USAGE OF REALITY MINING IN LOCATION BASED ACTIVITY RECOGNITION

Ms. Poonam Bhagwandas Godhwani*

Abstract

The most important changes may come, however, from the fact that mobile phones can be used for reality mining. Their reality mining functionality is mostly latent at this point, but already these devices are being used to measure population flows into cities and slums, to map the movement of populations during emergencies, to identify neighbourhoods where social services are inadequate. The reality mining functionality of mobile phone networks is what governments in countries such as India are using to track terrorists, and they claim that the vast majority of captured terrorists have been identified through mobile phone transactions. The ability of mobile phone networks to identify unusual patterns of movement and communication. We can also use reality mining of mobile phone GPS data, call logs, and email records to better understand the "traffic" within an organization. Analysis of these digital traces allows a detailed picture of face to face, voice, and digital communication patterns. These patterns, in turn, allow us a new level of insight into the problems of industry and government, including building customer relationships, resource management, transportation, and public health. In our studies of many different types of companies, we have found that the trade off between face-to-face and email communication is a critical predictor of both productivity and job satisfaction. In fact, variations in these patterns can account for more than 30 percent of the differences in productivity between different parts of the same company! Patterns of *change* within these communication networks are also a critical predictor of creative productivity, accounting for up to 40 percent of differences in creative output.4 In short, reality mining is beginning to provide the capacity to collect and analyze data about people with a breadth and depth that was previously inconceivable. Current work using

* Teaching Assistant, Department of Computer Science & Technology, UTU-Bardoli (Gujarat)

reality mining techniques is underway in a variety of applications. For example, reality mining studies of voice, communications, and mobility patterns have already demonstrated the ability to screen for depression, infer quality-of-life metrics, and develop financial indexes for individual neighbourhoods [4].

Keywords

Reality Mining, Socio-geographic, Ubiquity, GPS, Human-centric, supervised learning, topic model, Noise, Entropy

1. Introduction

Sensors are everywhere, continuously gathering information as we live our daily lives. We live in a technology driven society where, each one of us continuously leaves digital traces behind. Whether using email, the telephone, a bank machine, or even simpler activities such as driving, using a photocopy machine, and a camera, all of these activities leave traces of our behaviour. Recently, these devices have been viewed from an engineering perspective as sensors, capturing data which scientists in many disciplines are very excited about. This data potentially impacts every one of us as researchers begin to study the possibilities of their use. Applications to society as a whole are being investigated in terms of epidemiology and psychology [1], urban planning [2], security, and even in the analysis of poverty [3].

Reality Mining is the study of human social behaviour based on wireless mobile phone sensed data. Mobile phones are particularly promising as sensors due to their vast usage over the world on a daily continuous basis, and also due to the numerous types of sensors embedded in the device. Not only do people carry them around as they live their daily lives sensing their location, and motion (via the accelerometer), their interactions can also be captured by Bluetooth, not only with other individuals carrying phones, but also with computers in proximity. There are many other forms of human behaviours that can be sensed with mobile phones, one of which is the reason they were invented in the first place, for communication. The mobile phone has developed, due to its paramount nature, from a simple communication device to include many other tools such as cameras, browsers, games, calendars, alarm clocks, and will surely continue to develop in the future. All of these forms of data can be analyzed to reveal details about human behaviour.

The focus here is on large-scale socio-geographic data obtained by mobile phone sensors capturing real-life location and proximity data. Our goal is to mine human activities and routines from this socio-geographic data. We define routines to be temporal regularities in peoples' lives. Activity modelling is a relatively recent domain in computer science, and of great interest in a new discipline named Computational Social Science [5].

2. Division of Reality Mining Dataset

Reality mining project introduced for sensing complex social systems with data collected from 94 mobile phones. Reality Mining considers Mobile phones as wearable sensors allow studying both individuals and organizations. We can divide reality mining dataset [1] into six categories:

1) Subjects (Volunteers) data: related to individual personal information like working time, which group he belongs to.

2) Self Reported Data: related to surveys results which represent what people think.

3) Locations Data: in reality mining dataset locations associated with mobile phones cell tower IDs, as each cell tower represent unique location. So locations data related to which cell tower user belongs to during the time.

4) Applications Data: related to which application used during the time for each subject.

5) Communications Data: related to user communications log data including type of communication (i.e. voice call, short message...), direction (i.e. Outgoing / Incoming) and duration.

6) Bluetooth Scanning Data: related to observed devices by subject's mobile phone each Bluetooth scanning time.

Reality mining raise interested questions related to user modelling. From reality mining dataset structure we can see that each category represent a different perspective. Self reported data represent what users think about friendship and spending time in work and so on. Others phone data can represent the actual events and relations by extracting the model for each concerned pattern. The convenience and ubiquity of the mobile phone are changing the way in which we interact with the information world around us. Mobile phone are no longer simple calling devices, recent innovations have empowered mobile phones to monitor their children, being able to play 'treasure hunt games', access their email, and more recently search the world wide web through a

voice interface! With ubiquitous connectivity also comes the ability of cell phones to act as natural sensors. This is increasingly being used to get more accurate data and analysis of group dynamics than was ever possible before. With this context it is both an exciting and an important time to compare mobile social networks with the social networks created through face-to-face and Internet mediums.

3. Current Challenges

As large scale data collections on human behaviour become more readily available, the need for effective methods and mathematical models for analysis becomes crucial in order to make good use of the sources. In machine learning, algorithms have been developed to recognize complex patterns and make intelligent decisions based on data. Traditional machine learning models are recognized as useful tools for large scale data analysis [6, 7]. They have been used in the domain of human behaviour analysis, though their limitations with new types of data and human-centric questions become apparent. For example, many of the traditional machine learning models is supervised, requiring training data which is often impossible or illegal to collect on human subjects. Other specifications related to human-centric data include the multimodal aspect, the noise, the massive quantity, and the complex questions of interest. More specifically, data collected by mobile phone sensors include many types, ranging from GPS, Bluetooth, accelerometer, to voice features. Each of these sensors may be sampled with varying frequencies, each has varying timescales and differing characteristics, and each has its own sources of noise. For example, a person can interact with more than one other person at a given time; some potential features to model include the group size, identity, and relationship with those in proximity. Noise is inherent in human behaviour in relation to sensors. People forget their phones, lend them to friends, and set the time incorrectly, forget to turn on Bluetooth, and most importantly the phone is not always attached to the individual [8]. Each sensor has its particular problem. For example GPS does not work indoors and Bluetooth detects devices through thin walls. Continuous data on large populations pose many computational challenges. For example traditional techniques using linear algebra are not easily applicable due to large matrix operations. Finally, and most importantly, traditional machine learning techniques have not been designed to target the questions of interest. For example, a typical question of interest would be “what are the differences in the daily routines between two populations (e.g., of students)?” We do not know

what machine learning tools would best solve this problem. There are several state-of-the art classification tools to divide the groups with lowest error; however, we are interested in finding the similar and differentiating features within these groups and understanding how significant they are. We believe machine learning methods can provide useful algorithms for large-scale activity modelling, as we show in this thesis. However they do present their limitations in face of human-centric data and can be built upon. Though machine learning techniques are very useful techniques for large-scale human behaviour analysis, we believe they pose limitations and can be used as a basis for new methods targeting the field of Computational Social science. Here we choose to investigate probabilistic topic models as the basic tool. Topic models are chosen first and foremost for their unsupervised nature. They have several other advantages making them an attractive choice for the goals of this thesis. Their probabilistic, generative nature makes them attractive over discriminative approaches since we are interested in modelling the nature of the data. The input feature structure in topic models, called bag of words, is critical in removing redundant time information and is key in handling the large amounts of noise in the data. Very importantly, topic models can be applied to very large data collections. They also can be manipulated in various ways to integrate multiple data types.

4. Related Work

There has been a significant amount of work on social network analysis of online interactions. Studies have focused on a wide range of issues such as characterizing online interactions, effect of the internet on real life interactions, understanding real world social networks from online interaction structure and the study of email networks to characterize tie formation [9].

5. Co-relation between calls and Location

And finally we study correlations between calls the spatial location. The analysis includes studying the relation between the duration of phone calls and the location of the people the call is between. We haven't come across any research that explicitly studies that. Analyzing this can help us glean some insights at a macro level analysis between call duration and the location of the callers, and perhaps lead to some design ideas.

6. Location-Based Activity Recognition

Our goal is the discovery and analysis of human routines which characterize both individual and group behaviours in terms of location patterns. Our focus is the automatic discovery of human activities and routines from mobile phone location data collected by one hundred individuals over the course of a year. Automatic routine classification and discovery are in general challenging tasks as people's locations often vary from day to day and from individual to individual, and data from sensors can frequently be incomplete as well as noisy. A supervised learning approach to activity recognition would require prior knowledge in the form of predefined activity categories and labelled data [10]. In contrast, an unsupervised learning approach has the potential of automatic discovery of emerging routines of people not requiring training data. Through discovery, sifting through large amounts of noisy data becomes possible. Further, one can cluster data (i.e., people or days) corresponding to the most common routines (those of several people) and discover the dataset structure with minimal prior knowledge. We show that topic models prove to be effective in making sense

of behavioural patterns at large-scale while filtering out the immense amount of noise in real-life data. Here we present an overview of our approach to large scale human activity mining from location data. We present details of each component, beginning with the bag construction then topic models and finally we present the experiments and results. Our overall goal is to determine what human routines are contained in mobile tower connection data and how to discover them in an unsupervised manner. As described earlier, we represent a day in the life of an individual in terms of their locations obtained by cell tower connections and use this information to form a bag of location sequences. This bag representation was carefully designed to capture dynamics (i.e., location transitions) as well as both fine-grain (30 minute) and coarse-grain (several hour) time descriptions. Overall, we make an analogy between the bag of location sequences for mobile data and a bag of words for text documents, where a location sequence is analogous to a text word, a day in the life of a person is analogous to a document, and a person is analogous to the author of a document. We use two models to discover routines. Our proposed methodology based on a bag of location sequences structure is advantageous over [11] in that it contains both fine-grain and coarse-grain time considerations, which keeps into account transitions in location and is robust to variations in the data which may be due to noise or due to variations in the dataset, such as eating lunch at 11:30 am as opposed to 11:55 am. Further, due to our location sequence structure, we can discover routines characteristic of various intervals in the day. Our topic model methodology

clearly defines a mechanism to rank users and days (with probabilities always greater than zero), and with easily identifiable routines with semantic meanings which can be visualized comprehensibly over many of the discovered topics. Ranking allows us to see the raw data in a particular order (given by probabilities), giving structure to the data. This is true for both users and days and is useful for visualizing and structuring the data. We can perform several tasks with the discovered data, such as find users that go to work early, find groups of users that are at home during the day, or find users that turn their phones off in the morning.

7. Conclusion

We presented a novel application of pervasive sensing using mobile phones modelling the spread of political opinions in real-world face-to-face networks. Using mobile phone sensors, we determined user's behaviours discovered as topics, where features include opinions, amount of interaction, and amount of phone communication as well as relationship information. We consider groups that changed opinion versus those that did not, and observed statistically significant differences in the entropy of topic distributions. This indicates that people who changed preferred party often discussed face-to-face with their democrat political discussants, and their daily routines included heavy phone and SMS activity. We also found that people who decreased their interest in politics often interacted with people who have little or no interest in politics. One limitation of this methodology is that survey questionnaire results are noisy and can only be obtained in small scales in comparison to mobile sensor data. This limits the scope of the results and analysis. This work clearly represents a first attempt. We can anticipate several future extensions of this work. In addition to political opinions, it would be important to understand if pervasive sensing methods can help understand the propagation of other types of opinions and habits in face-to-face networks, e.g., those related to health or purchasing behaviour, both in our current dataset and also in other observational data. Overall, our quantitative analysis has the potential of shedding more light on long-standing open questions in political science and other social sciences, about the diffusion mechanism for opinions and behaviours, but further studies are obviously needed to realize the actual limitations of this approach.

8. Limitation

We have shown many insights into activity modelling, our work has some limitations. The first one relates to the scope of the data collection and features. However, the Reality Mining dataset did contain a mixture of business students as well as engineering students, and their routines are likely representative of many students and working professionals. Further, some colleagues have used our methods, confirming many of the dominant activities on data collection. Another limitation is that we consider location data from cell tower connections, and reduce over 32 000 possible locations into four categories, home, work, out, and no reception. The proximity data considered is also limited in that we only consider being in proximity with other individuals from the data collection. We began some initial investigation in considering proximity to known people, strangers, laptops, and computers, however it was not pursued due to lack of ground truth. The coarse-grain timeslots considered over several hour intervals are another limitation of the methods used. The second limitation of this work stems from the noise in the data. For example, what is sensed by Bluetooth proximity is often a small subset of real face-to-face interactions since most people are not carrying their mobiles in indoor settings. There is no way to account for this problem in the dataset we used and it is a limitation of all current Reality Mining studies. Other sources of noise relating to Bluetooth are the detection of other devices through thin walls where our results would reveal the two individuals are in proximity though they are not.

9. Future Work

The location and Bluetooth interaction features could be explored in different ways. For example, location information can also be obtained by GPS data and WLAN information. The development of methods to encapsulate richer location features, for example cell tower identities or GPS hot spots, would reveal more detailed location information. Bluetooth devices include several categories. A device in proximity may be that of a friend or a stranger. It may also be a laptop or a computer, either owned by you or a friend or stranger. An extension of this work would be to investigate the discovery of Bluetooth devices and their types given ground truth could be collected. Knowledge of Bluetooth devices and interacting user types would be of great interest in activity discovery. In future work, the methodology for data prediction could be further optimized to use the topics in a more sophisticated manner, and to include prediction on varying timescales, such as full days of missing data. It would also be very useful to take advantage of the other, often available data modalities of mobile sensor data for data prediction. For instance, one could predict

a user's location given the time of day and their interactions, the day of the week, or even using their phone call and SMS data. The Bluetooth proximity data is potentially a very rich source if one considers proximity to all other devices including laptops, computers, and anonymous cell phones in predicting missing data. This data in itself could be used to determine the semantic labels of an individual, such as if the user is at home (in proximity with their home computer), at work (in proximity with their work computer), or out (in proximity with strangers). In a different line of work, we would like to enrich the location vocabulary by refining the "other" category. This in principle could be done from the Reality Mining dataset, but handling sparse human annotation of places is in itself a research problem.

References

[1] Madan, A., Cebri´an, M., Lazer, D., and Pentland, A. (2010a). Social sensing for epidemiological behaviour change. In Ubiquitous Computing (UbiComp), Copenhagen, Denmark

[2] Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. (2008). Understanding individual human mobility patterns

[3] Wesolowski, A. and Eagle, N. (2010). Parameterizing the dynamics of slums. In AAAI Spring Symposium on Artificial Intelligence for Development.

[4] Pentland, A. 2008. Honest Signals: How they Shape our World. Cambridge, MA: MIT Press

[5] Lazer, D, Pentland, A., Adamic, L., Aral, S., Barabasi, A., Brewer, D., Christakis, N., Contractor, N, Fowler, J., Gutmann, M., Jebara, T., King, G., Macy, M., Roy, D., and Van Alstyne, M. (2009). Computional social science

[6] Duda, R., Hart, P., and Stork, D. (2000). Pattern Classification (2nd Ed.). Wiley-Interscience.

[7] Bishop, C. M. (2007). Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 1st ed. 2006. corr. 2nd printing edition.

[8] Patel, S., Kientz, J., Hayes, G., Bhat, S., and Abowd, G. (2006). Farther than you may think: An empirical investigation of the proximity of users to their mobile phones. In Proc. of Ubicomp, Orange County, California, USA.

[9] G. Kossinets and D. J. Watts. Empirical analysis of an evolving social network. Science, 311(5757):88-90, January 2006.

[10] Liao, L., Fox, D., and Kautz, H. (2006). Location-based activity recognition. In Advances in Neural Information Processing Systems (NIPS), Vancouver, Canada.

[11] Eagle, N. and Pentland, A. (2009). Eigenbehaviors: Identifying structure in routine. BehavioralEcology and Sociobiology.