Proc. 9th ACM Conference on Embedded Networked Sensor Systems, Seattle, WA, November 1-4, 2011.

Demo: Distilling Likely Truth from Noisy Streaming Data with Apollo

Hieu Le[1] hieule2@illinois.edu
Dong Wang[1] dwang24@illinois.edu
Hossein Ahmadi[1] hahmadi2@illinois.edu
Md Yusuf S Uddin [1] mduddin2@illinois.edu
Boleslaw Szymanski[2] szymab@rpi.edu
Raghu Ganti[3] rganti@us.ibm.com
Tarek Abdelzaher[1] zaher@illinois.edu

[1]University of Illinois at Urbana Champaign, Urbana, IL 61801
[2]Rensselaer Polytechnic Institute, Troy, NY 12180
[3]IBM Research, Yorktown Heights, NY 10598

Abstract

At CPSWeek 2011, the authors presented a demonstration of Apollo, a fact-finder for participatory sensing that ranks archived human-centric and sensor data by credibility. The current demonstration significantly extends our previous work by allowing Apollo to operate on live *streaming data*; in this case, *live* Twitter feeds. As the role of humans as sensors increases in emerging sensing applications, a principled approach becomes necessary to address the problem of ascertaining the veracity of sources and observations made by them. Participatory and social sensing applications may use potentially unreliable or unverified sources, such as a phone-based sensing application that grows virally in a large un-vetted population, a disaster-response application, where conflicting damage assessment reports may come from large numbers of different volunteers, or a military application, where friendly observers at a remote location may make hard-to-verify claims about local events. Apollo analyzes noisy data that increasingly plagues human-centric sensing to determine which items of information are more likely to be true.

Categories and Subject Descriptors

C.3 [**Special-purpose and Application-based Systems**]: Signal processing systems; J.7 [**Computers in Other Systems**]: Real time

General Terms

Measurement

Keywords

Data fusion, Participatory sensing, Quality of information

SenSys'11, November 1–4, 2011, Seattle, WA, USA.
ACM 978-1-4503-0718-5/11/11

1 Introduction

This demonstration illustrates an extended version of the general fact-finding tool, Apollo [2]. Apollo was designed to uncover most likely truth in noisy participatory sensing data.

As mobile and other information sharing services become commodity, human-centric participatory sensing applications become an attractive data collection paradigm. However, along with the increased availability of data comes the problem of ensuring quality of information. Services that rely on massive collection of human-centric data have little control over the reliability of participants in the data collection process. The ability to rank reported data and sources by credibility becomes an important aspect of systems that need to derive value from such data. This problem is commonly known as *fact-finding*.

While a complete description of the theory behind Apollo is beyond the scope of this abstract, informally, Apollo breaks down input data streams into *entities* and *claims*. An entity represents an information source such as a person or a sensor. A claim is an information item reported by some entity, such as a a tweet , a picture, or a sensor value. Apollo creates a logical network where nodes represent entities and claims. Network links connect the entities to the claims they make. Apollo also links mutually corroborating claims together, such as several messages with the same content, or several sensors reporting similar observations around the same location. The links in the network constructed by Apollo can be thought of as constraints on relations between possible truth values taken by the nodes linked. For example, a link between two claims may indicate that they are likely to be either both true or both false. A link between a source entity and a claim means that the probability of correctness (or accuracy) of the source is the same as the probability of correctness of the claim. This leads to the formulation of an expectation maximization problem where the goal is to find a maximum likelihood assignment of truth probabilities to claims and sources, given the source-claim network. This problem can be solved by an Expectation Maximization

(EM) algorithm. The solution yields, for each reported item of information, the probability that it is true, and for each entity, the probability with which it reports true data.

While fact-finding itself is not our contribution [1, 3], the formulation of the problem as one of expectation maximization (EM) is new. Our preliminary work, leading to this Apollo demonstration, has shown that the EM formulation significantly outperforms prior fact-finding approaches.

We demonstrate Apollo using real-time Twitter feeds. The application allows users to create *fact-finding tasks* by defining a query that allows the application to collect tweets on a subject of interest. A query includes optional keywords that tweets should have and a geographic region from which to collect the tweets. We show the user two scenarios. In the first, we illustrate results of existing real-time fact-finding tasks with pre-created queries . Users can see top-ranked tweets deemed probable, and those deemed not, then compare them to ground truth. In the second, users are allowed to create their own new queries at demonstration time and observe fact-finder outputs on topics of their choice.

In the next sections, we describe the general architecture of Apollo that supports real-time data processing capabilities.

2 System Overview

The input to Apollo can, in general, be a stream of reports, expressed as claims and their sources. The reported data can be in different formats such as images, GPS locations, scalar sensor readings, or text. The raw data is processed by an appropriate parser to produce a canonical input format. The data in the stream is then grouped into different *time windows* based on arrival time. Data in each time window is then converted into a graph of nodes and links and processed by the EM algorithm as described earlier. The EM algorithm then outputs claims and sources, sorted by likelihood. The results from different windows are combined into an exponential moving average, to allow evolution of credibility values over time. In general, the system has the following configurable parts:

- *The parser:* The parser breaks input data down into tuples of sources and claims. The claims are fed to a distance function to determine their similarity. Links are established between similar claims as well as between claims and sources.
- *The distance metrics:* Apollo provides a library of different basic distance functions for different types of data such as images, text, and numeric sensor data. The distance function takes two reported claims and produces a real-number in a predefined range. The resulting graph is then input to a generic clustering algorithm.
- *Clustering and source credibility ranking:* Claims are grouped into appropriate clusters by distance, and the network of sources and claims (or, rather, claim clusters) is processed by EM.
- *Smoothing:* Outputs of EM (the credibility values) are smoothed over time using an exponential moving average.

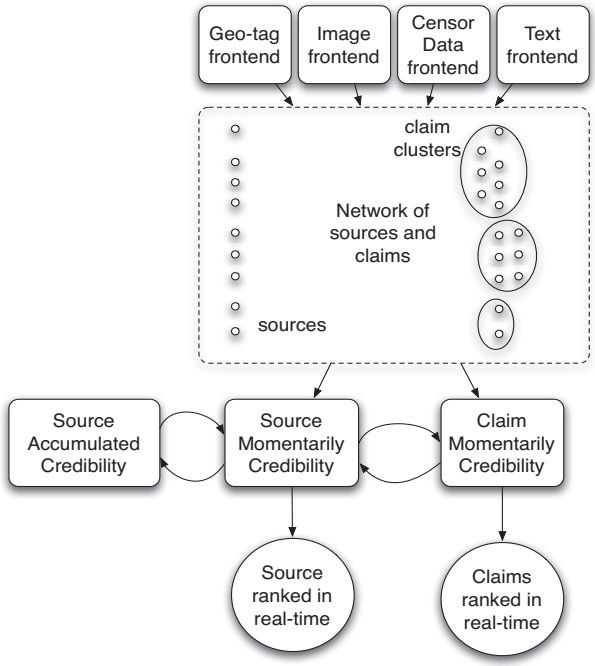

Figure 1. The architecture of Apollo.

3 The Demonstration Script

During the demonstration, a laptop will be used to connect to a remote server that runs the real-time fact-finding application. Attendees will be able to browse through existing results of previously created queries. These queries help users see what the results of a long-run query look like and compare to previously collected ground truth. Attendees will also be allowed to create new queries and observe the results in real-time. Users can pause, resume, and delete tasks. Users can also select different algorithm parameters such as distance functions.

4 Acknowledgments

Research reported in this paper was sponsored in part by NSF grant CNS 10-40380 and in part by the Army Research Laboratory, and was accomplished under Cooperative Agreement Number W911NF-09-2-0053.

5 Additional Authors

Omid Fatemieh[1], Hongyang Wang[1], Jeff Pasternack[1], Jiawei Han[1], Dan Roth[1], Sibel Adali[2], and Hui Lei[3].

6 References

[1] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *WSDM*, pages 131–140, 2010.

[2] H. K. Le, H. Ahmadi, D. Wang, O. Fatemieh, Y. Sarwar, M. Gupta, J. Pasternack, T. Abdelzaher, J. Han, D. Roth, B. Szymanski, S. Adali, R. Ganti, F. Ye, and H. Lei. Apollo: Towards factfinding in participatory sensing. In *IPSN (demo abstract)*, 2011.

[3] J. Pasternack and D. Roth. Knowing what to believe (when you already know something). In *COLING*, 2010.