Knowledge-Based Systems 000 (2018) 1–11

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Who to select: Identifying critical sources in social sensing

Dong Wang*, Nathan Vance, Chao Huang

Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, United States

ARTICLE INFO

Article history:
Received 10 May 2017
Revised 1 January 2018
Accepted 3 January 2018
Available online xxx

Keywords:
Source selection
Source dependency
Speak rate
Social sensing
Twitter

ABSTRACT

Social sensing has emerged as a new data collection paradigm in networked sensing applications where humans are used as "sensors" to report their observations about the physical world. While many previous studies in social sensing focus on the problem of ascertaining the reliability of data sources and the correctness of their reported claims (often known as *truth discovery*), this paper investigates a new problem of *critical source selection*. The goal of this problem is to identify a subset of critical sources that can help effectively reduce the computational complexity of the original truth discovery problem and improve the accuracy of the analysis results. In this paper, we propose a new scheme, *Critical Source Selection (CSS)*, to find the critical set of sources by explicitly exploring both *dependency* and *speak rate* of sources. We evaluated the performance of our scheme and compared it to the state-of-the-art baselines using two data traces collected from a real world social sensing application. The results showed that our scheme significantly outperforms the baselines by finding more truthful information at a higher speed.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

This paper develops a new scheme to solve the critical source selection problem in social sensing applications. Social sensing has emerged as a new networked sensing paradigm of collecting observations about the physical environment from humans or devices on their behalf. This paradigm is motivated by the proliferation of digital sensors in the possession of individuals (e.g., smartphones) and the wide adaptation of online social media (e.g., Twitter, Facebook). In social sensing applications, people can report certain observations about their environment such as traffic conditions at various locales [1], pothole information on streets [2], and available gas stations in the aftermath of a disaster [3]. One key challenge of using "humans as sensors" is to estimate the correctness of observations (i.e., *claims*) and the reliability of data sources without knowing ground truth about the situation *a priori*. We refer to this problem as the *truth discovery problem*.

In this paper, we study a new problem of *critical source selection* where the goal is to identify a subset of critical sources that can reduce the computational complexity of the original truth discovery problem and improve the accuracy of the analysis results. First, it is critical to consider the source dependency in solving this problem. In social sensing, it is not unusual for a human source to forward claims they received from others (e.g., friends from their social networks) [4]. Fig. 1 shows some simple examples extracted from real-world Twitter data where sources with social connections (i.e., following relationship) report the same claim. From a networked sensing perspective, such dependencies between sources can easily introduce correlation and redundancy between reported observations, which are shown to affect truth discovery results negatively if they are not appropriately modeled [5]. Previous works [5–8] have started to account for dependencies between sources in truth discovery tasks by partitioning them into independent groups where sources in different groups are considered to be independent. However, the complexity of their solutions grow exponentially with respect to the maximum size of the independent groups, making them impractical in many large-scale social sensing applications [6]. In this paper, we develop a new source selection scheme to explicitly consider the source dependency in the source selection process.

In addition to the source dependency, the speak rate of a source (i.e., how chatty a source is) is another important factor to consider in the critical source selection solution. In social sensing, different sources often report different numbers of claims. The speak rate of a source has a strong positive correlation with both the accuracy and the granularity of the source reliability estimation, which also directly affects the estimation of the claim correctness [9]. Therefore, the goal of our critical sensor selection scheme is to (i) maximize the average speak rate of the selected sources and (ii) minimize the dependency between them. However, those two objectives can be at odds with each other, which makes the critical sensor selection problem non-trivial to solve.

* Corresponding author.
E-mail address: dwang5@nd.edu (D. Wang).

https://doi.org/10.1016/j.knosys.2018.01.006
0950-7051/© 2018 Elsevier B.V. All rights reserved.

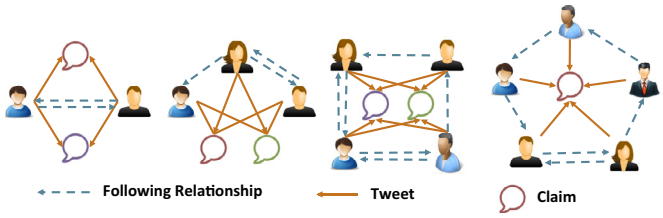

Fig. 1. Source dependency examples on Twitter.

Previous work has made significant progress towards source selection in sensor network and data fusion [10–15]. However, most current solutions ignore either the source dependency or the speak rate in their models, which has led to suboptimal source selection results because redundant sources or sources with inaccurate source reliability estimations are selected. In this paper, we present a Critical Source Selection (CSS) scheme that explicitly incorporates both the *source dependency* and the *speak rate* feature into the critical source selection process. In particular, we formulate our critical source selection problem as a constraint optimization problem with multiple objectives, and we develop an efficient algorithm to solve it. We evaluate our CSS scheme in comparison with the state-of-the-art baselines using two real-world social sensing data traces collected from Twitter (i.e., the Paris Attack event in 2015 and the Oregon Shooting event in 2015). The results show that our scheme significantly outperforms the baselines by finding more truthful information at a faster speed.

In summary, our contributions are as follows:

- We investigate the problem of critical source selection in social sensing to reduce the complexity of the truth discovery problem while simultaneously improving the accuracy of estimation results.
- We develop a new approach (CSS) that selects a critical set of sources by exploring both their source dependencies and their speak rates.
- We perform extensive experiments to compare the performance of our CSS scheme with state-of-the-art baselines using real-world social sensing data. The evaluation results demonstrate the effectiveness and efficiency of our scheme.

A preliminary version of this work has been published in [16]. This work significantly expands on our previous work and makes new contributions as follows. *First*, we extend our previous proposed model by developing a new annealing based process to increase the probability of reaching a globally optimal solution (Section 4). *Second*, we formally prove that the critical source selection problem in our work is NP-hard (Section 4). *Third*, we compare our scheme with more recent baselines using the real-world datasets and carry out a more comprehensive evaluation and comparison between the CSS scheme and the state-of-the-art techniques (Section 5). *Fourth*, we perform a set of new experiments to investigate the effect of parameters in the proposed model and study the robustness of the model with respect to the changes of the parameters (Section 5). *Finally*, we also evaluate the execution time of all compared algorithms to study their computational efficiency (Section 5).

The rest of this paper is organized as follows: we discuss the related work in Section 2. In Section 3, we present the problem of critical source selection. The proposed critical source selection scheme is discussed in Section 4. Experiment and evaluation are presented in Section 5. Finally, we conclude the paper in Section 7.

2. Related work

Social Sensing. has emerged as a new sensing paradigm which attracted much attention in sensor networks research [17], urban sensing [18], surrogate sensing [19], Internet of Things [20], and data distillation [21]. The key idea of social sensing is to use humans as sensors in many sensing applications such as participatory sensing [22] and opportunistic sensing [23]. In particular, human sensors can contribute their observations through "sensing campaigns" [24] or social data scavenging [25]. Current works in social sensing have addressed important challenges in many relevant fields such as privacy perseverance [26], truth estimation [27], social signal processing [28], social sensor profiling [29], semantics of the sensing content [30,31], and social interaction promotions [32]. However, source selection remains a critical and open research question in social sensing. In this work, we study the problem of *critical source selection* to reduce the computational complexity of the truth discovery problem and improve the accuracy of the analysis results.

Truth discovery in social sensing. Data quality and trustworthiness is a fundamental challenge in social sensing. Prior works in social sensing have made significant advances to infer the credibility of reported data [6,7,33,34]. For example, Ouyang et al. [33] investigated the potential of leveraging crowds as sensors to detect the true value of quantitative characteristics from noisy social sensing data. Huang et al. explored the *topic relevance of claims and arbitrary source dependency problem* in social sensing and developed a topic-aware truth discovery solution [6]. Zhang et al. developed a reliable truth discovery solution that is robust to sparse data and misinformation in social sensing [35]. Zhao et al. studied the problem of real-time truth discovery and developed a probabilistic model to efficiently handle streaming data [34]. Wang at al. considered source dependency by assuming that it can be represented by sets of disjoint trees [7]. All the above works solve the *truth discovery problem* and focus on modeling the relationship between source reliability and claim correctness. In contrast, this paper solves a new problem of *critical source selection* which can help improve both the effectiveness and the efficiency of the above truth discovery solutions.

In addition, a few recent truth discovery solutions focus on improving efficiency by using streaming approaches [36–38]. For example, Wang et al. developed a streaming truth discovery scheme to recursively update the estimation results by leveraging the previous estimation and the CRLBs of the estimation [36]. Zhang et al. proposed another category of streaming truth discovery approaches by explicitly addressing the scalability and physical constraints in social sensing application [37,38]. However, the above works did not consider the critical source selection problem in their truth discovery solutions. In sharp contrast to those works, this paper improves the efficiency of the truth discovery solutions by solving a new critical source selection problem.

Source selection in social sensing. There exists a good amount of work on the topic of *source selection* in networked sensing, data mining, and machine learning communities [10–13,39]. For example, Uddin et al. investigated the problem of diversifying the source selection in social sensing based on the social connections between sources. Rekatsinas et al. [11] studied the problem of source selection for dynamic sources whose contents change over time. Dong et al. [12] proposed an algorithm to select a subset of sources in data fusion applications by considering integration cost. Hosseini et al. selected the subset of data sources to predict the state of all other sources by considering source correlations [13]. Amintoosi et al. [39] proposed a privacy-aware participant selection framework that explicitly protects users' privacy in the social sensing applications. However, most current solutions ignore either the source dependency or the speak rate in their models. In contrast,

Please cite this article as: D. Wang et al., Who to select: Identifying critical sources in social sensing, Knowledge-Based Systems (2018), https://doi.org/10.1016/j.knosys.2018.01.006

Download English Version:

https://daneshyari.com/en/article/6861647

Download Persian Version:

https://daneshyari.com/article/6861647

Daneshyari.com