Social genome mining for crisis prediction

Peter Wlodarczak
University of Southern Queensland
West Street, Toowoomba Qld 4350
Australia
+61 7 4631 5543
wlodarczak@gmail.com

Siyu Qian
University of Wollongong
NSW 2522
Australia
+61 2 4221 3218
sq992@uowmail.edu.au

Mustafa Ally
University of Southern Queensland
West Street, Toowoomba Qld 4350
Australia
+61 7 4631 5543
Mustafa.Ally@usq.edu.au

Jeffrey Soar
University of Southern Queensland
West Street, Toowoomba Qld 4350
Australia
+61 7 4631 5543
Jeffrey.Soar@usq.edu.au

ABSTRACT

With the increasing digitalization of western societies, we leave more and more digital traces in countless databases. These digital information include health records, bank account information, school grades etc. and they form our social digital footprint. The digital traces of a society as a whole form what is called the social genome and its analysis in real time or near real time has the potential to transform healthcare, the way we do business and even our behavior. The social genome can also be used to predict crises and can possibly prevent conflicts in future hot spots. However, the analysis of these vast amounts of data poses technical challenges and raises serious privacy issues. SM analysis has become a very active area of research and SM is mined for marketing purposes, to detect new trends, to get user opinions on products and services and to predict crises. This paper describes some of the methods used to analyze SM posts for crisis prediction by determining the social genome.

General Terms

Algorithms, Management, Measurement, Security, Human Factors.

Keywords

Social digital footprint, social genome, population informatics.

1. INTRODUCTION

Digital information throughout our life is stored in numerous databases from birth certificates to obituaries. Collectively, these digital traces - across a group, town, county, state, or nation - form a population's *social genome*, the footprints of our society in general [1]. Analyzing the social genome can reveal crucial information about a society and transform healthcare, economics, education and homeland security. Social scientists can now analyze these large, distributed datasets in near real time and a new field of research called *population informatics* has emerged. Population informatics is the burgeoning field at the intersection of social sciences, health sciences, computer science, and statistics that applies quantitative methods and computational tools to answer questions about human populations [1].

This paper was presented at the First International Workshop on Population Informatics for Big Data (PopInfo'15), Sydney, 10 August 2015. Copyright of this work is with the authors.

In whole genome sequencing (WGS), human genes of individuals are sequenced with the goal of predicting the susceptibility of individuals to certain diseases. One application is predicting disorders based on mining the genotype and understanding how the interactions between genetic loci lead to certain human diseases [2]. Similarly to WGS in individuals, *population sequencing* analyses social genome data at the population level to gain insights into how people live, react to change, make decisions and can help understand the root causes of social and public health problems and predict crises before they occur.

Big Data has provided us with the technologies to analyze such large amounts of data from different data sources and in various formats (Distributed Data Mining, DDM). However, there are many challenges with maintaining privacy and confidentiality and access to many of the digital data sources is restricted, leaving the majority of the databases untapped. Also disambiguating individuals proses a major challenge. For instance identifying James in a healthcare database, Jim in an education record and Jim123 on Twitter as the same person is called *Record Linkage* (RL) and poses a major problem in DDM. RL is an important data pre-processing step in DDM used for data deduplication and reconciliation [20]. The process of ingesting, disambiguating, and enriching data from disparate data sources is referred to as *knowledge based synthesis*. In population informatics the goal is to transform raw administrative data into knowledge that can support decision making.

Probably one of the most well-known social genome projects is Google flu. It uses a combination of medical records and search query terms to predict flu outbreaks before they occur [19]. Praised as a success, some authors have doubted its effectiveness [3]. Google flu wrongly predicted several outbreaks and missed out on others.

A relative new area of social genome mining is predicting crises. Whereas forecasting of hot spots based on news and reports in political science is not new per se, only recently the amounts of digital available information and its accessibility has opened new possibilities in predicting future conflicts. A recent study analyzed news reports and the onset of war could be predicted in 85% of the cases [4]. The widespread use and the accessibility of Social Media (SM) such as Twitter and Facebook have made it a very popular area of research. SM has been used to organize the Arab spring (Arabellion) and to organize the

"Occupy Wall Street" protests in New York's Zuchotti park [5]. This suggests that analyzing SM, SM listening, could have the potential to detect and predict arising conflicts. SM are sensors for tensions and the information in SM posts could be used as an early warning system to prevent future crises. Using SM for crisis analysis and forecasting has attracted many researchers and several studies have been conducted recently [6], [7], [8], [18]. Traditional mass media sources distribute news based on information from local journalists and are based on observations by one or a few persons. Technological challenges in areas afflicted by crises (i.e., down satellite connections, etc.) may slow official news correspondent reports, but SM reports may be much more swiftly distributed [6]. SM posts distribute information from many observers on the same events from different angles. Some reports might be less credible and need to be weighted less. Perceived source credibility becomes an increasingly important variable to examine within SM, especially in terms of crisis and risk information [6]. Determining the SM genotype is a possibility to determine a user's credibility and possibly spam post. The genotype is a per-topic summary of a user's interest, activity and susceptibility to adopt new information [10].

When making crisis predictions based on news and SM, correctly categorizing the posts is a crucial step. The reports are likely to differ in content depending on the stage of the crisis: early signals, escalation, outbreak, etc. We also know that the information contained in the messages, and the proportion of messages that will be posted in different categories can vary significantly across crises, to the extreme that in some cases a category of information that is very common in one disaster can be almost completely absent in another [8].

A possible solution to the before mentioned problems is determining the SM genotype. The next chapter describes some of the methods that can be used to determine the SM genotype and make predictions on future hot spots.

2. METHODS

Every day mountains of data are produced on SM and we need appropriate methods for analyzing these vast amounts of data to find useful and actionable patterns. Data mining provides the necessary tools for discovering patterns in data [9]. Data Mining (DM) is also called Knowledge Discovery in Databases (KDD). To discover useful patterns the posts have to be classified. Genotype-based classifiers capture unique traits and variations in different genes (topics) similar to a biological genotype. Variations in genotype over time can then be used to determine if there are early signals of a conflict, if a conflict is in the offing etc. The classifiers are represented as a time series to predict if an escalation is to be expected. The process goes through three steps, a data collection step, a data analysis step and a predictive step. These three steps are described in the next chapters.

2.1 Data collection

Many SM sites, by far not all, provide Application Programing Interfaces (API) to query or automatically collect real time data. For instance Twitter offers a "firehose" API to collect 100%, a "gardenhose" API to collect 10% and a "spritzer" API to collect 1% of real-time data. It also offers a query API to search historic data. Most SM sites limit free access to their data. For instance Facebook limits timeline access to the last 100 posts, which makes it unsuitable for analyzing historic data. Twitter's "firehose" access is restricted to paying users, and the query API is restricted to 180 calls per 15 minutes.

The data can be captured programmatically using any programing language such as Java, Python or C# and many libraries exist that facilitate the programming task. For instance Twitter4j or Spring social are free libraries for Java developers.

Other options are screen scrapers, especially if no API is provided. Screen scrapers often plug into browsers and simulate a user e. g. scrolling down the browser to access a Facebook timeline and capturing the content data of the site.

2.2 Data analysis

To determine the genotype of SM posts, the data has to be classified according to the state at which a possible crisis currently is. There are many different types of crises: war, terrorism, epidemics etc. This means at the onset of a crisis an expert must select a handful of information categories into which information will be categorized [8]. The simplest methods use word frequency analysis, i. e. how often words such as "war", "attack", "Ebola" occur in a post. These methods are called *term frequency* analysis. Often they also weight words as in *Inverse Document Frequency* analysis. Term Frequency and Inverse Document Frequency (TF-IDF) is defined as:

$$w_{t,d} = tf_{t,d} \times \log(\frac{N}{df_t}) \qquad (1)$$

where $tf_{t,d}$ is the number of occurrences of term t in the document d, N is the number of document in the collection and df_t, is the number of documents, in which term t appears [17]. However, usually word frequencies disregard the context in which it is used and more sophisticated methods should be preferred.

2.2.1 Unsupervised approaches

Unsupervised approaches analyze text documents with the aim of discovering abstract, latent topics. An often used unsupervised approach for text categorization is Latent Dirichlet Allocation (LDA). LDA is a form of Latent Semantic Indexing (LSI). It is a generative, probabilistic topic modelling technique. A latent variable in statistics is a variable that cannot be directly observed. The Dirichlet distribution is defined as:

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{n} \alpha_i)}{\prod_{i=1}^{n} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1}, \qquad (2)$$

Where every document d is characterized by Dirichlet distribution Θ_d with parameter α, n is the number of predefined topics and $\alpha >$ 1, and $\sum_{i=1}^{n} \theta_i = 1$. Γ(x) is a Gamma function.

The documents d are in a specific category, for instance "genocide", n is the number of topics that needs to be extracted. The output of LDA is a matrix in which each cell indicates the extent to which a document d corresponds to a topic n. It represents the genotype of the document. To be useful for crisis prediction, the sequential structure needs to be represented too and sequential extensions to LDA have been proposed [16]. LDA has proven to be useful for exploratory as well as predictive text analysis.

2.2.2 Supervised approaches

Supervised approaches are used for classification and for regression. They are adopted when the class label is known. Unsupervised methods such as LDA can be used when the class label is unknown. Here the class label would be the latent topic. Supervised approaches cannot process text documents and take a feature vector as input. Text documents such as blog or forum posts are represented in the vector space using Vector Space Modelling (VSM) techniques. VSM translates documents into a vector with word statistic measurements such as word frequencies or part of speech (POS) tags. Creating a bag of words is a popular way of creating input vectors for supervised methods. A bag of words is a list of words disregarding word order or grammar. It is a method of *natural language processing* (NLP) and is used for instance in spam filtering or sentiment analysis.

Popular supervised approaches include *naïve Bayes classifiers, decision tree induction, k-nearest neighbor, artificial neural networks* (aNN) and *Support Vector Machines* (SVM). However there are many others. They are well documented in literature and are not further explained here. Supervised methods are useful when a problem cannot adequately be solved using a simple, probabilistic, rule-based model, when the model becomes too complex or when a model doesn't scale. They yield good approximations for very complex problems. Usually several models are trained and the best performing learner is chosen. For instance naïve Bayes classifiers are known to handle noise poorly and should be avoided in noisy environments. Decision trees are prone to overfitting and should be avoided when the decision rules are expected to become too complex.

Social genotypes are usually complex, high dimensional and fuzzy. Supervised approaches are well suited for handling the high dimensionality and fuzziness of genotype data. However, it should be noted that many classifiers output the probability *Pr*, that a region x_i with corresponding label y_i belongs to class *j* [2].

As mentioned before, genotype classification needs to identify the relevant categories [8]. To get good results for crisis prediction it is crucial to define appropriate categories. Typically supervised approaches go through many iterations using different categories until satisfactory results are obtained.

2.3 Predictive analysis

To be able to make predictions time series analysis can be used. They are useful to make predictions based on historic data about future events.

2.3.1 Time series analysis

Time series analysis aims to find correlations between variables in two time series. A time series contains for instance the frequency of certain topics related to disease outbreaks over a period of time. Granger causality analysis has been used for time series analysis. It analyses if a time series *X* has predictive information over a time series *Y*. For instance if in a historic time series certain topics in blogs showed an increased frequency before a crisis emerged, it might be an indicator that in the future, if a similar increase occurs, a new crisis is about to emerge.

For a strictly stationary bivariate process $\{(X_t, Y_t)\}$, $\{X_t\}$ is a Granger cause of $\{Y_t\}$ if past and current values of *X* contain additional information on future values of *Y* that is not contained in past and current *Y*-values alone [13]. The Granger causality test for two scalar-valued, stationary, and ergodic time series $\{X_t\}$ and $\{Y_t\}$ is defined as:

$$F(X_t \mid \mathrm{I}_{t-1}) = F(X_t \mid (\mathrm{I}_{t-1} - Y_{t-Ly}^{Ly})), t = 1,2... \qquad (3)$$

Where $F(X_t|I_{t-1})$ is the conditional probability distribution of X_t given the bivariate set I_{t-1} consisting of an L_x-length vector X_t and an L_y-length vector of Y_t. If the equality in equation (3) does not hold, then knowledge of past *Y* values helps to predict current and future *X* values, and *Y* is said to strictly Granger cause *X* [14].

The original Granger test only examined linear causality among time series, however non-linear and multivariate extensions exist. Also recently other methods such as the Phase Slope Index (PSI) have been preferred [15].

If there is a correlation between social genome data and real world events such as crisis or war outbreaks, social genome mining can be used to make predictions about future conflicts.

3. CHALLENGES

As mentioned before, social genome mining and population informatics pose serious privacy concerns and privacy-preserving data operations have become a hot topic. For instance privacy-preserving RL is a major challenge and an area where more research is needed [21]. RL can reveal someone's identity on a SM site in a country where unpopular comments on political events might entail serious consequences by the government. It is critical to distinguish between identity and sensitive attribute disclosure. Identity disclosure reveals the identity of an entity, (who is this individual?), attribute disclosure reveals the sensitive data, (is this individual HIV positive?). Secure Decoupled Linkage (SDLink) has been proposed for privacy-preserving record linkage [1]. SDLink decouples identifying data from attribute data through encryption and chaffing (adding fake data) mechanisms.

Even if a record by itself does not disclose sensitive information about an individual, a collection of several records potentially could. For instance, consider two queries, one for the total number of HIV infected citizens of a town and the number of HIV infected citizens other than James. Neither query discloses James HIV status, but combining the answers the result can be inferred.

Recent studies have shown that it is actually possible to deduce sensitive data from supposedly safe data sets by combining them together. For instance a study of credit card data showed that it takes only a tiny amount of personal information to de-anonymize people [11].

A possible solution is differential privacy (DP). DP, also called indistinguishability, aims to maximize the accuracy of queries in statistical databases while minimizing the chances of identifying the individual records. DP is a probabilistic model and hence it is necessarily random. Some rely on adding controlled noise, others sample from a problem-dependent distribution instead.

The freedom of speech in conflict areas is often curtailed and SM sites are censored which causes a shortage of news or SM posts to detect an arising conflict. Also, crises are often only reported if they reach a certain escalation level. Some researchers use "weak signals" in news to detect conflicts in areas where information is sparse [18]. Also, the network effect might influence the result. If a tool predicts a conflict and the press publishes it, the data analysis would be influenced. The resulting system would influence itself.

While social genome mining might have the potential to predict future crises, the appropriate answer to prevent the crises are still an open question that cannot be easily be responded to. Recent military interventions in North Africa and the near east following the Arabic spring have left a whole region in havoc and lead to refugee disasters in the Mediterranean and the Pacific Ocean. Early interventions following a crises prediction is no guarantee for success and can in the worst case just exacerbate the situation. Finding appropriate responses is a cross-disciplinary research area that can prevent possible human disasters. However, predictions might not be accurate and a lot of caution has to be applied in finding appropriate answers.

4. CONCLUTIONS

Social genome mining has the potential to predict future crises. Research on crisis prediction using social genome mining is still in its infancy and its accuracy for predicting future conflicts has still to be proven. As some studies have suggested, an early warning system could predict war up to one year before its outbreak, however, so far predictions have only be made successfully a posterior [4], [12].

Also there is no silver bullet for crisis prevention and finding the appropriate response is a very delicate matter. A wrong prediction might even prove to be harmful when triggering an inadequate response. Nevertheless early results have shown some promising results [4], [12],[18] and have the potential to prevent disasters such as refugee emergencies or disease spread.

Lastly social genome mining poses still many challenging but interesting research problems for new studies and more research would be highly desirable.

5. REFERENCES

[1] Kum, H.-C., Krishnamurthy, A., Machanavajjhala, A. and Ahalt, S. C. Social Genome: Putting Big Data to Work for Population Informatics. Computer, 47, 1 2014), 56-63.

[2] Wlodarczak, P., Soar, J. and Ally, M. Genome mining using machine learning techniques. Springer International Publishing, Switzerland, 2015.

[3] Lazer, D., Kennedy, R., King, G. and Vespignani, A. The Parable of Google Flu: Traps in Big Data Analysis. Science, 343, 6176 (March 14, 2014 2014), 1203-1205.

[4] Chadefaux, T. Early warning signals for war in the news. Journal of Peace Research, 51, (1 2014), 5-18.

[5] Gerbaudo, P. Tweets and the Streets: Social Media and Contemporary Activism. Pluto Press, 2012.

[6] Westerman, D., Spence, P. R. and Van Der Heide, B. Social Media as Information Source: Recency of Updates and Credibility of Information. Journal of Computer-Mediated Communication, 19, 2 2014), 171-183.

[7] Alexander, D. Social Media in Disaster Risk Reduction and Crisis Management. Sci Eng Ethics, 20, 3 (2014/09/01 2014), 717-733.

[8] Imran, M. and Castillo, C. Towards a Data-driven Approach to Identify Crisis-Related Topics in Social Media Streams. In Proceedings of the Proceedings of the 24th International Conference on World Wide Web Companion (Florence, Italy, 2015).

[9] Zafarani, R., Abbasi, M. A. and Liu, H. Social Media Mining. Cambridge University Press, Cambridge, 2014.

[10] Bogdanov, P., Busch, M., Moehlis, J., Singh, A. K. and Szymanski, B. K. The social media genome: modeling individual topic-specific behavior in social media. In Proceedings of the Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (Niagara, Ontario, Canada, 2013).

[11] Bohannon, J. Credit card study blows holes in anonymity. Science, 347, 6221 2015), 468-468.

[12] Löwer, C. Die Logik der Gewalt. Technology Review, 6, (2015).

[13] Diks, C. and Panchenko, V. A new statistic and practical guidelines for nonparametric Granger causality testing, Journal of Economic Dynamics and Control, vol. 30, no. 9–10, (2006), pp. 1647-1669.

[14] Hiemstra, C. and Jones, J. D. Testing for Linear and Nonlinear Granger Causality in the Stock Price- Volume Relation, The Journal of Finance, vol. 49, no. 5, (1994), pp. 1639-1664.

[15] Madan, A. Cebrian, M. Lazer, D. and Pentland, A. Social sensing for epidemiological behavior change, in Proceedings of the 12th ACM international conference on Ubiquitous computing, Copenhagen, Denmark, (2010), pp. 291-300.

[16] [20] Du, L. Buntine, W. Jin, H. and Chen, C. Sequential latent Dirichlet allocation, Knowledge and Information Systems, vol. 31, no. 3, (2012), pp. 475-503, 2012/06/01.

[17] Daniel, Z. Daniel, Z. Ján, S. Jozef, J. and Anton, C. Text Categorization with Latent Dirichlet Allocation, Journal of electrical and electronics engineering, vol. 7, (2014.), pp. 161-164, 05/01.

[18] Gao, J. Leetaru, K. H. Hu, J. Cioffi-Revilla, C. and Schrodt, P. Massive Media Event Data Analysis to Assess World-Wide Political Conflict and Instability, Lecture Notes in Computer Science, Springer International Publishing, Germany, 2013.

[19] Ginsberg, J. Mohebbi, MH. Patel, RS. Brammer, L. Smolinski, MS. and Brilliant, L. Detecting influenza epidemics using search engine query data, Nature, vol. 457, 2009.

[20] Vatsalan, D., Christen, P. and Verykios, V. S. A taxonomy of privacy-preserving record linkage techniques. *Information Systems*, 38, 6 (9// 2013), 946-969.

[21] Christen, P. Overview and taxonomy of techniques for privacy-preserving record linkage. *JSM*, 2013.