ARTICLE IN PRESS

Pervasive and Mobile Computing ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Pervasive and Mobile Computing

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

Fast track article

ClariSense+: An enhanced traffic anomaly explanation service using social network feeds

Prasanna Giridhar[a,*], Md Tanvir Amin[a], Tarek Abdelzaher[a], Dong Wang[b], Lance Kaplan[c], Jemin George[c], Raghu Ganti[d]

[a] Department of Computer Science, University of Illinois at Urbana–Champaign, Urbana, IL 61801, United States

[b] Department of Computer Science, University of Notre Dame, Notre Dame, IN 46556, United States

[c] Networked Sensing & Fusion Branch, US Army Research Laboratory, Adelphi, MD 20783, United States

[d] IBM Research, Yorktown Heights, NY, United States

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Anomaly detection
Sensors
Social data
Information gain
Credibility

ABSTRACT

The explosive growth in social networks that publish real-time content begs the question of whether their feeds can complement traditional sensors to achieve augmented sensing capabilities. One such capability is to *explain* anomalous sensor readings. In our previous conference paper, we built an automated anomaly clarification service, called ClariSense, with the ability to explain sensor anomalies using social network feeds (from Twitter). In this extended work, we present an enhanced anomaly explanation system that augments our base algorithm by considering both (i) the credibility of social feeds and (ii) the spatial locality of detected anomalies. The work is geared specifically for describing small-footprint anomalies, such as vehicular traffic accidents. The original system used information gain to select more informative microblog items to explain physical sensor anomalies. In this paper, we show that significant improvements are achieved in our ability to explain small-footprint anomalies by accounting for information credibility and further discriminating among high-information-gain items according to the size of their spatial footprint. Hence, items that lack sufficient corroboration and items whose spatial footprint in the blogosphere is not specific to the approximate location of the physical anomaly receive less consideration. We briefly demonstrate the workings of such a system by considering a variety of real-world anomalous events, and comparing their causes, as identified by ClariSense+, to ground truth for validation. A more systematic evaluation of this work is done using vehicular traffic anomalies. Specifically, we consider real-time traffic flow feeds shared by the California traffic system. When flow anomalies are detected, our system automatically diagnoses their root cause by correlating the anomaly with feeds on Twitter. For evaluation purposes, the identified cause is then retroactively compared to official traffic and incident reports that we take as ground truth. Results show a great correspondence between our automatically selected explanations and ground-truth data.

* Corresponding author.
E-mail addresses: giridha2@illinois.edu (P. Giridhar), maamin2@illinois.edu (M.T. Amin), zaher@illinois.edu (T. Abdelzaher), dwang5@nd.edu (D. Wang), lance.m.kaplan@us.army.mil (L. Kaplan), jemin.george.civ@mail.mil (J. George), rganti@us.ibm.com (R. Ganti).

http://dx.doi.org/10.1016/j.pmcj.2017.02.007

Please cite this article in press as: P. Giridhar, et al., ClariSense+: An enhanced traffic anomaly explanation service using social network feeds, Pervasive and Mobile Computing (2017), http://dx.doi.org/10.1016/j.pmcj.2017.02.007

1. Introduction

The proliferation of sensors used in *human spaces*, such as smart power meters, pollution meters, GPS devices, and vehicular traffic flow sensors, suggests that phenomena measured by such sensors will often be observed and reported socially as well. This is especially true of phenomena that deviate from the norm, hence attracting human attention. For example, the use of a mall's parking lot by freight trucks that increase local pollution, the closure of a freeway due to a forest fire, or the change in building occupancy patterns due to shutdown of a major local employer are events that leave a signature on both local sensors and social media. This leads to the idea of developing a service that explains anomalies seen by sensors using data feeds from social networks (e.g., Twitter [1]).

In this paper, we develop such an anomaly explanation service, with a focus on explaining *small-footprint* anomalies, defined as those that are localized in time and in space. We specifically apply this service to explain vehicular traffic flow abnormalities. The general working principle of such a service is conceptually simple. Given a sensor network, such as the network of traffic flow meters on city highways, and a social network, such as Twitter, the service (i) detects anomalies in sensor reports, (ii) detects anomalies in rates of different keywords in microblog items posted on the social network, (iii) filters out anomalous microblog items that lack sufficient corroboration, (iv) filters out anomalous microblog items, whose spatial signature in the blogosphere significantly exceeds that of the physical anomaly, and then finally (v) sorts the remaining set of microblog items by information gain and matches them to the set of sensors where anomalous readings are detected. A match occurs between a microblog item and a sensor if the microblog item contains keywords that refer to geographic landmarks relevant to the location of the sensor. Our results based on the traffic study show that the first one or two matching items, when considered in decreasing order of information gain, explain the sensor anomaly in most cases. Hence, a client of the service can see a map of anomaly locations (from the sensor network) and their automatically identified explanations (from the social network) that aim to clarify corresponding potential root causes. This paper describes the design and implementation of this service.

We first showcase the workings of this service for a few major real world events that caused deviation in sensor data readings. Twitter feeds corresponding to these event domains were collected and analyzed for possible explanations giving a proof that such a service is capable of identifying physical sensor anomalies with causes in any kind of settings. We then describe a case study of using this service to explain anomalies in vehicular traffic flows. Specifically, we consider real-time data feeds from traffic flow sensors in three major cities in California: Los Angeles, San Francisco, and San Diego. A public database containing this information is available at: http://pems.dot.ca.gov. The same database contains a significant number of traffic events (or *incidents*), such as vehicular accidents, reported road hazards, and scheduled road maintenance events, together with corresponding sensor locations and event durations.

Real-time traffic data were collected for 3 weeks at a period of one sample (per sensor) every 5 min. Simultaneously, Twitter feeds were collected and ranked based on inclusion of "anomalous" keywords (i.e., those that occurred with disproportionately different frequencies before and after the sensor anomaly). Anomalous sensor readings were then matched to anomalous tweets, as described in the paper, offering clarifications of possible causes of the anomaly. A comparison of these explanations to manually collected ground truth from the incidents database suggests that our system is very good at explaining truly unusual events. The more unusual the event, the better the chances that our service automatically explains it.

It is important at this point to make a disclaimer that recognizes a limitation of this work. Specifically, it should be noted that there is an inherent conflict between demonstrating the value of our service and being able to evaluate its accuracy in a systematic manner. To evaluate accuracy systematically on a sufficiently large dataset, we need to consider a case study that features anomalies for which ground truth explanations are readily available. However, the availability of ground truth explanations in such a case study is precisely what makes this case study a poor demonstration of value of the service. Said differently, since we can access police reports to understand why a traffic anomaly happened, why do we need ClariSense+? In this paper, we opted for the systematic evaluation. Consequently, while our contribution claims lie in developing an analytical framework and demonstrating its ability to explain traffic anomalies, the purpose of publishing it is to encourage subsequent development of other applications in domains where ground truth is not as easy to come by. Various military and disaster response scenarios come to mind, but of course until the system is tested in these scenarios as well, no general applicability claims can be made beyond the current evaluation domain.

With that disclaimer in mind, the remainder of this paper is organized as follows. We describe a brief feasibility argument in Section 2. We present the system design in Section 3. The evaluation of our system is discussed in Section 4, followed by related work in Section 5. In Section 6 we discuss about the things we learned from this work. Finally, conclusions are presented in Section 7.

2. A feasibility argument

On June 9th, 2013, an anomalous 10-mile traffic jam was detected on a major Southern California freeway. To explain it, we contrasted two Twitter feeds; Namely, (i) Twitter feeds with keywords "California", "Traffic", and "Jam" shortly after the anomaly and (ii) Twitter feeds with the same keywords from one week earlier. The comparison returned the words "Impeach" and "Obama" as discriminative (in that they occur very frequently in the current feed while being absent from the earlier one). Picking the most common tweets containing these keywords, we obtained the following actual tweets:

Please cite this article in press as: P. Giridhar, et al., ClariSense+: An enhanced traffic anomaly explanation service using social network feeds, Pervasive and Mobile Computing (2017), http://dx.doi.org/10.1016/j.pmcj.2017.02.007