About the model comparison, we see that clearly all models improve the random prediction accuracy, but not all improve the MODE accuracy. The MODE accuracy, achieving a mean of 0.38, is caused by the frequency of the most common visited place, which is not unusual to be significant in some cases. MC and PFA models improve MODE mean and median accuracy, but it is a small difference. Between the Bayesian models, we see that the ones with only temporal or spatio-temporal context (BN-T and BN-ST) perform poorly, worse than MODE. This can be caused by a low correlation between the place visited and the temporal and spatial context and by underfitting, as the temporal context may not be enough to guess the place and the spatial context is not always present (when the user is not inside a region-of-interest). The most complex model used (BN-A) gets better results than all previous models. In this case, we see how the semantic context leverage the prediction accuracy. This is no surprise, as the semantic context serves as a filter of the state (e.g., if the context is *Food*, the place cannot be *Cinema*). Nevertheless, the reduced DBN model gets the best mean accuracy, 0.58,

in spite of not using temporal and region attributes. From this, we conclude that in this data the context and the type of the previous place are sufficient and more useful to predict the next user state, and spatial or temporal attributes provide a low efficiency.

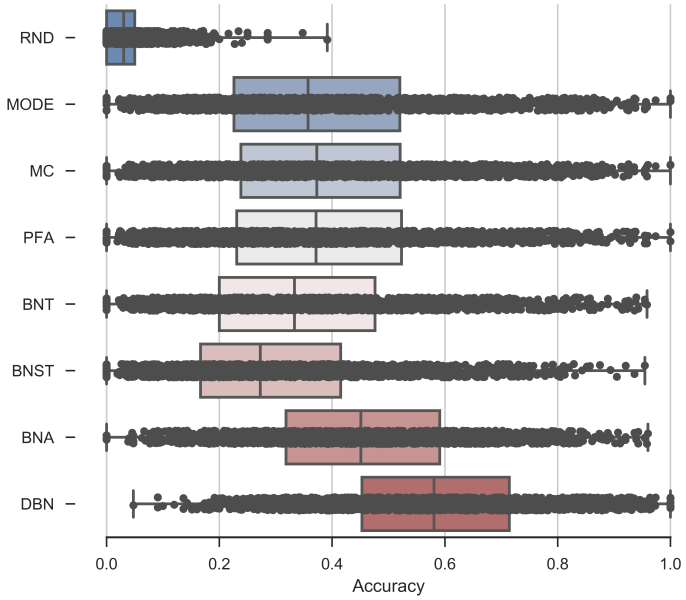

Figure 3. Prediction accuracy of the type of visited location.

5. Conclusions and Future Work

One of the main challenges in the research field of context-aware systems is the identification of contexts to get a better understand of the user and to adapt the services. This paper has presented our methodology to know the user's trends and predict their future behaviors using only the information of their interaction with an LBSN. Therefore, these predictions can be exploited to adapt the services to the user behavior. In particular, we observed that previous state and semantic context features were the most useful information to anticipate the services to the user in our case of study.

In the methodology, we establish some criteria to extract several user context information in order to train a discrete probabilistic model. We design a Bayesian network structure representing the multiple context interaction and a Dynamic Bayesian model capable of anticipating the user context. In addition, the DBN model showed promising results for the user modeling task, improving the performance of Markov Chains, Probabilistic Finite Automata, and Bayesian Networks.

As future work, we will design mechanisms that recognize similar user communities: an explicit (through direct user linkage) or implicit (by user behavior similarity) social context. Moreover, we will infer other types of latent or hidden contexts, such as intentionality. Finally, we also plan to analyze the results considering other user's features as regularity, high or low activity, etc.

Author Contributions: Sergio Salomón has designed the methodology and performed the experiments. Sergio Salomón and Rafael Duque have written the paper. José Luis Montaña have supervised the methodology.

Funding: This research was funded by Fondo Europeo de Desarrollo Regional (FEDER) and Sociedad para el Desarrollo Regional de Cantabria (SODERCAN) grant number TI16-IN-007 (within the program "I+C=+C 2016 - PROYECTOS DE I+D EN EL ÁMBITO DE LAS TIC, LÍNEA SMART"), and by Ministerio de Ciencia e Innovación (MICINN), Spain grant number MTM2014-55262-P (project PAC::LFO).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BN	Bayesian Network
DBN	Dynamic Bayesian Network
LBSN	Location Based Social Network
MC	Markov Chain
PFA	Probabilistic Finite Automaton

References

1. Jaroucheh, Z.; Liu, X.; Smith, S. Recognize contextual situation in pervasive environments using process mining techniques. *J. Ambient Intell. Hum. Comput.* **2011**, *2*, 53–69.
2. Aztiria, A.; Augusto, J.C.; Basagoiti, R.; Izaguirre, A.; Cook, D.J. Learning Frequent Behaviors of the Users in Intelligent Environments. *IEEE Trans. Syst. Man Cybern. Syst.* **2013**, *43*, 1265–1278.
3. Castillejo, E.; Almeida, A.; López-de Ipiña, D. *User, Context and Device Modeling for Adaptive User Interface Systems*; Ubiquitous Computing and Ambient Intelligence. Context-Awareness and Context-Driven Interaction; Urzaiz, G., Ochoa, S.F., Bravo, J., Chen, L.L., Oliveira, J., Eds.; Springer International Publishing: Cham, Switzerland, 2013; pp. 94–101.
4. Evers, C.; Kniewel, R.; Geihs, K.; Schmidt, L. Achieving User Participation for Adaptive Applications. In *Ubiquitous Computing and Ambient Intelligence*; Bravo, J., López-de Ipiña, D., Moya, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 200–207.
5. Weiser, M. Some Computer Science Issues in Ubiquitous Computing. *Commun. ACM* **1993**, *36*, 75–84.
6. Kuflik, T.; Kay, J.; Kummerfeld, B. Challenges and Solutions of Ubiquitous User Modeling. In *Ubiquitous Display Environments*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 7–30.
7. Hanani, U.; Shapira, B.; Shoval, P. Information Filtering: Overview of Issues, Research and Systems. *User Model. User-Adapt. Interact.* **2001**, *11*, 203–259.
8. Leake, D.B. *Case-Based Reasoning: Experiences, Lessons and Future Directions*, 1st ed.; MIT Press: Cambridge, MA, USA, 1996.
9. Burke, R. Knowledge-Based Recommender Systems. In *Encyclopedia of Library and Information Systems*; Marcel, D., Ed.; CRC Press: Boca Raton, FL, USA, 2000; p. 2000.
10. Rich, E. Stereotypes and User Modeling. In *User Models in Dialog Systems*; Kobsa, A., Wahlster, W., Eds.; Springer: Berlin/Heidelberg, Germany, 1989; pp. 35–51.
11. Roick, O.; Heuser, S. Location Based Social Networks—Definition, Current State of the Art and Research Agenda. *Trans. GIS* **2013**, *17*, 763–784.
12. Chorley, M.J.; Whitaker, R.M.; Allen, S.M. Personality and location-based social networks. *Comput. Hum. Behav.* **2015**, *46*, 45–56.
13. Jin, P.J.; Cebelak, M.; Yang, F.; Zhang, J.; Walton, C.M.; Ran, B. Location-Based Social Networking Data: Exploration into Use of Doubly Constrained Gravity Model for Origin–Destination Estimation. *Transp. Res. Rec.* **2014**, *2430*, 72–82.
14. Eagle, N.; Clauset, A.; Quinn, J.A. Location Segmentation, Inference and Prediction for Anticipatory Computing. In Proceedings of the AAAI Spring Symposium: Technosocial Predictive Analytics, Stanford, CA, USA, 23–25 March 2009.
15. Hasan, S.; Ukkusuri, S.V. Location Contexts of User Check-Ins to Model Urban Geo Life-Style Patterns. *PLoS ONE* **2015**, *10*, 1–19.
16. Cho, S.B. Exploiting Machine Learning Techniques for Location Recognition and Prediction with Smartphone Logs. *Neurocomputing* **2016**, *176*, 98–106.

17. Salomón, S.; Tîrnăucă, C.; Duque, R.; Montaña, J.L. Daily Routines Inference Based on Location History. In *Ubiquitous Computing and Ambient Intelligence*; Springer International Publishing: Cham, Switzerland, 2017; pp. 828–839.
18. Yang, D.; Zhang, D.; Zheng, V.W.; Yu, Z. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. *IEEE Trans. Syst. Man Cybern. Syst.* **2015**, *45*, 129–142.