Legal assessment of context prediction techniques

Christian Voigtmann and Klaus David
University of Kassel
Chair of Communication Technology
Email: {voigtmann,david}@uni-kassel.de

Hendrik Skistims and Alexander Roßnagel
University of Kassel
Chair of Public Law
Email: {h.skistims,a.rossnagel}@uni-kassel.de

***Abstract*—Sensors built into current smartphones or in modern facilities are becoming ubiquitous and influence our daily lives more and more. Hence, services that proactively adapt their behaviour to the user's needs is no longer a vision of the future. The proactiveness of such services is achieved by using context prediction algorithms. These algorithms are primarily applied to personal context data of a user, for example to user's motion activities or even to complete daily routines in order to predict a user's next context. Due to this fact the paper presents the most significant legal criteria that have to be considered by the prediction approaches while using personal context data. Following, the criteria are used to legally assess the context prediction approaches. Finally, the resulting consequences are discussed.**

I. Introduction

The collection of user related context information is important for the process of automatic adaption of location-based services. These services try to proactively adapt to users' needs. For example, location-based services that proactively heat up a person's flat or services that interpret device activities within an environment and provide automated assistance.

In order to proactively adapt to future behaviours, places or needs of users, context prediction approaches are utilised. Context prediction seeks to exploit regularities of a person's context history to forecast his or her most probable next context. In most cases, location data of users have been collected and utilised by context prediction approaches. Additionally, location related context data is popular due to the fact that current and future whereabouts of users offer meaningful context information. Further, location information has been often used due to the fact that the location of a person or an object can easily be obtained everywhere, using Wi-Fi and GPS technology or other ubiquitous sensors such as an accelerometer or a gyroscope integrated in modern smartphones. Implementations of location-based context prediction approaches can be found in Smart Home environments [1], [2] where sensors are used to observe and collect persons' behaviours and environmental contexts. Prediction has been also applied in the field of indoor location prediction to forecast the next room a user will enter [3] and in the field of outdoor location prediction [4] e.g., to predict a user's next mobile network cell to infer his or her next possible whereabout.

Location data that has been used in the different application fields to forecast a user's next context mostly represents sensitive personal user data. In most cases third parties could use the collected context information to identify the person the data belongs to. As an example, we can think of recorded GPS coordinates that represent a person's way from his or her home to work. This information taken alone is often sufficient to identify a specific person. If additional time stamped information is added the person can be surely identified.

In the past several research works were published that focussed on legal aspects of context aware techniques such as context prediction. In [5] the authors mention issues related to privacy and security that can be raised during the data acquisition phase of the context prediction process. Issues with regard to context prediction are also discussed in [6]. In addition to important aspects for a context prediction task such as prediction accuracy, fault tolerance, unobstrusive sensor technology for data collection and user acceptance, the author also mentions that legal aspects like privacy still need to be clarified. Further in [7], the German law of informational self-determination and the principle of European legislation have been applied to the process of context prediction and their implications have been examined.

A first discussion of the data protection law in terms of current context prediction approaches is outlined by the authors in [8]. The authors compare a collaborative-based prediction approach to a non collaborative-based prediction approach with respect to data protection concerns. Following this, suggestions to overcome the concerns are outlined. In this paper we examine and assess different state of the art context prediction approaches using the suggestions identified in [8].

The rest of the paper is organised as follows. In the next section, we briefly present the different prediction techniques, which will be evaluated. Then, we introduce the legal evaluation criteria used to assess the different algorithms. Section IV outlines and discusses the evaluation results. Finally, the conclusion is given in Section V.

II. Prediction Techniques

In this section the prediction algorithms that will be assessed from a legal point of view are introduced and it will be briefly discussed how they process user related content.

A. Alignment Predictor

Alignment [9] is a context time series prediction algorithm that is inspired by algorithms with a focus on computational biology and bases on local alignment techniques, such as the Smith and Waterman algorithm. Alignment compares two context sequences. Therefore, it belongs to the branch of

978-1-4673-1881-5/12/$31.00 ©2012 IEEE

pattern matching algorithms. The first sequence represents the context history of the user whose next context has to be predicted. The second sequence represents the current context pattern of the user. During the matching process a pattern in the history will be identified whose similarity to the given current pattern is the highest and therefore results in the lowest penalty costs for a given cost matrix. As a result, the context that follows the identified pattern in the history of the user will be predicted as the next context.

B. Active LeZi

Active LeZi [10] is based upon the Jacob Ziv and Abraham Lempel's LZ78 dictionary-based data compression algorithm that incrementally parses a given input sequence. Active LeZi further extends LZ78 by exploiting all the information in the input sequence using a sliding window. While Active LeZi parses the given history of a user it forms a trie and calculates the probabilities for every possible context transition. The maximum depth of the trie corresponds to the length of the longest context pattern in the history of a user that has been found by Active LeZi. To predict a user's next context the built up trie receives the current pattern as input and calculates the probability for all possible contexts that might follow after the given context pattern sequence. The context with the highest probability will be finally predicted next.

C. State Predictor

The State Predictor method presented in [11] is inspired by branch prediction techniques of microprocessors transformed to handle context prediction tasks. It distinguishes between a 1-state and a 2-state context predictor. A 1-state context predictor only uses the previous state respectively the user's last context information to predict the future context state. In contrast, a 2-state context predictor additionally distinguishes between strong and weak states. If the prediction of a user's next context is correct, then the state predictor switches into the strong state, otherwise it remains in the weak state. If a prediction is incorrect and the state predictor was in a strong state of a context, then it changes to the weak state of the context. If the predictor is already in a weak state of a context and its prediction of a user's next context is incorrect it remains in the weak state of the context.

D. CCP

The idea of the Collaborative Context Predictor (CCP) [12] is not only to be limited to using the user's history whose next context has to be predicted, but also to take advantage of information stored in the histories of other users. This additional information increases the information space and therefore provides predictions even if the user's own context history does not provide suitable information which can be the case if the user changes his or her habits. Instead of only concatenating the additional history information to the user's history, CCP utilises existing direct and indirect relations between different context histories of the users.

E. Machine learning approaches

In addition to the algorithms mentioned above that are specifically developed for context prediction tasks, several techniques from the field of machine learning, have also been used to predict future contexts of a person. The most common techniques are Bayesian Networks or Decision Trees.

1) Bayesian Networks: To anticipate a user's next context Bayesian Networks have been used, for example in [13]. The user's personal context information is represented and stored in nodes using a Bayesian Network. A Bayesian Network is a directed acyclic graph where the edges between the nodes represent dependencies between the different context information. With the help of these dependencies it is possible to calculate the conditional probability of a certain next context given the current context sequence of a user.

2) Tree-based Classifier: A Decision Tree classifier for context prediction is employed in [14]. Context predictors based on Decision Trees, e.g., the C4.5 classifier (which enhances the earlier ID3 algorithm) stores context information in nodes similar to Bayesian Networks. Using these nodes and their corresponding transition probabilities it is possible to find the path in the tree that most probably matches a given context sequence. The leaves of a decision tree represent the possible future contexts.

III. Legal evaluation criteria

The technique, which is unable to cause privacy problems, is the most effective ensuring data protection. The need to enforce law would be reduced.[1] To enhance privacy by design, privacy protection rules should be used to create legal requirements. The context prediction algorithms must ensure compliance with data protection law. Normally for this purpose the separate laws of each country have to be considered. To enlarge the scope to a European range, mainly the data-protection rules of the European Union will be contemplated in the following. These data-protection rules are expressed by several data-protection principles. These principles were first developed by the German Constitutional Court in the final census decision.[2] They can be considered as the key principles of data protection in Europe, since they were all implemented in the Data Protection Directive 95/46/EC [15].

In general most algorithms used for context prediction tasks process personal data. The predictions of contexts are related to an identified or identifiable natural person. The personal information embodied in the prediction concerns the plausibility of each of several behaviors. Therefore context can be described as personal data as defined in Art. 2 (a) of the Data Protection Directive 95/46/EC.

In the following we will outline which steps could be taken, to enhance the user's privacy by design while using each specific algorithm. For the sake of the principle of data

[1]Roßnagel, in: Roßnagel Handbuch Datenschutzrecht, Mnchen 2003, Kap. 3.4, Rn. 47.

[2]Federal Constitutional Court of Germany (Bundesverfassungsgericht). Holdings of the Federal Constitutional Court of Germany (BVerfGE 65, 1), 1983.

minimization, an anonymous or pseudonymous processing of data has to be considered. The anonymization of personal data requires that it is impossible to establish a relation between the information and the affected person. Information that cannot be linked to a person by legal definition cannot violate personal privacy. Despite this desirable goal it is unlikely that anonymization could be implemented in practice. The purpose of context prediction is to support the user on the one hand or to automatically adapt services with regard to given context information on the other hand. Anonymous processing of the context would hinder the support or prevent adaption by the application. This does not mean that the prediction output could not be used in a pseudonymous way. Pseudonymisation refers to replacing the identifiers with pseudonyms known only by the processor. The collected context-information itself could be pseudonymized. This pseudonymization should prevent third parties from reconstructing the behavior of an identified or identifiable natural person. Unfortunately there is no anonymization or pseudonymization that would satisfy legal requirements. Nevertheless this method would hinder conclusions regarding to the user's behavior and hence, enhances privacy. In the following this type of processing will be called content-data pseudonymization.

It is in the user's interest that as little personal data as possible is collected. Even of this collected data, data that has no effect on the context prediction should be erased. This would fulfill the obligation in Art. 6 (e) Data Protection Directive 95/46/EC which requires that personal data must be kept in a form which permits identification of data subjects for no longer than necessary for the purposes data was collected for. Moreover this would meet the principle of necessity of data processes and its binding to a specific purpose, normed in Art. 6 (b) Data Protection Directive 95/46/EC. The reduction of the data volume suggests processing the data on the client side as there is no legal limit to the amount of data that users can keep about themselves. The alternative of processing the data on the server side raises several data protection issues. For example, the external processing of the data requires the transmission of the raw data to the server of the service provider. The context prediction will be done on the server. Subsequently the results of the processing will be transmitted to the device of the user. Unfortunately this external processing raises the possibility of an unauthorized third party gaining access to the data. The service provider may not store the data for purposes other than providing the service to which the user has requested, as any other use of this data would be illegal (unless there is a court order to provide this data to law enforcement or other government agency).

In order to diminish the possibilities of abuse, measures must be taken to fulfill the requirements of the security processing as required in Art. 17 Data Protection Directive 95/46/EC, by providing technical measures to protect personal data against accidental or unlawful destruction.

IV. Evaluation

In the following we apply the data protection issues outlined in the previous section to the algorithms presented in Section II. Further we evaluate the results and discuss the possible consequences for the used prediction approaches.

Table I shows the different prediction approaches and the criteria used to assess these approaches from a legal perspective. If a prediction approach satisfies a legal criterion it gains one point. If an approach partially satisfies a criterion it gains half a point. The maximum score that can be obtained by the examined algorithm is five points. The more points a prediction approach receives the more it satisfies the criteria outlined in Section III. A prediction approach receives a point if it is able to handle pseudonymised context data, if its necessity of data is low and if it utilises only the context data of the person whose next context has to be predicted. Further, an approach receives points if its model can be trained on a person's personal phone due to the fact that its run-time behaviour is low and if it supports indexing to be able to automatically delete context data that is not frequently used. The points that have been received by a prediction approach in total are shown by Σ.

A. Content-data pseudonymization

A content-data pseudonymization results from only pseudonymising the user's context history. This implies that arbitrary placeholders are used to replace the contexts stored in a user's context history. The name of the user the history belongs to will not be replaced. The effects on prediction results while using pseudonymised context data was elaborated using the freely available Augsburger dataset[3]. The prediction models have been trained on the original data and on context histories whose contexts have been replaced by pseudonyms. The number and order of context information was unchanged. The results showed that the prediction results for all approaches remained the same. However, it should be noted that the Augsburger dataset consists of nominal data, which represents a person's current location. The simple replacement of context data through pseudonyms is not possible if the data is stored ordinal or numerically. If that was the case, because context information are represented in GPS coordinates one solution could be to add a global threshold that disguises the data. However, transformed contexts affect the resulting prediction accuracies but this has not been evaluated in this work.

B. Necessity of data

It is a general rule that the greater the amount of context data gathered from a person that can be utilised to train a context prediction approach the more accurate the approach can potentially be. Nevertheless, there are approaches that need a larger amount of training data to be able to make reliable predictions and there are approaches that can work with a smaller amount of training data. With regard to the

[3]http://www.pervasive.jku.at/Research/Context_Database/index.php

TABLE I
LEGAL ASSESSMENT OF DIFFERENT CONTEXT PREDICTION APPROACHES.

	content-data pseudonymization	necessity of data	collaborativeness	data processing	indexing	Σ
Alignment	yes	mid	non	runnable on smartphones	yes	4.5
ActiveLeZi	yes	high	non	not runnable on smartphones	yes	3.0
State Predictor	yes	high	non	runnable on smartphones	yes	4.0
CCP	yes	high	collaborative	runnable on smartphones	yes	3.0
Tree-based	yes	low	non	runnable on smartphones	yes	4.5
Bayesian network	yes	low	non	runnable on smartphones	yes	4.5

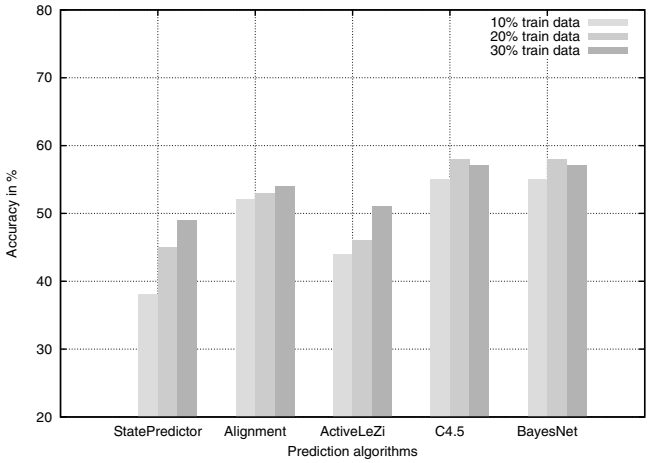

Fig. 1. Testing the necessity of data of the different prediction approaches using small training set to classify a given test dataset.

legal assessment of the prediction approaches those which can achieve higher prediction accuracy on a smaller size of training data, perform better. In order to find the approach which performs best using only a small amount of data to train its prediction model the Augsburger dataset was used again. The different sub datasets of the Augsburger dataset were merged into one dataset. The resulting dataset has been split into a test dataset (10%) and a train dataset (90%). Then three small training data sets with a size of 10%, 20%, and 30% were randomly drawn from the 90% training set. This selection was performed five times for all data set sizes in order to obtain a mean and variance of the result. Afterwards, the classifiers were trained using a small dataset and used to classify the test datasets. Figure 1 presents the averaged results obtained by the different classifiers using the datasets with varying sizes. The results obtained by the Tree-based classifier (C4.5) and by the classifier based on a Bayesian Network are quite similar. Both achieve a prediction accuracy up to 58%. Alignment achieved a prediction accuracy up to 53% and ActiveLezi up to 51%. The most inaccurate classifier is the StatePredictor whose accuracy rate was lower than 50%. CCP was not evaluated. Due to its collaborative character it is only suitable for using multiple context histories, which automatically signifies a high necessity of personal context data. However, if the needed "personal" context data is appropriately anonymized their usage may be lawful.

C. Collaborativeness

Basically, all presented prediction algorithms can be used in a collaborative manner by simply concatenating the context histories of different users. However, the Collaborative Context Prediction (CCP) approach is the only approach that specifically requires context histories of additional users. Consequently, the approach explicitly relies on personal context information of other persons. For this reason the CCP approach has to be assessed more poorly with respect to this criterion than the other prediction approaches that also work on a user's personal context history only.

D. Data processing

From the perspective of law it would be best to process context data directly on the user's smartphone in order to ensure that his or her data keeps private. To elaborate whether the presented prediction approaches are also suited to be executed directly on a smartphone, a benchmark with respect to the following aspects has to be performed: measuring the memory consumption on a smartphone when a prediction approach is trained and used to predict a user's next action directly on a smartphone; measuring the runtime behaviour of a prediction approach, how much time does a single prediction need to be calculated; measuring the implication of a prediction approach that is directly executed on a smartphone to its battery life.

Altogether, such a benchmark has to show a cross section of the three measured aspects to provide a reliable statement how well suited modern smartphones are to be used directly for prediction tasks. In this paper, the suitability of the prediction algorithms to be executed directly on a person's mobile is only discussed theoretically. Basically, the performance aspects of algorithms depend on their way of implementation and on the dataset that has to be processed. The more attributes a dataset has and the more different characteristics each attribute has the more complex the creation of a reliable prediction model will be. Regarding to the used Augsburger dataset all prediction approaches can surely be used directly on a smartphone because each instance (2120 training instances and 200 test instances) consists of four attributes whereas each attribute can have 16 different characteristics. Table II shows the measured averag training times and measured average classification times of the approaches using the Augsburger dataset. If we take into account that current smartphones perform ten times slower than typical PCs the Alignment approach which requires no training phase, the State Predictor, CCP the Tree-based approach and the approach that is based on Bayesian Networks

can run directly on current smartphones. Only the ActiveLeZi approach needs too much time to be trained (build up its trie) on a current smartphone and is therefore inappropriate.

TABLE II
AVERGARED TRAINING AND PREDICTION TIMES OF THE APPROACHES USING THE AUGSBURGER DATASET.

	avg. training time	avg. prediction time
Alignment	n/a	1.935 ms.
ActiveLeZi	20.67 sec.	27.27 ms.
State Predictor	5 ms.	0.025 ms.
CCP	8.68 ms.	91.05 ms.
Tree-based	400 ms.	1.12 ms.
Bayesian Network	100 ms.	0.88 ms.

E. Indexing

In order to store as little personal context data of a person as possible the indexing of context data is considered. The idea is to mark context data that is frequently used by prediction approach to forecast a next context as important. In contrast, context data that is not often used is marked as less important. The easiest way is to implement a counter that remembers how frequently a certain context is used in the prediction process. Using the example of a decision tree the frequency of traversing a certain node or a sub-tree in order to achieve a prediction can be counted. With regard to the algorithms examined in this paper it is possible to implement additional code in all algorithms using a technique called "hooking" to enhance the predictors by providing a counter functionality to additional indexing context information. Subsequently, the indexed context information can be used to weaken certain contexts during the prediction process or to pre-filter less important context information before the prediction process. Table I indicates that all algorithms can support a functionality to index context information.

The evaluation of the different legal aspects shows that Alignment, the Tree-based approach and the approach based on Bayesian Networks receive the highest scores (comp. Table I). That implies that these prediction approaches mostly satisfy the legal aspects demanded in Section III. CCP and the ActiveLeZi approach violate the identified legal criteria the most. The reason for this is the collaborative character of CCP and the high necessity of data of the slow learning time of ActiveLeZi.

V. CONCLUSION

In this paper, we investigated different well-known context prediction techniques with regard to their compatibility to the current Data Protection Directive. Based on this directive we first identified several legal criteria. Subsequently, the context prediction approaches were evaluated against these criteria. The evaluation shows that the Alignment predictor and predictors based on a Decision Tree and on a Bayesian Network most likely fulfill the required criteria. The CCP and ActiveLeZi approach most violate the legal criteria.

ACKNOWLEDGMENT

The authors are involved in the VENUS research project. VENUS is a research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University. We thank Hesse's Ministry of Higher Education, Research, and the Arts for funding the project as part of the research funding program "LOEWE - Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz". For further information, please visit: http://www.iteg.uni-kassel.de/venus.

This work was partially performed by the project SEAM4US "Sustainable Energy mAnageMent for Underground Stations" (FP7-ICT, EEB-ICT-2011.6.4), which is partly funded by the European Commission. The authors would like to acknowledge the contributions of their colleagues.

REFERENCES

[1] M. Mozer, "The neural network house: An environment that adapts to its inhabitants," *Proceedings of the American Association for Artificial Intelligence*, 1998.

[2] D. Cook, M. Youngblood, I. Heierman, E.O., K. Gopalratnam, S. Rao, A. Litvin, and F. Khawaja, "MavHome: an agent-based smart home," in *Pervasive Computing and Communications, 2003. (PerCom 2003). Proceedings of the First IEEE International Conference on*, Mar. 2003, pp. 521–524.

[3] D. Kelly, R. Behan, R. Villing, and S. McLoone, "Computationally tractable location estimation on WiFi enabled mobile phones," in *Signals and Systems Conference (ISSC 2009), IET Irish*, Jun. 2009, pp. 1–6.

[4] S. Scellato, M. Musolesi, C. Mascolo, V. Latora, and A. T. Campbell, "NextPlace: a spatio-temporal prediction framework for pervasive systems," in *Pervasive*, 2011, pp. 152–169.

[5] P. Nurmi, M. Martin, and J. A. Flanagan, "Enabling proactiveness through context prediction," in *In: CAPS 2005, Workshop on Context Awareness for Proactive Systems. (2005*, 2005.

[6] R. Mayrhofer, "Context prediction based on context histories: Expected benefits, issues and current state-of-the-art," *COGNITIVE SCIENCE RESEARCH PAPER-UNIVERSITY OF SUSSEX CSRP*, vol. 577, p. 31, 2005.

[7] C. Voigtmann, J. Zirfas, H. Skistims, K. David, and A. Roßnagel, "Prospects for context prediction despite the principle of informational Self-Determination." IEEE, 2010.

[8] H. Skistims, C. Voigtmann, A. Roßnagel, and K. David, "Datenschutzgerechte gestaltung von kontextvorhersagenden algorithmen," vol. 36, 2012, pp. 31–36.

[9] S. Sigg, S. Haseloff, and K. David, "An alignment approach for context prediction tasks in ubicomp environments," *Pervasive Computing, IEEE*, vol. 9, no. 4, pp. 90–97, october-december 2010.

[10] K. Gopalratnam and D. J., "Active lezi: An incremental parsing algorithm for sequential prediction," in *In Sixteenth International Florida Artificial Intelligence Research Society Conference*, 2003, pp. 38–42.

[11] J. Petzold, F. Bagci, W. Trumler, and T. Ungerer, "Global and local state context prediction," in *In Artificial Intelligence in Mobile Systems 2003 (AIMS 2003), Seattle, WA, USA*, 2003.

[12] C. Voigtmann, S. L. Lau, and K. David, "An approach to collaborative context prediction," in *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2011 IEEE International Conference on*, march 2011, pp. 438–443.

[13] S. Lee and K. C. Lee, "Context-prediction performance by a dynamic bayesian network: Emphasis on location prediction in ubiquitous decision support environment," *Expert Syst. Appl.*, vol. 39, no. 5, pp. 4908–4914, Apr. 2012.

[14] H. E. BYUN and K. CHEVERST, "Utilizing context history to provide dynamic adaptations," *Applied Artificial Intelligence*, vol. 18, no. 6, pp. 533–548, 2004.

[15] G. Hornung and C. Schnabel, "Data protection in germany: The population census decision and the right to informational self-determination," vol. 25, no. 1, pp. 84–88, 2009.