Expert Systems With Applications 64 (2016) 500–511

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Using context for online customer re-identification

U. Panniello [a,*], S. Hill [b], M. Gorgoglione [a]

[a] Politecnico di Bari, Department of Mathematics, Mechanical and Business Engineering, Viale Japigia 182 – 70126 Bari (Italy)
[b] Operations and Information Management Department, The Wharton School of the University of Pennsylvania, 3730 Walnut Street, Suite 500, Philadelphia, PA 19104 (USA)

ARTICLE INFO

Article history:
Received 25 December 2015
Revised 24 March 2016
Accepted 1 August 2016
Available online 5 August 2016

Keywords:
Customer identification
Context
Customer modeling

ABSTRACT

The task of re-identification consists of linking records for individuals with no identifying information to records with identifying information (i.e., name or social security number) in order to identify individuals within the anonymous data. This task is important for business since firms want to precisely identify consumers for several reasons, such as targeting advertisements to them or labeling them as fraudulent users. For these reasons, companies strive to improve their re-identification techniques. In addition, the re-identification task is relevant from a research prospective, and many algorithms and techniques have been proposed to improve existing re-identification models. However, no previous research has studied whether the use of contextual variables can improve re-identification performance. Context can be defined as the circumstances under which transactions take place. To date, contextual information (i.e., the time of day when or the location where digital data was created) has been used successfully in other modeling tasks such as in the recommender system domain, where its ability to improve the accuracy of lists of items suggested to website users has been demonstrated. Including contextual information in a re-identification model is not a trivial task for several reasons. In this paper, we discuss the main issues regarding the use of context for the re-identification task, namely, when incorporating context is expected to help re-identification and when it is expected to hurt. We propose contextual re-identification models and a framework for deciding when to use these and determining the best performing contextual method for the re-identification task. We test our contextual models using three different case studies. Our findings have a significant impact on expert and intelligent systems since they provide the first evidence of the possibility of including contextual variables for improving the results of the re-identification process. The results also have a relevant impact for businesses since they can help managers decide when and how to include a contextual variable into the re-identification task and contextualize subsequent actions after the re-identification task.

1. Introduction

People express many different aspects of their lives online, and it is common to segregate different aspects in different places. People may write movie reviews on their blog under a pseudonym while participating in a scholarly forum under their real name. It is possible to match these separate identities by using a re-identification method, which is the process of linking an individual's prior and current actions or characteristics to a specific identity. Re-identification methods can be used when data associated with one set of actions is purposely incorrect or missing (Delgado, Ros, & Vila, 2009).

Algorithms for re-identification have been effective for a number of data sources in important domains (Maiorana, 2010), including using DNA sequences to identify patients (Malin & Sweeney, 2001), email addresses for alias detection (Holzer, Malin, & Sweeney, 2005), IP addresses to identify online consumers (Malin, 2005), citations for author attribution (Bagavandas, Hameed, & Manimannan, 2009), and phone calls to identify fraudulent consumers (Hilas & Sahalos, 2005). We focus on identifying customers in three different domains, namely, e-commerce purchasing, phone networks, and web navigation.

The re-identification task belongs to the customer modeling research stream, where researchers have developed effective algorithms for a wide range of predictive tasks (Konstan & Riedl, 2012; Mendonça, Sousa, & Sá da Costa, 2012; Van Vlasselaer et al., 2015; Zheng, et al., 2015) using different types of available data. In particular, recent research has studied the significance of the context in which a transaction occurs when building customer behavior

* Corresponding author.
E-mail addresses: umberto.panniello@poliba.it (U. Panniello), shawndra@wharton.upenn.edu (S. Hill), michele.gorgoglione@poliba.it (M. Gorgoglione).

http://dx.doi.org/10.1016/j.eswa.2016.08.004
0957-4174/

predictive models, such as recommending items to users (Panniello & Gorgoglione, 2012; Panniello, Tuzhilin, & Gorgoglione, 2014), and demonstrated that using contextual knowledge results in better performance (Palmisano, Tuzhilin, & Gorgoglione, 2008). Although the beneficial effects of using a contextual variable in a predictive model have been widely demonstrated in several modeling algorithms, few works have studied the effect of context within the re-identification task. In particular, Ding, Meng, Chai, and Tang, (2011) used contextual variables to re-identify users in the web navigation domain albeit only to define initial conditions and demonstrate how these and different approaches influence results. No contextual algorithms for re-identification have been proposed previously because the inclusion of contextual information in a re-identification algorithm is not a trivial task. On one hand, labeling a user's transactions or behaviors by context creates overlap with prior behavior in a specific context, which should make it easier to predict that customer's identity. On the other hand, the behavior of a customer in a given context may be similar to that of another customer, thus increasing overlap with other consumers and making it more difficult to distinguish matching from non-matching consumer behavior.

This research aims to investigate if inclusion in a customer behavior model of the context in which a transaction takes place increases re-identification performance. We compare contextual and un-contextual models using three datasets and varied settings such as amount of data per user, classification algorithms, and granularity of the contextual variable. Finally, we propose a conceptual framework for determining when to incorporate context into modeling for re-identification and selecting the best performing contextual method for the re-identification task.

2. Prior work

Re-identification detects the presence of the same individual in different data sources or across different time periods and is the process of linking a user's prior and current actions or characteristics to one identity. The process plays a central role in many fields. Re-identification in the online shopping domain is important because firms are increasingly striving for effective targeted marketing and personalization. While general targeted marketing aims to effectively segment customers for classification, re-identification seeks to uniquely identify customers for one-on-one targeted marketing (Schafer, Konstan, & Riedl, 1999). The results of re-identification can be used to schedule automatic marketing actions defined by personalization tools such as recommender systems. Rawat, Nayak, and Li, (2011), Barla, Tvarozek, and Bielikova, (2009) and Yang, (2010) proposed the use of re-identification models to improve the results of the recommendation task. Lourenco and Belo, (2011) proposed an approach to usage profiling based on clickstream data to build predictive models that can identify user profiles and prevent potential site hazards.

Re-identification is also crucial in web-surfing applications. Bianco, Mardente, Mellia, Munaf\, and Muscariello, (2009), Huang, Peng, An, and Schuurmans, (2004), and Herrmann, Gerber, Banse, and Federrath, (2012) proposed techniques to re-identify web-user sessions while Rao, Kumari, and Raju, (2010) proposed an effective pre-processing approach to improve web-user re-identification.

Several re-identification algorithms have also been proposed in the domain of phone networks (Kianmehr & Alhajj, 2009). Identifying fraudulent consumers is important in the telecommunications industry because repetitive defaulters use fake identities to obscure their true identity (Hill & Nagle, 2009). Ding et al., (2011) proposed a model to determine a user's identity based on instant messages while Kwapisz, Weiss, and Moore, (2010) described and evaluated a system that uses phone-based acceleration sensors to identify and authenticate cell-phone users.

The most successful re-identification methods are probabilistic record linkage and distance-based record linkage (Domingo-Ferrer & Torra, 2002). In both cases, each record corresponds to a customer profile, and available information about the customer, such as demographics, psychographics, and browsing and purchasing history, is collected from multiple sources, processed, and transformed into actionable knowledge stored in profiles. Recently, scholars have observed that the task of re-identification can be approached by building appropriate probabilistic models of customers rather than computing the similarity between their profiles (Yang & Padmanabhan, 2010). More generally, the re-identification task belongs to the customer modeling research stream, where researchers have developed effective algorithms for a wide range of predictive tasks using different types of available data. In particular, recent research has studied the significance of the context in which a transaction occurs when building predictive models for customer behavior (Hong, Suh, & Kim, 2009), such as those for recommending items to users (Champiri, Shahamiri, & Salim, 2015; Panniello & Gorgoglione, 2012; Panniello et al., 2014) or segmenting users (Faraone, Gorgoglione, Palmisano, & Panniello, 2012; Lombardi, Gorgoglione, & Panniello, 2013), and demonstrated that using contextual knowledge results in better performance (Broelemann, Jiang, & Schwering, 2016; Palmisano et al., 2008).

Many interpretations of context have emerged in diverse fields such as psychology, philosophy, and computer science. Context can be defined as information used to characterize and interpret the situation in which a user interacts with an application at a certain time (Bazire & Brézillon, 2005). Others (Schmidt, Beigl, & Hans-W, 1999) have defined context as a key issue in interaction between humans and computers that describes the surrounding facts that add meaning to the interaction. Scholars have observed that the key to delivering successful solutions in e-commerce has shifted from content to context (Rayport & Sviokla, 1994). In the data mining community, context is defined as the events that characterize the lives of customers and can determine a change in their preferences or status and accordingly affect customers' value to a company (Berry, 1997). Context-aware systems literature defines context as the location of a user, the identity of people near the user, the objects surrounding the user, and the changes in these elements (Schilit & Theimer, 1994). While the beneficial effects of using a contextual variable in predictive models have been widely demonstrated in several modeling algorithms (Bettini et al., 2010; Kaschek, Schewe, Thalheim, & Zhang, 2004; Zimmermann, 2003), few works have studied the effect of context within the re-identification task. In particular, Ding et al., (2011) used contextual variables to re-identify users in the web-navigation domain albeit only to define initial conditions and demonstrate how these as well as different approaches influence results. We aim at filling this gap; in particular, we investigate whether the inclusion of a contextual variable in re-identification models results in better performance. To do so, we propose contextual re-identification models and compare them with a re-identification model that does not include contextual information. Finally, we propose a conceptual framework for deciding when to use a contextual re-identification model.

3. Case studies

To explore our research issue, we used three datasets from three different domains. The first dataset (DSet1) was taken from the study described in Palmisano et al., (2008). The dataset was collected from a special-purpose browser developed for users to navigate Amazon.com in real-time and simulate purchases of products. A group of students was asked to simulate purchases over four months. Once a product was selected for purchase, the browser recorded the selected item, its price, and other charac-