A Framework for Identification and Monitoring of Profiles and Behaviors of Users Based on Mobile App Usage

Nielsen Luiz Rechia Machado[1], **Duncan Dubugras Alcoba Ruiz**[1]

[1]School of Technology, Pontifical Catholic University of Rio Grande do Sul - PUCRS
Computer Science Ph.D. Program, 90619-900 – Porto Alegre – RS – Brazil

nielsen.machado@acad.pucrs.br, duncan.ruiz@pucrs.br

Level: Ph.D.
Enrollment in the program: March 2015
Proposal's defense date: July 2017
Conclusion expected to: March 2019
Concluded steps: Credits; Bibliographic revision; Qualifying exam; Implementation of first Framework prototype, Analyses of experimental results;
Future steps: Improve the framework; Perform new experimental results, Evaluate the framework, Write thesis; Ph.D. Defense.
Publications: [Machado and Ruiz 2017]

Resumo. *Nos último anos, o uso de dispositivos móveis, bem como de seus aplicativos (apps) cresceu significativamente. Além disso, a disputa de mercado faz com que empresas deste ramo foquem na fidelização de seus clientes. Estes clientes realizam diariamente muitas atividades por meio de apps, o que gera, em tempo real, uma grande quantidade de eventos. Diante disso, a identificação e o monitoramento de perfis e comportamento de tais clientes podem contribuir para minimizar situações de risco, como a perda destes clientes. Assim, esta pesquisa propõe um framework para identificação e monitoramento de perfis e comportamentos de uso de apps em dispositivos móveis por usuários, visando segmentá-los em perfis de uso, para identificar comportamentos infrequentes que representem situações de risco para fabricantes de dispositivos móveis.*

Abstract. *Over the last years, the use of mobile devices and applications (apps) significantly grow. In addition, the technological innovation and fierce dispute to conquer the mobile market make companies increase their attention to their clients' loyalty. These clients perform daily many activities through apps generating a large number of events in real time. In this sense, the identification and monitoring of profiles and behavior of these clients can contribute to minimize risk situations, such as turnover. Therefore, this study proposes a new framework for the identification and monitoring of profiles and behaviors based on users' app usage. It aims to segment users into profiles seeking to identify infrequent behaviors that could represent risk situations for mobile device manufacturers.*

1. Introduction and Background

Mobile phones have evolved from simple communication devices to dynamic tools that provide advanced functionality to assist users in their daily activities [Hamka et al. 2014].

People use several applications (*apps*) for a variety of goals, such as read books, watch videos, take pictures, and so on. The number of apps available between 2012 and 2017 grew five times (i.e., from 675,000 to 3,300,000)[1]. The considerable amount of data issued by the use of apps can be categorized as a Data Stream (DS). Gama [Gama 2010] defines a DS as stochastic processes in which events occur continuously and independently of each other. Such data usually have large Volume (Big Data), large Variety (different kinds of data), and are produced at high Velocity (ongoing and in real time). Further, it should be collected from a considerable amount of users in real time. In a DS scenario, it is important to investigate the changes in the data distribution, namely Concept Drift. Indeed, new concepts arise and known concepts may evolve or disappear, which are discovered applying Novelty Detection techniques [Gama 2010]. In this sense, an effective tool, able to perform information analysis and capable of helping researchers and companies to extract knowledge about these app DSs, is almost mandatory.

Advances in mobile device industries, such as new services and technologies, as well as advances in the areas of data mining and machine learning, increased the competition in the market. Mobile device companies seek to keep their customers loyal and engaged because they know that the cost of attracting a new customer is six times greater than the amount spent to retain the old customers. Thus, these companies are under intense pressure to identify and monitor customers' behaviors. Since customers are the "source material", and the market is saturated, new methods for the management of these customers are vital for the survival and development of such companies [Almana et al. 2014]. However, tasks to identify app usage profiles and to monitor customers' behaviors are difficult. Moreover, it is still difficult to access data from mobile devices as app usage. The first mobile datasets were made available only in the last few years [Wagner et al. 2013]. Nevertheless, even considering the main datasets for extracting information from customers, such data have few or no detailed information of app usage and are usually private or not available due to privacy-preserving policies.

The main objective of this research is *to develop a framework for identification and monitoring of profiles and behaviors of users based on app usage on mobile devices*. We aim to segment the users into usage profiles seeking to identify infrequent behaviors that could represent risk situations for mobile device manufacturers. Specifically, our proposed framework is designed to (a) deal with different types of data from mobile app usage, (b) identify usage patterns, (c) group users in usage profiles, and (d) monitor profiles and behaviors of users aiming to segment them according to their variations through time. In a Computer Science view, we are developing an application for mining and monitoring app usage DSs that may have an impact on the mobile device industry.

2. Related Work

We carried out a Systematic Review [Kitchenham 2004] of literature to investigate studies seeking to identification and the monitoring of usage profiles. This review explores 20 related work to our research and is under review by the journal Wiley Interdisciplinary Reviews. Some studies seek to the analysis of most used apps in several contexts [Xu et al. 2011, Li et al. 2015]. However, such works do not aim to identify or monitor profiles and behaviors of users over time. Other studies have proposals for predic-

[1]Apps available in leading app stores - *goo.gl/3FLn4f*

ting users likely to churn (i.e turnover) [Rehman and Raza Ali 2014, Backiel et al. 2016]. Since the available data are in a batch fashion and the users are considered as independent individuals, most of these works consider the churn prediction task as a classification problem. Thus, usage profiles are not investigated as well as profile monitoring is not performed. These studies address different types of data, such as call data records (CDR), billing data and personal information (i.e. age and gender), not using app usage data. On the other hand, some studies propose the identification of usage profiles, which is carried out in several areas, such as Communication [Pyo et al. 2015] and Mobile [Rehman and Raza Ali 2014, Hamka et al. 2014]. Some of the studies from the mobile area do not use app usage data [Rehman and Raza Ali 2014] while others complement their datasets with this data [Hamka et al. 2014] but investigate only the number of apps used by users. Finally, in the last decade, some approaches were proposed for DS scenarios seeking to the identification and monitoring of clusters (i.e. profiles) on several areas [Spiliopoulou et al. 2006, Oliveira and Gama 2010]. However, such works are only intended to analyze changes between the clusters and not the behaviors of the objects (i.e users). We found a single study aiming to monitor usage profiles using a mobile DS [Pereira and Mendes-Moreira 2016]. However, such work does not use app data, applying only CDR.

3. Proposed Framework

In a real-world scenario, users perform a set of activities through a mobile device. Figure 1 (a) demonstrates the apps being run in the foreground by the user over time for three different devices. Each data event corresponds to an activity performed by a single user through the use of an app on a mobile device. Hence, a mobile DS contains million of activities or app events, which are produced from thousands of devices using one of the thousands of apps available in different time spans. A single activity, such as the app *WhatsApp*, is carried out repeatedly over time and such activities may be performed by several users and by different amounts of usage time (e.g. seconds or minutes). On the other hand, activities may or may not be carried out in the same time window (e.g., day or week) as shown in Figure 1 (b). According to the used time window, more or fewer events are processed and summarized. Thus, we can not consider each event as a complete representation of an independent object (device). It is how traditional data stream algorithms interpret them [Gama 2010].

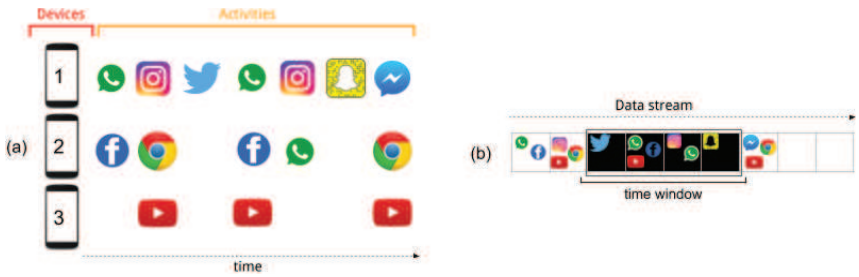

Figura 1. App Data Stream overview.

In this sense, the proposed framework aims to: collect app usage patterns, provide a limited number of usage profiles, and facilitate the monitoring of profiles and behaviors of users in a real-world scenario. To this end, our framework is composed by two main steps, namely *Data Stream Mining* and *Data Stream Monitoring* (see Figure 2).

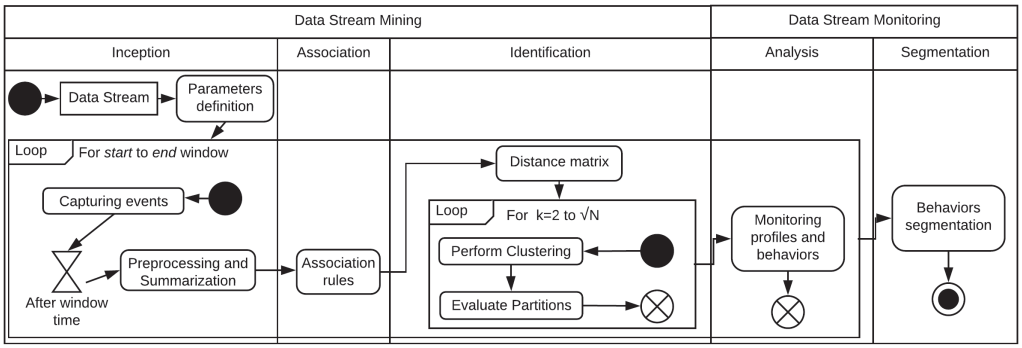

Figura 2. The Activity Diagram of the proposed framework.

3.1. Step 1: Data Stream mining

Inception phase - In the Inception phase, all events captured in the current time window are continuously preprocessed and summarised to handle with the several activities performed by each single user, as well as space and memory constraints. With a large number of app usage events, it is necessary to define a time window or a quantity of events which will be stored and preprocessed. This type of imposition is needed given the billion of events generated representing hundreds of terabytes stored in the physical memory. Time windows are the simplest way to maintain a practicable amount of data in physical memory. Such windows can help in the transition between data from the recent past and data from a distant past. Moreover, given a huge quantity of events, we require a data summarisation process to preserve the meaning of events without actually storing them [Gama 2010]. To this end, some parameters are defined in this phase: (i) the size of the time window that will cover the most recent data, (ii) a starting time, and (iii) an ending time for the data to be captured. For each window, we select the *most popular apps* and the *popular apps* based on the number of unique devices that use such apps. After, we perform the *Intuitive Partitioning* (IP) [Han et al. 2011] seeking to discretize the continuous values of the *popular apps*. The IP technique was more effective than others dividing such values into either a small or a large number of intervals. After the app usage mining, such data is summarized into a matrix composed of objects (devices) and attributes (apps) to be used as input for the Association rule algorithm in the next phase.

Association phase - We perform the Association Rules task [Tan et al. 2006] with the Apriori algorithm seeking to obtain positive correlations of app usage. We use these patterns to measure app usage similarity between the users aiming to use that in the identification of usage profiles. To the execution of this phase, it is necessary to receive as input the previously generated summary statistics as a transaction set. In addition, we need to define thresholds for support and all-confidence measures, which are used to generate itemsets, and for confidence and lift measures, which are used to generate the final rules.

Identification phase - Each itemset that composes one or more of the final rules has different users as support. For a user to be considered a support for a particular itemset, she should make use of all apps present on such itemset. In this sense, the distance between two users $u1$ and $u2$, who have their respective set of supported itemsets $i1$ and $i2$, is computed by the size of the intersection divided by the size of the union of the sample sets, as shown in Equation 1. In case of no existence of intersection between $i1$ and $i2$, the distance tends to infinity.

$$dist(u1, u2) = -log\left(\frac{|i1 \bigcap i2|}{|i1| + |i2| - |i1 \bigcap i2|}\right) \quad (1)$$

Such distance is computed for each pair of objects in a multidimensional space. Them, we obtain a matrix $y = [N \text{ x } N]$, where N is the number of observed users. In summary, the main goal of the first step is to make similar to each other all the users with the same usage patterns. Them, we carry out the Clustering task [Tan et al. 2006] aiming to identify usage profiles. The WARD algorithm receives as input a distance matrix between users based on the itemsets that generated the final rules in the previous phase. WARD produce several partitions, with a varying number of groups, and these partitions are evaluated using *Silhouette* and *Gap statistic*. We perform Data Mining and Unsupervised Machine Learning tasks aiming to extract the best patterns of the analyzed data regarding the huge amount of events produced by app usage. Thus, at the end of this step, all users have been mapped to one of the obtained clusters that are further investigated once they may change, evolve or disappear throughout time.

3.2. Step 2: Data Stream Monitoring

Analysis phase - In several real-world DS scenarios, it is necessary to monitor and investigate changes in the profiles (e.g. Concept Drift or Concept Evolution) and also of the behaviors of individuals composing these profiles. Indeed, it is necessary to distinguish or find same profiles in different time windows. It is feasible by tracking such concepts (i.e profiles) [Spiliopoulou et al. 2006, Oliveira and Gama 2010]. In this sense, we aim to perform Novelty Detection to discovery these variations that occur over the DS as well as to monitor users' behaviors given such changes.

In summary, we carry out Concept Drift and Concept Evolution approaches addressing the representation of each profile by the *enumeration* [Spiliopoulou et al. 2006] technique. Therefore, we monitor a profile by investigating the frequency distribution of its objects (users) in the next window profiles. On the other hand, it is necessary to understand the evolution of users through these profiles over time. In this sense, while each cluster label represents one profile on a single window, different labels may represent the same profile on several windows. Thus, in the next phase, we propose a new plan to analyze the changes in users' behaviors regarding the changes and evolutions of concepts found.

Segmentation phase - In this phase, we propose a sequence for the monitoring step aiming to track the changes in users' behaviors. We aim to find similar users' practices that allow segmenting such behaviors. The users are investigated to understand when and what their behavioral changes occur over the windows. In this sense, we define *life curves* seeking to represent the behavior of the users. For example, curves may show the continuity of a user in one profile, his change to other profile or the disappearance of such user. Therefore, here we aim to acquire knowledge to help us in the understanding of all users' behaviors. Moreover, we seek to segment users with same behavior and identify behaviors that could represent risk situations for mobile devices manufacturer companies.

Given a real scenario, a *life curve* is a temporal representation based on users that have the same behavior in the whole monitor process even in different profiles. To this end, some kinds of behaviors are designed (L, C, M, and O). The *L* action happens when

a user belongs to a profile c_{i+1} obtained in the window t_{i+1}, which represents the same profile c_i, to which this same user was grouped in the window t_i. The *C* action occurs when a user belongs to a profile c_{i+1}, obtained in the window t_{i+1}, which not represents the same profile c_i, to which this same user was grouped in the window t_i. The *M* action happens, when a user does not generate app usage events and such user is not grouped into the profiles obtained in the window t_{i+1}. And the *O* action occurs when a user is not supported by any itemsets obtained in the window t_{i+1}. Such behaviors are detected according to the evolution of the profiles helping us to analyze and understand how users behave throughout time.

4. Results and Discussion

We are improving our initial framework [Machado and Ruiz 2017] regarding to new experiments performed based on a real DS, as described next.

Step 1: We use a private DS provided by our sponsor, having 1,045,013,673 app events from 34,552 devices and 60,116 apps that were monitored for 140 days. A week time window (e.g. 7 days) was chosen for our experiments. Among all apps found in each week, it is possible to observe the existence of a substantial number of apps used by only a single or few devices. In this sense, we systematically explored how to define the *most used apps* for mining. To this end, we define the *most used apps* based on a minimal number of unique devices. Such apps are those used by 1% or more devices ($\overline{x}$ = 149). In addition, we define the *popular apps*, which are those used by 10% or more devices($\overline{x}$ = 33). *Popular apps* are widely used by users motivating their discretization. After the discretization process with IP approach the data are summarized into a matrix that is transformed in a transactinal set. At this step, we define thresholds for the abovementioned association measures. Such definitions are based on the percentage of users used to define the *most used apps*. In this sense, the *support* is 0.01 and the *confidence* is 0.10. In addition, *all-confidence* and *lift* measures are both computed as the mean of all values found. This way, we obtain the patterns of app usage ($\overline{x}$ = 2,000), which are used accordingly to the Equation1 to compute the distance matrix. Finally, with the combination of *WARD*, *Silhouette* and *Gap statistic* we found the usage profiles for each window ($\overline{x}$ = 6).

Step 2: With the obtained result from step one, our framework begins to monitor the profiles and their objects aiming to detect changes in the learned concepts and in the users' behaviors. In order to distinguish potential risk situation, we design all possible *life curves*, based on the *life curve* of each user. Implementations of the approaches of this step have been improved based on additional experiments, with new results expected soon. In summary, we found 706 different *life curves* performing the *enumeration* approach. In this sense, the most frequent *curve* (e.g. more users) presents only *L* actions. Such *curve* indicates that most users have the same behavior and that behavioral changes are infrequent. Finally, we are performing experiments with new parameters to the *enumeration* approach seeking to improve the results for this step and evaluate our framework.

Discussion: Since the app usage activities are typically DS events, with changes in the data distribution, such data require accurate analysis, which does not occur in batch scenarios. In this scenario is crucial to use such DS-oriented analysis and also time win-

dows, allowing the understanding of customers characteristics that may be of interest to mobile device manufacturers and other stakeholders. For example, we have been able to identify customers profiles, variations in concepts and changes in users' behaviors. Thus, such knowledge becomes relevant and may be used strategically in the decision-making process by companies seeking to understand the behavior of their users. In this sense, such knowledge may indicate investment perspectives to keep the users loyal to their brands.

Acknowledgments: We gratefully acknowledge Motorola Mobility for its support to this research.

Referências

Almana, A. M., Aksoy, M. S., and Alzahrani, R. (2014). A survey on data mining techniques in customer churn analysis for telecom industry. *IJERA*.

Backiel, A., Baesens, B., and Claeskens, G. (2016). Predicting time-to-churn of prepaid mobile telephone customers using social network analysis. *JORS*, 67(9).

Gama, J. (2010). *Knowledge discovery from data streams*. CRC Press, Boca Raton.

Hamka, F., Bouwman, H., et al. (2014). Mobile customer segmentation based on smartphone measurement. *Telematics and Informatics*, 31(2):220–227.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33:1–26.

Li, H., Lu, X., et al. (2015). Characterizing smartphone usage patterns from millions of android users. In *ACM SIGCOMM*, pages 459–472.

Machado, N. L. and Ruiz, D. D. (2017). Customer: A novel customer churn prediction method based on mobile application usage. In *IEEE IWCMC)*, pages 2146–2151.

Oliveira, M. D. and Gama, J. (2010). Mec-monitoring clusters' transitions. In *STAIRS*, pages 212–224.

Pereira, G. and Mendes-Moreira, J. (2016). Monitoring clusters in the telecom industry. In *Springer WorldCIST*, pages 631–640. Springer.

Pyo, S., Kim, E., et al. (2015). Lda-based unified topic modeling for similar tv user grouping and tv program recommendation. *IEEE CYB*, pages 1476–1490.

Rehman, A. and Raza Ali, A. (2014). Customer churn prediction, segmentation and fraud detection in telecommunication industry. *ASE BD/SI/PASSAT/BMC Conf.*

Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., and Schult, R. (2006). Monic: modeling and monitoring cluster transitions. In *ACM SigKDD*, pages 706–711.

Tan, P.-N., Steinbach, M., Kumar, V., et al. (2006). *Introduction to data mining*. Pearson Addison Wesley Boston.

Wagner, D. T., Rice, A., and Beresford, A. R. (2013). Device analyzer: Understanding smartphone usage. In *Springer MobiQuitous*, pages 195–208.

Xu, Q., Erman, J., et al. (2011). Identifying diverse usage behaviors of smartphone apps. In *ACM SIGCOMM*, pages 329–344.