Mining RADIUS Data: How to detect movement patterns?

Maribel Yasmina Santos
Algoritmi Research Centre, University of Minho
Campus de Azurém
Guimarães, Portugal
maribel@dsi.uminho.pt

Abstract

This paper presents some of the challenges that arise in the analysis of data associated to RADIUS logs, looking at movement patterns that can be identified using data mining algorithms. When talking about movement, space is inherent to the places visited by groups of individuals. However, no absolute reference to a geographic position exists in these data. The position of the Wi-Fi nodes gives some semantic about the places where people are, but does not allow the verification if two nodes are, for example, near each other. This paper describes the work developed so far in the process of analysis of such complex data set. The steps of the knowledge discovery in databases process were applied in order to verify if traditional data mining algorithms can be used in the analysis of RADIUS logs to identify places that are visited in the same sequence by several users. The obtained results show that the association rules technique can be used to identify places that are visited in a similar way by groups of individuals.
Keywords: data mining, radius data, association rules, movement patterns.

1 Introduction

Data Mining has been extensively applied to several types of data with the aim of identifying patterns or trends in data [1]. Independently of the data type to be analysed, several challenges are usually associated to the analysis of complex data sets. This is the case of the data obtained through the RADIUS (Remote Authentication Dial In User Service) networking protocol.

As this protocol offers a wide support, it is often used by ISP (Internet Service Providers) to manage access to the Internet or internal networks, wireless networks, and integrated e-mail services. It is commonly used to facilitate roaming between ISP [2]. In the particular case of this work, data of the Eduroam network are used. This network allows the authentication of users by their home institution when visiting a collaborating institution. Collaborating institutions provide their own credentials to their own users.

For the Eduroam network, this work looks at RADIUS data collected at the "Campus de Azurém" of the University of Minho in Portugal during ten days. In ten days of data, more than 684,000 records are available. In one day almost 50,000 logs are recorded.

When a user (device) connects to the network, the RADIUS service logs the identification of the device, the identification of the Access Point (AP), and a timestamp. With these data it is possible to know who the user is, where the connection was established and when the user connected or disconnected from the network.

Looking at these RADIUS data, this work investigates if traditional data mining techniques can be used to identify movement patterns, which means that we look for sequences of places that are visited in a similar way by the users.

The work here described is undertaken considering the steps of the knowledge discovery in databases process [1], which includes the selection, treatment, pre-processing and mining of the data, and the interpretation of results. Following these steps, several challenges have to be handled:

- The huge amount of data that is collected every day by each collaborating institution in the Eduroam network;
- The considerable large number of Access Points (APs) available in each Campus' facilities;
- The lack of information about the geographic position of the WI-Fi APs, representing the lack of a geographic reference for the presence of people in space.

Considering the identification of movement patterns, as sequences of places that are visited in a similar way by different users, the association rules technique was selected as the data mining algorithm to be applied at the data mining step of the knowledge discovery process. Association allows the identification of rules that relate items (places) that appear together in an event (connections of the users to the Eduroam network in a time interval).

The results obtained so far allowed the identification of several rules expressing the sequence of places visited by the Eduroam users.

After an introduction to the content of this paper and the motivation for undertaking this work, this paper proceeds with a brief overview on related work (Section 2); the description of the available data for analysis (Section 3); the obtained results using data mining to identify movement patterns (Section 4); and, a brief conclusion and guidelines for future work (Section 5).

2 Related work

The analysis of mobility data associated with Wi-Fi networks has been addressed by several authors. Yoon et al. [3] presented a trace-driven framework capable of building

mobility models for mobile systems simulation studies. The proposed framework combines wireless traces, association data between WiFi users and APs, with a map of the space over which the traces were collected. A probabilistic mobility model that produces user movement patterns that are representative of real movement was generated. For this, several heuristics allowed the inference of the paths users take between APs.

Bhattacharjee et al. [4] focused their work on obtaining mobility models for the proper evaluation of ad hoc protocol performance. They developed a mobility model by observing the actual movement patterns of people on-campus and then post processing the data to verify the spatio-temporal distributions.

In [5], Kim and Kotz presented a methodology for extracting mobility information from wireless network traces, and for classifying mobile users and APs. The authors used the Fourier transform to convert time-dependent location information to the frequency domain, then selected the two strongest periods and used them as parameters to a classification system that uses the Bayesian theory. To classify mobile users, the authors computed the maximum distance between any two APs visited by a user during a fixed time period and observed how this value changes or repeats over time. The authors found that user mobility had a strong period of one day, but there was also a large group of users that had either a much smaller or much bigger period.

The analysis of mobility data with the aim of identification of movement patterns has been applied in several contexts and with diverse purposes. Previous works call our attention to several applications. Independently of the methodology or adopted analysis techniques, movement patterns were identified.

In the work presented in this paper no new data analysis technique is proposed. The aim is to verify how traditional data mining algorithms can be used to extract movement patterns from RADIUS data. For this, the traditional steps of the knowledge discovery in databases process are followed. In the data mining step, association rules were chosen as the data analysis technique.

Knowledge Discovery in Databases (KDD) is a complex process concerning the discovery of relationships and other descriptions from data [6]. The steps of this process include Data Selection, Data Treatment, Data Pre-processing, Data Mining and Interpretation of Results.

For Data Mining, association rules are models that examine the extent to which values of one field depend on, or are predicted by, values of another field. Association identifies rules about items that appear together in an event [7].

3 RADIUS Data

The first step of the KDD process is associated with data selection. This step allows the selection of relevant data needed for the execution of a defined data mining task. In the scope of this work, the data mining task derives from the objectives of the SUM (Sensing and Understanding human Motion dynamics) project, in which it is expected to find if state-of-the-art machine learning and pattern recognition techniques can be used to uncover mobility patterns in the eduroam-collected data.

A RADIUS server collects a wide range of attributes. In this work, a small subset was selected. It is expected that these attributes are suitable for the task of movement analysis as they include the location (AP) where users connected to or disconnected from the Eduroam network.

The selected attributes and their meaning are listed in Table 1. Beside the location (where), they allow the analysis of when an event took place and who is associated with it.

Table 1: Subset of data selected for analysis

Attribute	Description
Timestamp	Time stamp of the register
Acct_Status_Type	Type of event
Acct_Session_Time	Time of session (in Seconds)
Called_Station_Id	Identification of the connecting AP
Calling_Station_Id	Identification of the connecting device (user)
WISPr_Location_Name	Descriptive location of the AP

If we look at the Timestamp attribute, in a day of access to the network, several different values are available as the information includes the hour, minute and second of the access. For the analysis of the access time, temporal classes need to be created and the study of temporal series also needs to be considered. As this paper addresses in a first place movement, time will be considered in future works.

For the Acct_Status_Type attribute, three values are available: Start (if the user is starting a session), Alive (if the user is maintaining a session) and Stop (if the user is ending a session). This information is relevant as it indicates for how long a device (person) remained connected to one AP. Since the coverage area of each AP is small, we can assume that a person remained within that area between the time instants going from a Start to a Stop record.

The Acct_Session_Time registers the duration (in seconds) of the session. This attribute could be used to group users attending to their behavior in terms of duration of the sessions. For example, what is the expected time of connection at a given location?

Called_Station_Id identifies the different APs that can be reached by the mobile users. In the context of the work described in this paper, and considering the University Campus where the data were collected, 143 different called stations are available.

Calling_Station_Id identifies the different devices that can be used to access the several APs. They correspond to the MAC addresses of the devices. In a day of data collection, almost 1,000 different devices were identified. In ten days, we have 2,629 different devices.

The WISPr_Location_Name represents a textual location of the place where the AP is located.

In the data treatment step of the KDD process, the cleaning of the selected data, which allows for the treatment of corrupted data and missing data fields, is undertaken. In the context of this work, data are collected in an automatic way decreasing errors and missing data fields. At this stage, no cleaning was needed.

In the data pre-processing step, it is possible the reduction of the sample destined for analysis. Two tasks can be carried

out here: i) the reduction of the number of rows or, ii) the reduction of the number of columns. In the reduction of the number of rows, data can be generalized according to the defined hierarchies or attributes with continuous values can be transformed into discreet values according to the defined classes.

In this work, hierarchies were used for the WISPr_Location_Name attribute as 123 different descriptions are available. For the analysis of movement patterns [8], to know what the movement among locations (and not among APs) is, this large amount of descriptive locations does not facilitate the identification and comprehension of movement patterns. Also, it allows the identification of movement patterns that can be associated to the "ping-pong" phenomenon. This happens when a user is within the coverage of two or more APs. For example, the log data may show that user A was associated with AP_1 at time t_1 and with AP_2, located a short distance away, after 1 second. Two or three seconds after, the log registers that the user is associated again with AP_1. In many cases this happens because the user was somewhere within the range of both APs. When this happens, the log may have many entries that do not correspond to a movement of the user or to a significant movement of the user [5].

To pre-process the WISPr_Location_Name attribute and overcome the previous problems, a hierarchical aggregation of locations was carried out. APs belonging to the same department, cantina, bar, among others, and having different descriptions were transformed in order to address the same location. Some examples of this transformation can be found in Table 2. Through this process, the 123 original descriptions were reduced to 37 descriptions.

Table 2: Transformation on the places' descriptions

Original description[1]	Assigned description
UMINHO-bar-0a	UMINHO-bar
UMINHO-bar-0b	
UMINHO-cantina-1a	UMINHO-cantina
UMINHO-cantina-0a	
UMINHO-cantina-1b	
UMINHO-cantina-2a	
UMINHO-dsi-1a	UMINHO-dsi
UMINHO-dsi-1c	
UMINHO-dsi-1b	
UMINHO-dsi-1d	

The original data were also pre-processed in order to create an intermediary table (DifferentLocals10Days) that has, for each user, the descriptions of the places that were visited in a specific day or period of time. We consider 5 hours as the maximum absent period in the network. After 5 hours we start a new sequence of visited places. For the 2,629 different devices, we obtained 11,478 movement traces indicating that a user moved from LocalA to LocalB, from LocalB to LocalC, and so on (Table 3).

Table 3: Users and visited places

User[2]	Visited places
Calling_Station_ID= 001xx01x1111	UMINHO-dec ↓ UMINHO-dem
Calling_Station_ID= 001yyyy22y33	UMINHO-sdum ↓ UMINHO-blocoa ↓ UMINHO-blocob
Calling_Station_ID= 000zz100z200	UMINHO-quimica ↓ UMINHO-dps ↓ UMINHO-mct ↓ UMINHO-dsi

The DifferentLocals10Days table has sequences of places visited by the users. After the aggregation of the WISPr_Location_Name attribute, each sequence was analysed in order to remove repeated places originated, for example, by the "ping-pong" phenomenon. If a user is being intermittently associated to two APs of the same location, the original sequence is transformed into a new one without duplicated locations.

Consider for example the original sequence:

UMINHO-blocob, UMINHO-blocob, UMINHO-blocoa, UMINHO-cantina

This sequence is transformed to:

UMINHO-blocob, UMINHO-blocoa, UMINHO-cantina

stating that the user started the connection in building B (UMINHO-blocob), then proceeded to building A (UMINHO-blocoa) and then ended at the cantina (UMINHO-cantina).

Next section proceeds with the presentation of the results obtained in the data mining step and, also, with the interpretation of results.

4 Results

The data mining step looks for association rules [9] that relate the presence of users in the same places with a certain probability.

Using the DifferentLocals10Days table as input for the learning process and an association rule algorithm made available by the SQL Server Business Intelligence Studio 2008©, we identified several association rules that exemplify the places that are visited by the users. Table 3 shows some of these rules.

[1] For confidentially and privacy reasons, the original descriptions were modified for presentation in this paper.

[2] For privacy reasons, all the presented MAC addresses are fictitious.

Table 3: Some of the obtained Association Rules

Probability	Rule
0.651	UMINHO-piep → UMINHO-reaz
0.579	UMINHO-dsi, UMINHO-ee → UMINHO-dps
0.579	UMINHO-blocob, UMINHO-sdum → UMINHO-bar
0.560	UMINHO-fisica → UMINHO-mct
0.522	UMINHO-dep → UMINHO-dps
0.500	UMINHO-blocob, UMINHO-bar → UMINHO-sdum

A small set (15) of rules was obtained and all presenting a small confidence, as it is possible to see in Table 3. The obtained rules, although the low probability, make sense as show how a user behave inside a building. For example, the rule:

UMINHO-blocob, UMINHO-sdum → UMINHO-bar

- Represents a typical behaviour in students that start a day having classes in building B (UMINHO-blocob), then go to the library to study or pick up some book (UMINHO-sdum) and finally go to the bar to eat something (UMINHO-bar).

Transforming the obtained rule to the format IF ... THEN, we have that:

IF UMINHO-blocob AND UMINHO-sdum THEN UMINHO-bar

An interesting point is that some variants of the rules are also detected by the algorithm, when the sequence of the movement suffers changes. This means that users can start the movement by visiting building B, then go to the bar, and only afterwards go to the library. This variant of the former rule, also present in Table 3, is represented by:

UMINHO-blocob, UMINHO-bar → UMINHO-sdum
or
IF UMINHO-blocob AND UMINHO-bar THEN UMINHO-sdum

5 Conclusion and Future Work

This paper presented the analysis of RADIUS data in order to identify movement patterns that although having an explicit reference to a location, cannot be analysed by spatial data mining algorithms as the spatial reference is qualitative.

The use of an association rules algorithm allowed the identification of several rules expressing the sequence of places visited by the Eduroam users. Although the identified rules show some interesting patterns, as we can recognise in them usual users' movements, weak correlations were obtained.

This low probability of the rules might be explained by the sample of data used for analysis. Ten days of data seem to be not enough for establishing a correlation among the places that are usually visited by the users.

In order to improve the results, future work includes the analysis of more data, with a bigger sample of a month, and the analysis of different data sets, for different months, to verify if the obtained movement patterns present a temporal component. In this last case, time sequences would be obtained.

As other data mining techniques need to be explored, future work also includes the use of decision trees to predict the time interval during which a user is connected to the network and how this affects the movement patterns. Additionally, the use of clustering is foreseen to group users with similar behaviour.

Acknowledgements

This project was supported by the Portuguese Science and Technology Foundation (FCT - *Fundação para a Ciência e Tecnologia*), under project SUM (Sensing and Understanding human Motion dynamics, PTDC/EIA-EIA/113933/2009). This work was partly funded by FEDER funds through the Operational Competitiveness Program (COMPETE) and by FCT with the project: FCOMP-01-0124-FEDER-022674.
I wish to thank Adriano Moreira for his comments and suggestions.

References

[1] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, 2001.

[2] J. Hassell. RADIUS: securing public access to private resources. O'Reilly Media, 2003.

[3] J. Yoon, B. D. Noble, M. Liu and M. Kim. Building Realistic Mobility Models from Coarse-Grained Traces, in Proceedings of ACM MOBISYS'2006. 2006. Sweden.

[4] D. Bhattacharjee, A. Rao, C. Shah, M. Shah and A. Helmy. Empirical modeling of campus-wide pedestrian mobility observations on the USC campus, in Proceedings of the 60th IEEE Vehicular Technology Conference. 2004.

[5] M. Kim and D. Kotz. Classifying the Mobility of Users and the Popularity of Access Points, in Proceedings of the International Workshop on Location- and Context-Awareness (LoCA). 2005. Germany: Springer-Verlag.

[6] U. Fayyad and R. Uthurusamy. Data Mining and Knowledge Discovery in Databases. *Communications of the ACM*, 39(11): 24-26, 1996.

[7] R. Agrawal, T. Imielinski and A. Swami. Mining Association Rules between Sets of Items in Large Databases, in Proceedings of the 1993 ACM SIGMOD Conference on Management of Data. Washington DC. 1993.

[8] N. Andrienko, G. Andrienko, N. Pelekis and S. Spaccapietra. Basic Concepts on Movement Data, in *Mobility. Data Mining and Privacy*, F. Giannotti and D. Pedreschi, Editors. 2008, Springer-Verlag. p. 15-38.

[9] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules, in Proceedings of the 20th VLDB Conference, Santiago, Chile, 1994.