Reality Mining: Data Mining or something else?

PETR BERKA
University of Finance and Administration, Prague, Czech Republic

Abstract: Reality mining is defined as "collection and analysis of machine-sensed environmental data pertaining to human social behavior, with the goal of identifying predictable patterns of behavior". As such it can be understood as a particular application area of data mining. Or is there more behind this new notion? And will reality mining become a new subfield of data mining in a same way as text mining or web mining? The paper looks on similarities and differences between reality mining and "classical" data mining and reviews some applications presented by their authors as reality mining.

Keywords: reality mining, data streams, sensor data, ubiquitous knowledge discovery.

1 Introduction

Knowledge discovery in databases (KDD) or data mining (DM) is aimed at acquiring implicit knowledge from data and using it to build classification, prediction, description, etc. models for decision support. As more data is gathered, with the amount of data doubling every three years, data mining becomes an increasingly important tool to transform this data into knowledge.

In 1990th data mining was used to analyze tabelar data, where each row corresponds to a single object (market basket, patient, bank client, etc.) and each column corresponds to an attribute (categorical or numerical) describing a property of the objects. Such data can be found in many many application areas: banking and finance, medicine, manufacturing, sociology, etc. Nevertheless, this type of data is not the only one to be analysed. Nowdays we can see data mining applied to time series, graphs, dat streams, texts, images. And data mining on particular data types became its own name. So text mining became the name for data mining on unstructured textual documents and web mining became the name for data mining on data gathered from web. These subfields of data mining are already well established and accepted both by the data mining/scientific comunity and users of the data mining results.

A decade ago, a new term "reality mining" occurs in the data mining community. How is this new term related to "classical" data mining? And what are characteristical features of reality mining apllications and projects? The paper looks on similarities and differences between reality mining and "classical" data mining and reviews some applications presented by their authors as reality mining.

2 Reality Mining vs. Data Mining

The term "reality minig" was coined by Nathan Eagle and Alex Pentland from Media Laboratory, Massachusetts Institute of Technology (MIT) about 10 years ago. According to them, reality mining is the collection and analysis of machine-sensed environmental data

pertaining to human social behavior, with the goal of identifying predictable patterns of behavior.

Reality mining studies human interactions based on the usage of wireless devices such as mobile phones and GPS systems providing a more accurate picture of what people do, where they go, and with whom they communicate with rather than from more subjective sources such as a person's own account. As such it can be understood as a particular application area of data mining, where the analysed data are (usually) collected from wireless devices (like mobile phones or GPS receivers).

From the data mining point-of-view, reality mining deals with the most challenging data mining problems as defined in [12]. In particular, it tackles the issues of "scaling up for high dimensional data/high speed streams", "mining sequence data and time series data", and "data mining in a network setting". Le's review some of these issuse in this section.

2.1 Sensor Networks

Sensor networks consist of distributed autonomous devices that cooperatively monitor an environment. Each node in a sensor network is able to sense, process and act. Sensors are equipped with capacities to store information in memory, process information and communicate with neighbors. They have strong constraints on resources such as energy, memory, computational speed and bandwidth.

Typical applications of sensor networks include monitoring, tracking, and controlling. Some of the specific applications are habitat monitoring, object tracking, nuclear reactor controlling, fire detection, traffic monitoring etc. In a typical application, a wireless sensor network (WSN) is scattered in a region where it is meant to collect data through its sensor nodes. Sensor nodes can be imagined as small computers, equipped with basic capacities in terms of their interfaces and components. Sensors act in dynamic environments, under adversarial conditions.

With the evolution of sensing hardware, sensor networks have attracted the attention of data stream researchers as a platform for many significant applications.

2.2 Mining Data Streams

Many sources produce data continuously. Examples include sensor networks, wireless networks, radio frequency identification (RFID), customer click streams, telephone records, multimedia data, scientific data, sets of retail chain transactions etc. These sources are called data streams. A data stream is an ordered sequence of instances that can be read only once or a small number of times using limited computing and storage capabilities. These sources of data are characterized by being open-ended, flowing at high-speed, and generated by non stationary distributions in dynamic environments.

Data streams are everywhere these days. Examples of applications include network monitoring, user modeling in web applications, sensor networks in electrical networks, telecommunications data management, surveillance systems, prediction in stock markets, monitoring radio frequency identification, smart houses and assistive living etc.

Extracting knowledge from multiple, possibly distributed data streams, is one of the most significant challenges that we face today. In these applications it is not feasible to load the arriving data into a traditional data base management system. Also the use of conventional data mining algorithms for analyzing static data sets that work based on many assumptions

such as stationary distribution, ability to store all the data in memory at the same time, and centralized collection of all the data for subsequent analysis are questionable in this new domain.

2.3 Ubiquitous Knowledge Discovery

The concept of reality mining is also closely related to the concept of ubiquitous knowledge discovery. Ubiquitous Knowledge Discovery can be defined as "Knowledge discovery process in mobile, distributed, dynamic environments, in presence of massive amounts of data". Knowledge discovery in ubiquitous environments is an emerging area of research at the intersection of the two major challenges of highly distributed and mobile systems and advanced knowledge discovery systems. Research areas as defined in the EU funded project KDUbiq (2005-2008 FP6 FET IST) are [3]:

- data mining in mobile systems, wireless communication networks, calm technologies,
- distributed architectures: distributed data mining, grid, P2P, autonomic computing,
- agents,
- learning components: statistical learning (incl. online learning), evolutionary computing,
- anytime algorithms data types: spatio-temporal, stream, multimedia,
- security and privacy: privacy preserving data mining, intrusion detection,
- HCI and cognitive modelling: user interfaces of ubiquitous discovery systems.

This multi-disciplinary approach constitutes a paradigm shift for the field of knowledge discovery since the idea of a standalone (desktop or workstation) analysis tool is abandoned in favour of process integrated, distributed and autonomous analysis systems. In a fully ubiquitous setting, the learning typically takes place inside the small devices. Instead of off-line learning in a batch setting, sequential learning, anytime learning, real-time learning, online learning eytc. under real-time constraints from ubiquitous and distributed data is needed. Instead of learning from stationary distributions, concept drift is the rule rather than the exception. Instead of large stand-alone workstations, learning takes place inside small, unreliable devices.

3 Example Reality Mining Projects

Let's have a closer look on some applications presented by their authors as reality mining.

3.1 Complex social systems

A pioneering project in the area of reality mining was realized by Eagle and Pentland in 2004 [2]. They collected data from 100 mobile phones of students and researchers at MIT over the course of 9 months. This makes about 45 000 hours of communication logs, location and proximity data. Seventy-five users were either students or faculty in the MIT Media Laboratory, while the remaining twenty-five were incoming students at the MIT Sloan business school adjacent to the Laboratory. Of the 75 users at the Lab, 20 were incoming master's students and 5 were incoming MIT freshman. The collected information includes call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (such as charging and idle). Capturing mobile phone usage patterns of 100 people for an extended period of time can provide insight into both the users and the ease of use of the device itself. By continually logging and time-stamping information about a user's activity,

location, and proximity to other users, the large-scale dynamics of collective human behavior can be analyzed. If deployed within a group of people working closely together, correlations between the phone log and proximity log could also be used to provide insight behind the factors driving mobile phone use. The initial study presented a simple model of behavior in three states: home, work, and elsewhere, the ultimate goal was to create a predictive classifier that can perceive aspects of a user's life more accurately than a human observer (including the actual user). So one central intent of this research was to verify the accuracy of automatically collected data from mobile phones for quantifying social networks. The labels for the classifier came from a survey taken by all the experimental subjects at the end of two months of data collection. The survey asked who they spent time with, both in the workplace and out of the workplace, and who they would consider to be within their circle of friends. These labels were compared with estimated location (using cell tower distribution and static Bluetooth device distribution), proximity (measured from Bluetooth logs), and time of day. Workplace colleagues, outside friends, and people within a user's circle of friends were identified with over 90% accuracy [2].

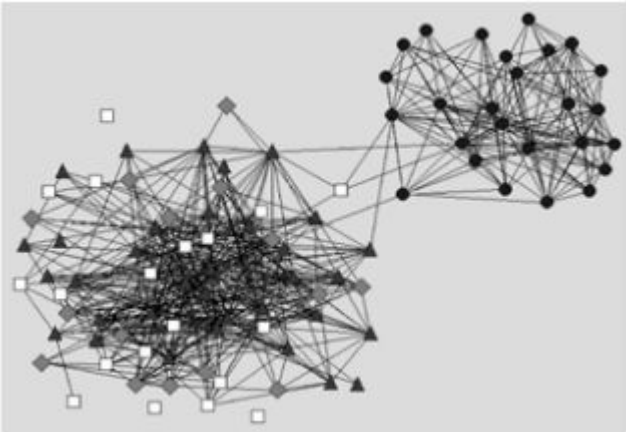

Fig. 1 Proximity pattern found in data. Source: [2]

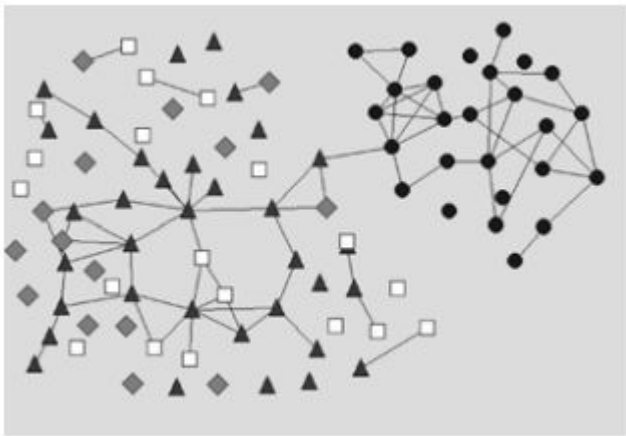

Fig. 2 Friendship network inferred from proximity pattern. Source: [2]

Weselowski et al. [13] analysed the daily travel patterns of 14 816 521 (approx. 38% of the Kenyan population) individuals across Kenya from June 2008 to June 2009 using nearly 12 billion calls and text messages to estimate daily locations for each one. They used a common measure of individual mobility, the radius of gyration, to examine how mobility patterns varied across the country on the district and population levels. This measure reflects both the frequency of travel and distance. For each person, they calculated a radius of gyration value over the year and then aggregated these population values to the district level. They couple this analysis with the results from a survey of socioeconomic status, mobile phone ownership and usage patterns across the country, providing regional estimates of population distributions of income, reported airtime expenditure and actual airtime expenditure across the country. For each district, they used the CDR (call data records) data to calculate the average mobility of individuals within discretized ranges of actual airtime expenditures. In the majority of districts individual mobility and airtime expenditure were positively correlated in the CDR data, also the monthly income and expenditure on airtime reported in survey were positively correlated.

Miklas et al. [8] examined how mobile systems could exploit people's social relations, in particular interactions between friends, that is people who meet more regularly and for longer periods of time, and interactions between strangers, that is people who meet sporadically, by passing each other by. The authors performed a social-based analysis of a trace of Bluetooth activity to annotate it with the required information. By studying the frequency of encounters, they could annotate this trace with social information by classifying pairs of people who encounter frequently as "friends", whereas pairs of people encountering sporadically are classified as "strangers". They used a 101-day trace of encounters between people equipped with Bluetooth-enabled mobile phones collected by the "Reality Mining" project at the MIT Media Lab [2] with the goal to characterize the key temporal and social parameters of people's encounters from this trace.

3.2 Public health and medicine

Mobile phones can be particularly useful in gathering health - related information. Data on a sample population over a given period (a week or a month) can be used, and then assuming some of people are sick, a model to predict how an illness spreads can be created.

A well known example of this kind is the work by Buckee et al. [1]; a study of the role of human mobility in the dissemination of malaria parasites in Kenya. Local and regional movements between areas with different malaria risks have several consequences for the transmission and epidemiology of the disease: individuals from low malaria risk regions traveling to high risk regions are particularly susceptible to disease because they lack well-developed immune responses, individuals from high risk regions traveling to low risk regions may carry parasites with them, the movement of people between different endemic regions can bring together populations of parasites that would otherwise remain genetically distinct. Traditional approaches to measuring human movements on these regional scales rely on survey data from national censuses or other household surveys. Mobile phones offer individual-level information on a scale previously impossible, providing a "big data" approach to understanding human mobility. The authors discussed the opportunities for measuring human mobility using data from mobile phones, as well as some of the issues

associated with combining mobility estimates with malaria infection risk maps to meaningfully estimate routes of parasite importation. When compared the data from CDR (for 34861 phone users) with travel data collected as part of a malariometric survey (for 2650 persons), they found out, that mobile phones allow to collect more accurate data in significantly larger scale and thus allow to create better models.

3.3 Trafic monitoring and control

After the September 11 attacks, U.S. officials scrambled to secure other national landmarks that might become future targets. One of them was the California's Golden Gate Bridge. Might terrorists try to destroy the passageway between San Francisco and Marin County, and if so, how widespread would the fallout be? To answer those questions, the Homeland Security Dept. turned to a small company called Inrix that uses GPS-enabled mobile phones and tracking devices installed on commercial vehicles to monitor traffic conditions. Inrix used its models to predict that the loss of the 1.7-mile bridge would result in immediate transport chaos. But Inrix found that within four days the situation tends to stabilize because people know what is happening and adjust their plans [6].

The *Mobile Century* experiment, carried out at the University of Berkeley on February 8, 2008 was intended as a proof-of-concept of traffic monitoring system based on vehicles equipped with GPS-enabled mobile phones. The goal of this controlled field experiment was to test traffic data collection from GPS-equipped mobile phones driving on a stretch of a highway located in the San Francisco Bay Area. One hundred vehicles carrying the GPS-enabled Nokia N95 drove along a 10-mile stretch of I-880 from 9:30am to 6:30pm. The experiment proved that data from GPS-enabled mobile phones alone were sufficient to infer traffic features, i.e., to construct an accurate velocity map over time and space [5]. The *Mobile Millennium* project grew from Mobile Century with a strategic objective to demonstrate the potential of GPS in mobile phones to alter the way traffic data is collected, by leveraging the existing mobile phone infrastructure to collect data and transmit it directly back to drivers. The public-private research partnership — UC Berkeley, Nokia Research Center, and NAVTEQ, with sponsorship from the California Department of Transportation — launched the pilot program from the Berkeley campus on November 10, 2008. It ran for exactly 12 months. During that time, more than 5,000 users downloaded the Mobile Millennium traffic software onto their phones (http://trafic.berkekey.edu).

Fig. 3 Snapshot of Mobile Millennium Traffic in San Francisco and the Bay Area. Source: [5]

Santani et al. presented an empirical analysis of the GPS-enabled taxi dispatch system in Singapore [11]. The data set records the movement of 6,230 GPS-enabled taxicabs over a 24-hour period, as well as the time and location of 38,048 booking requests. The main goals of the analysis were (a) to characterize the efficiency of the taxi system, and (b) to explore the sources of inefficiency. The two types of analysed data (taxi locations, booking requests) were firstly used to comparing the density of taxis with the frequency of bookings in each of four geographic zones. The results of this first step are then used to identify times and locations in which the demand for taxis exceeded the supply or vice versa. They then analysed the efficiency of the taxi system by defining driver satisfaction (as ratio of time the taxi was occupied and time the taxi was free or occupied) and passenger satisfaction (as median waiting time in given zone and given hour).

3.4 Smart homes and ambient assisted living

Smart Home and Ambient Assisted Living systems utilize advanced and ubiquitous technologies including sensors and other devices that are integrated in the residential infrastructure or wearable, to capture data describing activities of daily living and health related events (for a generic structure of such a system see Fig. 4). The used technologies can allow for the detection of emergencies and provide the means to increase social interaction and minimize isolation for residents. Smart Home and Ambient Assisted Living systems are composed of different components one of them being a reality mining component. To illustrate, consider two examples of how reality mining can be used within ambient assisted living system. By taking advantage of special sensors in mobile phones, such as the microphone or the accelerometers, important diagnostic data can be captured. Clinical pilot data demonstrate that it may be possible to diagnose depression from the way a person talks--a depressed person tends to speak more slowly, a change that speech analysis software on a phone might recognize more readily than friends or family do. Similarly, monitoring a phone's motion sensors can also reveal small changes in gait, which could be an early indicator of ailments such as Parkinson's disease [10].

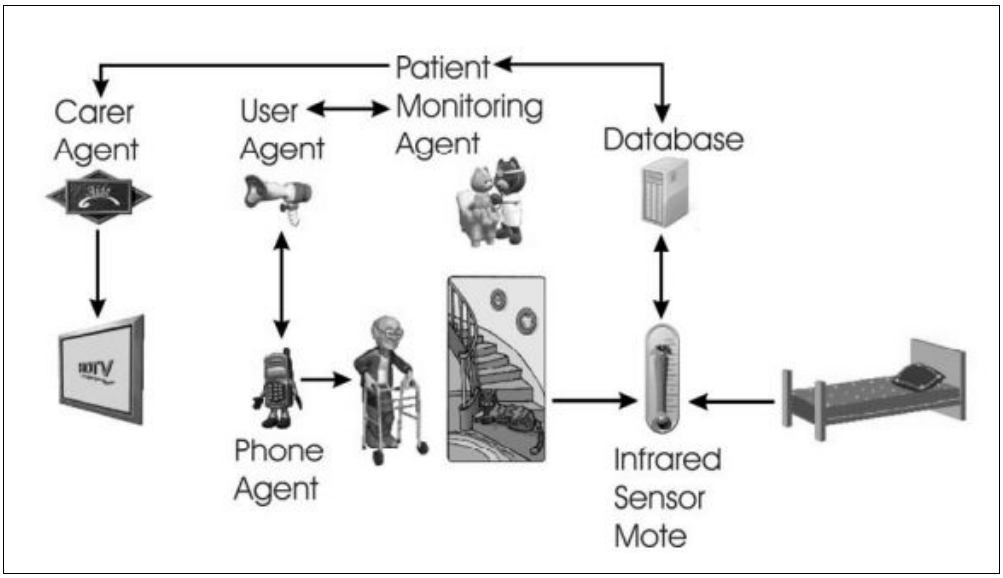

Fig. 4 A generic scheme of Ambient Assisitve Living Systems. Source: [9]

3.5 Environmental monitoring

Noise pollution is a major problem in big cities around the world. It affects human behaviour, well-being, productivity and long-term health. Recognising noise pollution as an important issue, the European Commission adopted the European Noise Directive, which requires major cities to establish a noise management policy. The first step in the establishment of such a policy is to assess the current situation by gathering real-world data and building strategic noise maps. A standard way how to get data is using sensor networks. However, this approach has several limitations: sparsity of network, fixed location of sensors and high cost. An alternative is thus to use mobil phones to collect the data about noise pollution.

Maisonneuve et al. propose a low-cost approach involving the general public to monitor noise pollution using their mobile phones as noise sensors [7]. The *NoiseTube* is a research project started in 2008 at the Sony Computer Science Laboratory in Paris in collaboration with Vrije Universiteit Brussel. The NoiseTube platform consists of two components, the mobile application and the Web-based community memory. The mobile application can be downloaded for free (e.g. from http://play.google.com for Android based phones) to turn mobile phones into mobile noise sensors. It collects information from different sensors (microphone, GPS receiver, user input) which is logged locally and/or sent to NoiseTube community memory server in real-time. The collected data are thus composed not only of the sound level recorded by the phone's microphone, but also of geographical coordinates, and annotations added by the users. The community memory operates on a central Web-server and collects and post-processes all gathered noise pollution measurements and runs a website which lets users explore, visualise, analyse and search through the data. The main post-processing performed on the server is aggregating info obtained fromdifferent users, automatic contextual tagging and GPS correction. NoiseTube thus goes beyond traditional noise maps due to the new nature of the collected data – real, local and personal exposure measurements with additional semantic information.

Fig. 5 Collective noise map for part of Paris, France. Source: [7]

4 Conclusions

The concept of reality mining is closely related to ubiquitous knowledge discovery (and one of the first papers on reality mining was published in ubiquitous computing journal). Anyway, the term reality mining sounds better.

The main source of data for reality mining applications are mobile phones equipped with GPS receivers. Researchers say they can get a more accurate picture of what people do, where they go, and with whom they communicate from a device they carry than from more subjective sources, including what people say about themselves. In short, people lie - mobile phones don't.

Although the example applications of reality mining discussed in section 3 do not use soffisticated machine learning algorithms, they still can be cosidered as data mining tasks aiming (at lest) at data description and summarization.

Reality mining seems to be a perspective field in which some applications (helth status monitoring, environmental monitoring or traffic monitoring) can directly influence our daily life. But if reality mining belongs to 10 most important technolgies, as published in MIT Technology Review [4] remains an open question.

Literature

1. Buckee, C. O., Wesolowski, A., Eagle, N., Hansen, E. and Snow, R. W., 2011. Mobile phones and malaria: modeling human and parasite travel. *Travel Med Infect Dis,* 11(1), pp. 15-22.
2. Eagle, N. and Pentland, A.S., 2006. Reality mining: sensing complex social systems. *Personal Ubiquitous Computing*, 10(4), pp. 255–268.

3. Gama, J. and May, M., 2011. Ubiquitous Knowledge Discovery, *Intelligent Data Analysis,* 15(2011) pp. 1–2.
4. Greene, K., 2008. 10 Breakthrough technologies 2008: TR10 Reality Mining. *MIT Technology Review*, March/April 2008. Available through http://www2.technologyreview.com/ [Accessed 28 August 2014].
5. Herrera, J. C., Work, D. B., Herring, R., Ban, X. J., Jacobson, Q. and Bayen, A. M., 2010. Evaluation of Traffic Data Obtained via GPS-Enabled Mobile Phones: the Mobile Century Field Experiment. *Transportation Research C*, 18, pp. 568–583.
6. Hesseldahl, A., 2008. There's Gold in 'Reality Mining', Bloomberg Businessweek, Available through http:// http://www.bloomberg.com [Accessed 28 August 2014].
7. Maisonneuve, N., Stevens, M. and Ocha, B., 2010. Participatory noise pollution monitoring using mobile phones. *Information Polity* 15 (2010), pp. 51–71.
8. Miklas, A. G., Gollu, K. K., Chan, K.K.W., Saroiu, S., Gummadi, K. P. and de Lara, E., 2007. Exploiting Social Interactions in Mobile Systems. In: *9. Int. Conf. on Ubiquitous Computing,* Springer LNCS 4717, pp. 409-427.
9. O'Grady, M. J.,. Muldoon, C., Dragone, M., Tynan, R. and O'Hare, G. M. P., 2010. Towards evolutionary ambient assisted living systems. *Journal of Ambient Intelligence and Humanized Computing* 1 (1), pp. 15-29.
10. Pentland, A., Lazer, D., Brewer, D. and Heibeck T., 2009. Using reality mining to improve public health and medicine, *Stud Health Technol Inform*. 149, pp. 93-102.
11. Santani, D., Balan, R. K. and Woodard, C. J., 2008. Spatio-Temporal Efficiency in a Taxi Dispatch System. *Research Collection School Of Information Systems,* Paper 9. Available through http://ink.library.smu.edu.sg/ [Accessed 3 September 2014].
12. Yang, Q. and Wu, X., 2006. 10 Challenging Problems in Data Mining. *International Journal of Information Technology & Decision Making,* Vol. 5, No. 4, pp. 597–604.
13. Weselowski, A., Eagle, N., Noor, A. M., Snow, R.W. and Buckee, C. O., 2013. The impact of biases in mobile phone ownership on estimates of human mobility. *J. R. Soc. Interface*, 10(2013). Available through http://rsif.royalsocietypublishing.org [Accessed 2 September 2014].

Contact data:

Petr Berka, prof. Ing. CSc.

University of Finance and Administration, Estonská 500, 10100 Praha 10, Czech Republic

berka@vse.cz