Proc. Information Processing in Sensor Networks, Chicago, IL, April 12-14, 2011, pp.129-130

Apollo: Towards Factfinding in Participatory Sensing

H. Khac Le, J. Pasternack, H. Ahmadi, M. Gupta, Y. Sun, T. Abdelzaher, J. Han, D. Roth
University of Illinois at Urbana Champaign
Urbana, IL 61801

B. Szymanski, S. Adali
Rensselaer Polytechnic Institute
Troy, NY 12180

ABSTRACT

This demonstration presents Apollo, a new sensor information processing tool for uncovering likely facts in noisy participatory sensing data[1]. Participatory sensing, where users proactively document and share their observations, has received significant attention in recent years as a paradigm for crowd-sourcing observation tasks. However, it poses interesting challenges in assessing confidence in the information received. By borrowing clustering and ranking tools from data mining literature, we show how to group data into sets (or *claims*), corroborating specific events or observations, then iteratively assess both claim and source credibility, ultimately leading to a ranking of described claims by their likelihoold of occurrence. Apollo belongs to a category of tools called *fact-finders*. It is the first fact-finder designed and implemented specifically for participatory sensing. Apollo uses Twitter as the underlying engine for sharing participatory sensing data. Twitter is widely popular, can be interfaced to cell-phones that share sensor data, and comes with a powerful search API, as well as a publish-subscribe mechanism. We evaluate it using a participatory sensing application that collects and posts noisy vehicular traffic data on Twitter, as well as a set of 60,000 (human) tweets collected during the Haiti tsunami and a set of 500,000 tweets collected about Cairo during its recent unrest. Viewers of the demonstration will interact with Apollo for various fact-finding tasks.

Categories and Subject Descriptors

D.2.5 [**Software Engineering**]:

General Terms

Design, Reliability, Experimentation

Keywords

Data fusion, Participatory sensing, Quality of information

IPSN 2011 Chicago, IL USA

1. INTRODUCTION

This demonstration illustrates a fact-finding tool designed to uncover most likely truth in noisy participatory sensing data. Participatory sensing [1], where participants proactively report their observations, has become an increasingly important data collection paradigm, where humans act as the sensors, or employ devices they own (such as cell phones) to perform sensing tasks. Its growing importance stems from the large penetration of sensor-enabled devices with communication capabilities in human populations. However, it poses challenges in that data are collected from individuals and devices who may not always be accurate. Filtering the deluge of data for correctness is an important challenge, commonly known in machine learning and knowledge discovery literature as *fact-finding*. Apollo is the first fact-finder designed specifically for participatory sensing data.

A fact-finder [6] maintains the abstraction of *sources* and *claims*. In general, starting with *no a priori information*, it iteratively computes their credibility: The credibility of claims depends on the credibility of sources that make them. Similarly, the credibility of sources depends on the credibility of claims they make. Iterations continue until they converge. Enhancements of this basic iterative scheme include a more general notion of claim assertion, where weights describe how confidently a source asserts a claim [5], incorporation of prior knowledge in the analysis [4], and clustering of claims by *subject* [3] since the credibility of a source may depend on the subject matter. The impact of dependence between sources on credibility (e.g., when one spreads claims made by others) can also be accounted for [2].

Apollo is a general framework for fact-finding in participatory sensing data. Data-type-dependent modules first convert source data (a set of *source, observation* tuples) into a common representation of sources and claims. Clustering is performed on input tuples first, by similarity of their observations, to generate a smaller number of claims for scalability. Their credibility is then assessed in an iterative fashion, together with the credibility of each source. Figure 1 illustrates the architecture of Apollo.

We evaluate our fact-finding algorithm in three distinct scenarios. In the first, a controlled experiment is conducted where a set of participants report (tweet) traffic speed data from GPS sensors in vehicles. We intentially corrupt a fraction of observations. We also allow some participants to report traffic information they heard from others, as op-

[1] Named after the Greek god of light and truth, among other designations

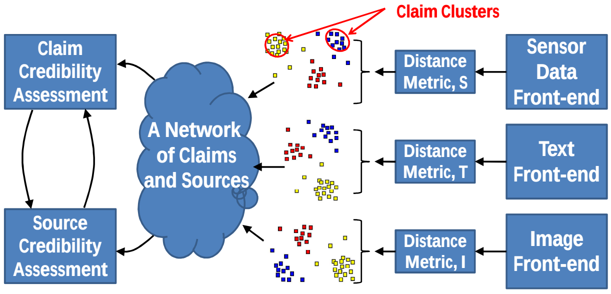

Figure 1: The Architecture of Apollo

posed to reporting their own (i.e., spread rumors). We show that computing average statistics from such noisy data is not always accurate. However, when data tuples are clustered and ranked by Apollo, the quality of reported observations increases considerably. The second and third scenarios use Twitter data sets collected during the Haiti Tsunami (60,000 reports) and the recent Cairo unrest (500,000 reports), respectively. In each case, the significant number of reports describe a much smaller number of events, some real and some rumored. We use our algorithm to identify the distinct events and rank them by credibility. The results are then compared to media reports.

2. SYSTEM OVERVIEW

We consider a system composed of human or sensory data sources that send their observations as Twitter feeds. The sink has two modes of operation. A *follower* mode, where it follows only explicitly named sources, and a *crawler* mode, where it uses the Twitter search API to download all tweets on the topic of a query (subject to API-specific volume constraints). A fact-finding algorithm is then applied to clean up the data. The algorithm has the following reconfigurable parts:

- *The parser:* It uses a configuration file (that describes the format of the input data stream) to convert input data into a standard JSON format. Conceptually, the data stream is composed of *source, observation* tuples, where *source* identifies the data source (e.g., the Twitter user ID), and the observation content may be structured, as in the case of phones sharing sensor values, or unstructured, as in the case of people sharing text. The configuration file describes the observation format.

- *The distance metrics:* A library of different distance metrics is provided for clustering of observations. For unstructured text, these metrics reflect text similarity. For structured data, metrics compute differences in data vectors. Data can be multidimensional. For example, when cell-phones report sensor values at given locations, both measurements and locations can be elements of the data vector.

- *The cluster credibility metrics:* When observations are clustered by the distance metric, one needs to rank different clusters in terms of credibility. A library of ranking functions are provided, inspired by current fact-finding literature, ranging from simple ones (how many tuples are in the cluster) to more complex ones (that take into account the social network relation between sources). For example, the rank of a cluster may be lower if observations in the cluster are reported by a group of tweeters who are followers of the same source.

- *Source credibility metrics:* More credible sources are those whose observations are found more often in more credible clusters. This basic intuition leads to a library of simple credibility metrics to rank sources.

With the above parameters, the rest of the algorithm simply iterates on computing clusters, cluster credibility metrics, source credibility metrics, until stable results are obtained.

3. THE DEMONSTRATION SCRIPT

During the demo, a laptop will be used that runs Apollo on three different sets of Twitter data; a set obtained in a controlled vehicular traffic speed measurement experiment, a set collected from Haiti during its tsunami, and a set collected from Cairo during its recent unrest. Users will be allowed to reconfigure the experiments by changing the used distance and credibility metrics, clustering algorithm parameters, as well as the amount of data operated on, and observing their effects on final results. This will help answer questions such as: what is the impact (in the given scenario) of considering social network relations between data sources on improving quality of information in participatory sensing systems? What are the trade-offs between different distance metrics among data vectors? How coarse-grained can observation clustering be without impacting correctness of results? How much data is needed to get reliable results? Results will be displayed as lists of credible events and credible sources. "Ground truth" data will also be shown (which is known for our controlled experiment and estimated from other media in the other two case studies).

4. REFERENCES

[1] J. Burke, D. Estrin, M. Hansen, A. Parker, N. Ramanathan, S. Reddy, and M. B. Srivastava. Participatory sensing. In *In: Workshop on World-Sensor-Web (WSW): Mobile Device Centric Sensor Networks and Applications*, pages 117–134, 2006.

[2] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. *Proc. VLDB Endow.*, 2:550–561, August 2009.

[3] M. Gupta, Y. Sun, and J. Han. Trust analysis with clustering (poster paper). In *World Wide Web Conference (WWW)*, Hyderabad, India, March 2011.

[4] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *Proc. the International Conference on Computational Linguistics (COLING)*, Beijing, China, August 2010.

[5] J. Pasternack and D. Roth. Generalized fact-finding (poster paper). In *World Wide Web Conference (WWW)*, Hyderabad, India, March 2011.

[6] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. on Knowl. and Data Eng.*, 20:796–808, June 2008.