On Mining Mobile Users by Monitoring Logs

Dmitry Namiot
Lomonosov Moscow State University
119991, GSP-1, 1-52, Leninskiye Gory
Moscow, Russia
+7-495-9392359
dnamiot@gmail.com

ABSTRACT
This paper considers a new model of data analysis for monitoring of mobile devices. Passive monitoring of mobile devices is based on ideas of network proximity and uses network protocol analysis for Wi-Fi and Bluetooth to gather presence information on mobile visitors. This is a direct analogue for web log and web site usage data, but we can deal with real visitors (with their mobile devices), rather than with abstract requests for web pages. In this paper, we propose a new model for processing of these data, which can detect some form of relationships between mobile users.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]

General Terms
Algorithms, Experimentation.

Keywords
Mobile monitoring, Wi-Fi, clustering, data mining.

1. INTRODUCTION
Monitoring the presence of mobile users (subscribers) is one of the most interesting and useful sources of information in smart cities data processing [1]. Monitoring of mobile users (in fact, it is monitoring for mobile devices) supplies data to evaluate the mobility of residents, planning transport routes, etc. [2]. On the lower level, we can talk, for example, about retail applications where analysis of the presence of mobile subscribers can be used to improve service, evaluation of marketing campaigns, planning, etc. [3]. At this moment, we can list well known and commonly used methods for determining the location of mobile devices based on the location of Wi-Fi access points [4]. Mobile operating systems (mobile applications) can use the information about the objects of the network infrastructure for verifying (or even determine) the true state of the subscriber. By analyzing the signal strength and visibility of access points we can build various metrics about the location of mobile devices (mobile users) [5, 6].

Passive Wi-Fi monitoring is one of the commonly approaches [7]. It lets anonymously collect data about mobile users (mobile devices) in proximity of some metering device [8].

ECIR'14 Information Access in Smart Cities Workshop (i-ASC 2014).
April 13, 2014, Amsterdam, the Netherlands.

This paper presents a new model for processing data collected during the passive monitoring for Wi-Fi (Bluetooth) devices. The rest of the paper is organized as follows. In Section 2 we describe the mobile monitoring and collected datasets. In Section 3 we describe exiting approaches for data processing as well as our data mining approach.

2. MOBILE MONITORING
Collecting traces of Wi-Fi beacons is the well-know approach for getting the locations of mobile devices. Beacon frames are used to announce the presence of a Wi-Fi network. As a result, an 802.11 client receives the beacons sent from all nearby access points. The client receives beacons even when it is not connected to any network. In fact, even when a client is connected to some particular Access Point (AP), it periodically scans Wi-Fi channels to receive beacons from other nearby APs [9]. It lets clients keep track of networks in its vicinity. But at the same time a Wi-Fi client periodically broadcasts an 802.11 probe request frame. The client expects to get back an appropriate probe request response from Wi-Fi access point. As per Wi-Fi spec, a station (client) sends a probe request frame when it needs to obtain the information from another station [10]. Figure 1 illustrates data flow for Probe Requests.

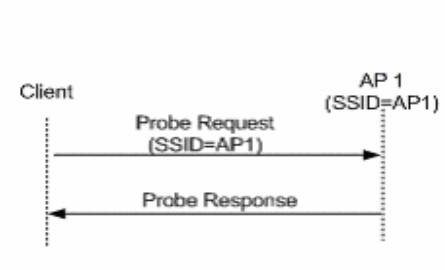

Figure 1. Wi-Fi Probe request/response

Technically, probe request frame contains the following information:

- source address (MAC-address)
- SSID
- supported rates
- additional request information
- extended support rates
- vendor specific information

Our metering device (it could be a Wi-Fi router, for example) can analyze received probe requests. Obviously, any new request (any new MAC-address) corresponds to a new wireless customer

nearby. Note, that Bluetooth devices could be monitored by the same principles.

Wi-Fi based device detection uses only a part of the above mentioned probe request. It is a device-unique address (MAC address). This unique information lets us re-identify the devices (mobile phones) across our monitors. The sequence of sequential requests (records) with the same MAC-address forms a session (similar to HTTP session in web applications).

Technically, data collected during passive Wi-Fi monitoring is similar to data collected in web statistics. Web statistics (web logs mining) are based on the standard format (formats) for log-files. The common standard is provided by W3C [11]. An extended log file contains a sequence of lines containing ASCII characters terminated by either the sequence LF or CRLF. Each line there corresponds to one request. Each line may contain either a directive or an entry [12]. The typical records contain the following fields: host address (IP address), user name, date, time, time zone information, URI (request), HTTP protocol version, status code, size of response in bytes.

For mobile monitoring, we can use the following fields: MAC-address for the device (hash-code for the privacy replacement), date, time, time zone info, signal strength (RSSI), name of the access point. The key missed point is the request (URI). It is obvious, that there are simply no requests for presence records. It is a key point, because many of exiting processing models can use URI data (e.g., for clustering).

Also, we should note, that passive Wi-Fi detection is not 100% reliable. Mobile phones (mobile OS, actually) can actually transmit probe requests at their discretion. Our own experiments with commercially available Wi-Fi probe scanners confirm data from [13]. The monitor detects in average about 70% of passing smartphones.

3. DATA PROCESSING

Technically, most of the monitoring systems for mobile devices treat collected data as some form of web log and provide appropriate statistics. The typical explanation of the existing systems is something like “Google Analytics for the real world” [14]. The typical analytical issue contains the number of visitors during the period, their timing, the number of unique visitors, the estimate for the number of regular visitors, etc. Figure 2 demonstrates a distribution of visits by type of mobile devices.

Extracting information from a Web log is fairly well-known research topic [14, 15], and consequently, the different software products. Usually, the study (analysis) can be classified into the following categories: content analysis, analysis of the structure and usage analysis. Analysis of the usage, in turn, may include personalization system, recommendations for modification sites, and business intelligence.

From the analysis of different patterns allocated when analyzing Web logs, we have identified one direction, which is almost not covered in this context. It is the mining of user groups. Actually, it is explainable for web statistics. Stable group of users for web access is several visitors browsed the site in parallel (approximately parallel) mode. This may make sense if we are talking, for example, about search bots. Yes, they can demonstrate sometimes correlation in time visit. For routine visits such grouping is rather artificial. Vice versa, for mobile monitoring, where each hit (each record in the log file) is some real visit, time based grouping makes sense.

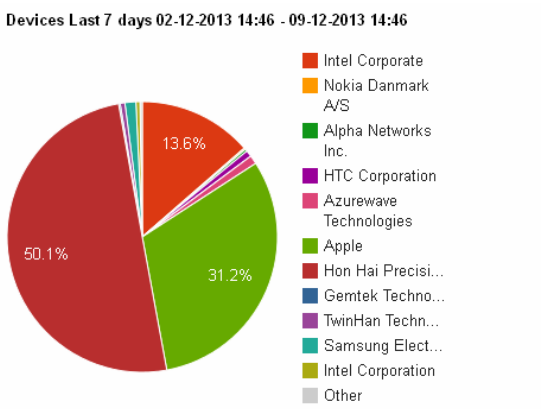

Figure 2. Mobile devices for visitors

There are some papers describing grouping for moving objects (for trajectories) [16, 17] Yes, it could be reproduced for proximity data too. For example, our paper [18] describes relationships mining for proximity data and models like Spotex [19]. But for such kind of tasks we need several metering devices. In this paper, we deal with the classical schema – one metering device and one log file.

For the typical web statistics, frequent visitors, for example, are IP addresses recorded (logged) every day for 7 (week) or 30 days (month). We want to extend this pattern to groups. Let us see a practical example. There is some group of friends, which occurs within a certain time in a cafe (co-working space, etc.). Not all of the members are present at each meeting, not all of them, as usually, arrive simultaneously (Figure 3). Can we discover such a group (groups) by proximity log?

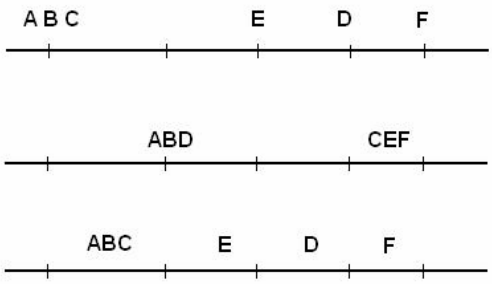

Figure 3. Visitors (A B C D E F) for 3 days.

One of the possible approaches for time-based analysis and events clustering is presented in [20]. It is based on the temporal similarity matrix. For two events i and j with timestamps t_i and t_j similarity for time interval K is:

$$S_k(i,j) = \exp\left(-\frac{|t_i - t_j|}{K}\right)$$

Authors present a method that first calculates the temporal similarity between all pairs of events (originally – photographs).

The calculated values are stored in a chronologically ordered matrix. And cluster boundaries are determined by calculating novelty scores for each set of similarity matrices. The authors assume that the events (in the original paper – photos) at cluster boundaries (in the original paper – event boundaries) separate two adjacent groups of events with high intra-class temporal similarity and low inter-class similarity.

In our research, we've followed to another approach. As it is mentioned in [21], time based clustering could be different from the traditional K-means clustering [22]. K-means clustering might find cluster centers with an idea to minimize some cost function. A traditional clustering algorithm (K-means might) find clusters and cluster–centers for the given K. As a basic point it uses the fact the cost function would change if those cluster–centers are moved. Cost function is the distance of data points to cluster centers. For our time stamped events we are not concerned with finding cluster centers at all. Really, the exact value for time any group is collected should be irrelevant. Our algorithm should only assign collected points to clusters and as long as the segmentation remains the same. We need segmentation or splitting of the time sequence. The key question is how to split our events in time, so that intra-cluster variance is reduced.

And our idea of mining groups is based on two sequential steps:

- find clusters for the each day
- detect the sequences of clusters across all days with some minimum set of common members

It is illustrated on Figure 4.

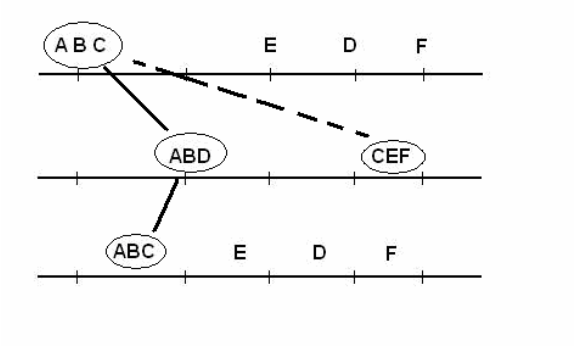

Figure 4. Clusters and groups

For getting clusters we followed to the algorithm originally developed for clustering photos [21]. It offers dynamic clusters with automatically detected boundaries (splitting points). It is based on two assumptions.

- Increased (long) time interval without registrations usually marks the end of cluster (all members of the group are in already). "Long" is defined as being either large relative to the extent of the cluster currently being examined (collected), or large relative to the average inter-group interval.
- Changed frequency of registrations corresponds to the start of a new group

As soon as clusters are detected, we can examine them for common elements (MAC-addresses), presented in the majority of per-day clusters. It means that group detection is always associated with predefined percentage of visits. E.g. visitors participated at least in 75% of meetings.

For this examination let us present each group as a string, where each element corresponds to the unique MAC-address (see Figure 4). Now we need to find a common subsequence of strings (groups) across all days (see solid line on Figure 4).

String C is a common subsequence of strings A and B if C is a subsequence of A and also a subsequence of B. String C is a longest common subsequence of string A and B if C is a common subsequence of A and B of maximal length. It means that there is no common subsequence of A and B that has greater length [23]. The typical algorithm for finding the longest common subsequence could be obtained from papers [23, 24].

The proposed system has been implemented in connection with Wi-Fi scanner from Libelium. During the testing stage we've successfully identified 8 groups from 11 (café in office building).

4. CONCLUSION

In this paper, we propose a new model for analyzing web-logs collected by the mobile phones monitoring systems. From the analysis of different patterns of web logs mining, we have identified one direction, which is almost not covered in this connection. It is the mining of user groups. In our paper, we propose two step algorithm for grouping mobile visitors. It could be used in Smart City projects as well as in retail information systems.

5. REFERENCES

[1] Murty, R., Gosain, A., Tierney, M., Brody, A., Fahad, A., Bers, J., & Welsh, M. (2008, May). CitySense: A vision for an urban-scale wireless networking testbed. In Proceedings of the 2008 IEEE International Conference on Technologies for Homeland Security, Waltham, MA.

[2] Cornelius, C., Kapadia, A., Kotz, D., Peebles, D., Shin, M., & Triandopoulos, N. (2008, June). Anonysense: privacy-aware people-centric sensing. In Proceedings of the 6th international conference on Mobile systems, applications, and services (pp. 211-224). ACM.

[3] Ryder, J., Longstaff, B., Reddy, S., & Estrin, D. (2009, August). Ambulation: A tool for monitoring mobility patterns over time using mobile phones. In Computational Science and Engineering, 2009. CSE'09. International Conference on (Vol. 4, pp. 927-931). IEEE.

[4] Namiot, D., and Sneps-Sneppe, M. (2012, April). Proximity as a service. In Future Internet Communications (BCFIC), 2012 2nd Baltic Congress on (pp. 199-205). IEEE. DOI: 10.1109/BCFIC.2012.6217947.

[5] Lassabe, F., Canalda, P., Chatonnay, P., & Spies, F. (2009). Indoor Wi-Fi positioning: techniques and systems. Annals of telecommunications-Annales des télécommunications, 64(9-10), 651-664.

[6] Zàruba, G. V., Huber, M., Kamangar, F. A., & Chlamtac, I. (2007). Indoor location tracking using RSSI readings from a single Wi-Fi access point. Wireless networks, 13(2), 221-235.

[7] Labiod, H., Afifi, H., & De Santis, C. (Eds.). (2007). Wi-Fi, Bluetooth, Zigbee and WiMAX. Springer.

[8] Namiot D. and Sneps-Sneppe M. Geofence and Network Proximity. In Internet of Things, Smart Spaces, and Next Generation Networking, Lecture Notes in Computer Science. Volume 8121, 2013, pp. 117-127, DOI: 10.1007/978-3-642-40316-3_11.

[9] Dmitry Namiot and Manfred Sneps-Sneppe. "Local messages for smartphones". Future Internet Communications (CFIC), 2013 Conference on (pp. 1-6). IEEE. DOI: 10.1109/CFIC.2013.6566322.

[10] M.Gast 802.11 Wireless Networks: The Definitive Guide O'Reilly Media, Inc., 2005, 654 p.

[11] Jansen, Bernard J., Amanda Spink, and Isak Taksai. Handbook of research on web log analysis. London: Information Science Reference, 2009.

[12] W3C log: http://www.w3.org/TR/WD-logfile.html Retrieved: Jan, 2014

[13] A. Musa and J.Eriksson, "Tracking Unmodified Smartphones Using Wi-Fi Monitors", SenSys'12, November 6–9, 2012, Toronto.

[14] Yang, Q., Zhang, H. H., & Li, T. (2001, August). Mining web logs for prediction models in WWW caching and prefetching. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 473-478). ACM.

[15] Grace, L. K., Maheswari, V., & Nagamalai, D. (2011). Analysis of web logs and web user in web mining. arXiv preprint arXiv:1101.5668.

[16] M. Andersson, J. Gudmundsson, P. Laube, and T. Wolle. Reporting leaders and followers among trajectories of moving point objects. GeoInformatica, 12(4):497–528, 2008.

[17] Li, Z., Ding, B., Wu, F., Lei, T. K. H., Kays, R., & Crofoot, M. C. (2013). Attraction and Avoidance Detection from Movements. Proceedings of the VLDB Endowment, 7(3).

[18] Namiot, D. (2013). Mining Relationships in Proximity Movements. Applied Mathematical Sciences, 7(144), 7173-7177.

[19] Namiot, D. (2012, September). Context-Aware Browsing--A Practical Approach. In Next Generation Mobile Applications, Services and Technologies (NGMAST), 2012 6th International Conference on (pp. 18-23). IEEE. DOI: 10.1109/NGMAST.2012.13

[20] Cooper, M., Foote, J., & Girgensohn, A. (2003, September). Automatically organizing digital photographs using time and content. In Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on (Vol. 3, pp. III-749). IEEE.

[21] Gargi, U. (2003). Consumer media capture: Time-based analysis and event clustering. HP-Labs Tech Report.

[22] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) 28.1 (1979): 100-108.

[23] Hirschberg, Daniel S. "Algorithms for the longest common subsequence problem." Journal of the ACM (JACM) 24.4 (1977): 664-675.

[24] Hunt, J. W., & Szymanski, T. G. (1977). A fast algorithm for computing longest common subsequences. Communications of the ACM, 20(5), 350-353.