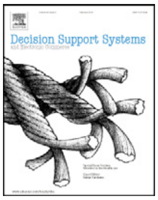
Decision Support Systems 61 (2014) 115–125

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss



Predicting crime using Twitter and kernel density estimation

Matthew S. Gerber *

Department of Systems and Information Engineering, University of Virginia, P.O. Box 400747, Charlottesville, VA 22904-4747, USA

ARTICLE INFO

Article history:
Received 11 September 2013
Received in revised form 28 December 2013
Accepted 13 February 2014
Available online 22 February 2014

Keywords:
Crime prediction
Twitter
Topic modeling
Density estimation

ABSTRACT

Twitter is used extensively in the United States as well as globally, creating many opportunities to augment decision support systems with Twitter-driven predictive analytics. Twitter is an ideal data source for decision support: its users, who number in the millions, publicly discuss events, emotions, and innumerable other topics; its content is authored and distributed in real time at no charge; and individual messages (also known as tweets) are often tagged with precise spatial and temporal coordinates. This article presents research investigating the use of spatiotemporally tagged tweets for crime prediction. We use Twitter-specific linguistic analysis and statistical topic modeling to automatically identify discussion topics across a major city in the United States. We then incorporate these topics into a crime prediction model and show that, for 19 of the 25 crime types we studied, the addition of Twitter data improves crime prediction performance versus a standard approach based on kernel density estimation. We identify a number of performance bottlenecks that could impact the use of Twitter in an actual decision support system. We also point out important areas of future work for this research, including deeper semantic analysis of message content, temporal modeling, and incorporation of auxiliary data sources. This research has implications specifically for criminal justice decision makers in charge of resource allocation for crime prevention. More generally, this research has implications for decision makers concerned with geographic spaces occupied by Twitter-using individuals.

1. Introduction

Twitter currently serves approximately 140 million worldwide users posting a combined 340 million messages (or tweets) per day [1]. Within the United States in 2012, 15% of online adults used the Twitter service and 8% did so on a typical day, with the latter number quadrupling since late 2010 [2]. The service's extensive use, both in the United States as well as globally, creates many opportunities to augment decision support systems with Twitter-driven predictive analytics. Recent research has shown that tweets can be used to predict various large-scale events like elections [3], infectious disease outbreaks [4], and national revolutions [5]. The essential hypothesis is that the location, timing, and content of tweets are informative with regard to future events.

Motivated by these prior studies, this article presents research answering the following question: can we use the tweets posted by residents in a major U.S. city to predict local criminal activity? This is an important question because tweets are public information and they are easy to obtain via the official Twitter service. Combined with Twitter's widespread use around the globe, an affirmative answer to this question could have implications for a large population of criminal justice decision makers. For example, improved crime prediction performance could allow such decision makers to more efficiently allocate police patrols and officer time, which are expensive and thus scarce for many jurisdictions.

However, there are many challenges to using Twitter as an information source for crime prediction. Tweets are notorious for (un)intentional misspellings, on-the-fly word invention, symbol use, and syntactic structures that often defy even the simplest computational treatments (e.g., word boundary identification) [6]. To make matters worse, Twitter imposes a 140-character limit on the length of each tweet, encouraging the use of these and other message shortening devices. Lastly, we are interested in predicting crime at a city-block resolution or finer, and it is not clear how tweets should be aggregated to support such analyses (prior work has investigated broader resolutions, for example, at the city or country levels). These factors conspire to produce a data source that is not only attractive – owing to its real time, personalized content – but also difficult to process. Thus, despite recent advances in all stages of the automatic text processing pipeline (e.g., word boundary identification through semantic analysis) as well as advances in crime prediction techniques (e.g., hot-spot mapping), the answer to our primary research question has remained unclear.

We pursued three objectives: (1) quantify the crime prediction gains achieved by adding Twitter-derived information to a standard crime prediction approach based on kernel density estimation (KDE), (2) identify existing text processing tools and associated parameterizations that can be employed effectively in the analysis of tweets for the purpose of crime prediction, and (3) identify performance bottlenecks that most affect the Twitter-based crime prediction approach. Our

* Tel.: +1 434 924 5397; fax: +1 434 982 2972.
E-mail address: msg8u@virginia.edu.

http://dx.doi.org/10.1016/j.dss.2014.02.003
0167-9236/© 2014 Elsevier B.V. All rights reserved.

results indicate progress toward each objective. We have achieved crime prediction performance gains across 19 of the 25 different crime types in our study using a novel application of statistical language processing and spatial modeling. In doing so, we have identified a small number of major performance bottlenecks, solutions to which would benefit future work in this area.

The rest of this article is structured as follows: in Section 2, we survey recent work on using Twitter data for predictive analytics. In Section 3, we describe our datasets and how we obtained them. In Section 4, we present our analytic approach for Twitter-based crime prediction, which we evaluate in Section 5. In Section 6, we discuss our results and the runtime characteristics of our approach. We conclude, in Section 7, with a summary of our contributions and pointers toward future work in this area.

2. Related work

2.1. Crime prediction

Hot-spot maps are a traditional method of analyzing and visualizing the distribution of crimes across space and time [7]. Relevant techniques include kernel density estimation (KDE), which fits a two-dimensional spatial probability density function to a historical crime record. This approach allows the analyst to rapidly visualize areas with historically high crime concentrations. Future crimes often occur in the vicinity of past crimes, making hot-spot maps a valuable crime prediction tool. More advanced techniques like self-exciting point process models also capture the spatiotemporal clustering of criminal events [8]. These techniques are useful but carry specific limitations. First, they are locally descriptive, meaning that a hot-spot model for one geographic area cannot be used to characterize a different geographic area. Second, they require historical crime data for the area of interest, meaning they cannot be constructed for areas that lack such data. Third, they do not consider the rich social media landscape of an area when analyzing crime patterns.

Researchers have addressed the first two limitations of hot-spot maps by projecting the criminal point process into a feature space that describes each point in terms of its proximity to, for example, local roadways and police headquarters [9]. This space is then modeled using simple techniques such as generalized additive models or logistic regression. The benefits of this approach are clear: it can simultaneously consider a wide variety of historical and spatial variables when making predictions; furthermore, predictions can be made for geographic areas that lack historical crime records, so long as the areas are associated with the requisite spatial information (e.g., locations of roadways and police headquarters). The third limitation of traditional hot-spot maps – the lack of consideration for social media – has been partially addressed by models discussed in the following section.

2.2. Prediction via social media

In a forthcoming survey of social-media-based predictive modeling, Kalampokis et al. identify seven application areas represented by 52 published articles [10]. As shown, researchers have attempted to use social media to predict or detect disease outbreaks [11], election results [12], macroeconomic processes (including crime) [13], box office performance of movies [14], natural phenomena such as earthquakes [15], product sales [16], and financial markets [17]. A primary difference between nearly all of these studies and the present research concerns spatial resolution. Whereas processes like disease outbreaks and election results can be addressed at a spatial resolution that covers an entire city with a single prediction, criminal processes can vary dramatically between individual city blocks. The work by Wang et al. comes closest to the present research by using tweets drawn from local news agencies [13]. The authors found preliminary evidence that such tweets can be used to predict hit-and-run vehicular accidents and breaking-and-entering crimes; however, their study did not address several key aspects of social-media-based crime prediction. First, they used tweets solely from hand-selected news agencies. These tweets, being written by professional journalists, were relatively easy to process using current text analysis techniques; however, this was done at the expense of ignoring hundreds of thousands of potentially important messages. Second, the tweets used by Wang et al. were not associated with GPS location information, which is often attached to Twitter messages and indicates the user's location when posting the message. Thus, the authors were unable to explore deeper issues concerning the geographic origin of Twitter messages and the correlation between message origin and criminal processes. Third, the authors only investigated two of the many crime types tracked by police organizations, and they did not compare their models with traditional hot-spot maps.

The present research addresses all limitations discussed above. We combine historical crime records with Twitter data harvested from all available Twitter users in the geographic area of interest. We address some of the difficult textual issues described previously (e.g., symbols and nonstandard vocabulary) using statistical language processing techniques, and we take full advantage of GPS location information embedded in many tweets. Furthermore, we demonstrate the performance of our approach on a comprehensive set of 25 crime types, and we compare our results with those obtained using standard hot-spot mapping techniques.

3. Data collection

Chicago, Illinois ranks third in the United States in population (2.7 million), second in the categories of total murders, robberies, aggravated assaults, property crimes, and burglaries, and first in total motor vehicle thefts (January–June, 2012 [18]). In addition to its large population and high crime rates, Chicago maintains a rich data portal containing, among other things, a complete listing of crimes documented by the Chicago Police Department.[1] Using the Data Portal, we collected information on all crimes documented between January 1, 2013 and March 31, 2013 ($n = 60{,}876$). Each crime record in our subset contained a timestamp of occurrence, latitude/longitude coordinates of the crime at the city-block level, and one of 27 types (e.g., ASSAULT and THEFT). Table 1 shows the frequency of each crime type in our subset.

During the same time period, we also collected tweets tagged with GPS coordinates falling within the city limits of Chicago, Illinois ($n = 1{,}528{,}184$). We did this using the official Twitter Streaming API, defining a collection bounding box with coordinates [−87.94011,41.64454] (lower-left corner) and [−87.52413,42.02303] (upper-right corner). Fig. 1 shows a time series of the tweets collected during this period and Fig. 2 shows a graphical KDE of the tweets within the city limits of Chicago. As shown in Fig. 2, most GPS-tagged tweets are posted in the downtown area of Chicago.

4. Analytic approach

To predict the occurrence of crime type *T*, we first defined a one-month training window (January 1, 2013–January 31, 2013). We then put down labeled points (latitude/longitude pairs) across the city limits of Chicago. These points came from two sources: (1) from the locations of known crimes of type *T* within the training window (these points received a label *T*), and (2) from a grid of evenly spaced points at 200-meter intervals, not coinciding with points from the first set (these points received a label *NONE*). Using all points, we trained a binary classifier with the following general form:

$$Pr\left(Label_p = T | f_1(p), f_2(p), \ldots, f_n(p)\right) = F(f_1(p), f_2(p), \ldots, f_n(p)). \quad (1)$$

[1] City of Chicago Data Portal: https://data.cityofchicago.org.