Social and Sensor Data Fusion in the Cloud

Surender Reddy Yerva, Jonnahtan Saltarin, Hoyoung Jeung, Karl Aberer

École Polytechnique Fédérale de Lausanne (EPFL)

{surenderreddy.yerva, jonnahtan.saltarin, hoyoung.jeung, karl.aberer}@epfl.ch

***Abstract*—This paper explores the potential of fusing social and sensor data in the cloud, presenting a practice—a travel recommendation system that offers the predicted mood information of people on where and when users wish to travel. The system is built upon a conceptual framework that allows to blend the heterogeneous social and sensor data for integrated analysis, extracting weather-dependant people's mood information from Twitter and meteorological sensor data streams. In order to handle massively streaming data, the system employs various cloud-serving systems, such as Hadoop, HBase, and GSN. Using this scalable system, we performed heavy ETL as well as filtering jobs, resulting in 12 million tweets over four months. We then derived a rich set of interesting findings through the data fusion, proving that our approach is effective and scalable, which can serve as an important basis in fusing social and sensor data in the cloud.**

I. Introduction

Mobile phones increasingly become multi-sensor devices, accumulating large volumes of data related to our daily lives. At the same time, mobile phones are also serving as a major channel for recording people's activities at social-networking services in the Internet. These trends obviously raise the potential of collaboratively analyzing sensor and social data in mobile cloud computing—where applications running in the cloud are accessed from thin mobile clients, providing virtually unlimited processing power, and promising cross-device platform compatibility.

The two popular data types, social and sensor data, are in fact mutually compensatory in various data processing and analysis. Participatory sensing, for instance, enables to collect people-sensed data via social network services (e.g., Twitter) over the areas where physical sensors are unavailable. Simultaneously, sensor data is capable of offering precise context information, leading to effective analysis of social data. Obviously, the potential of blending social and sensor data is high; nevertheless, currently they are typically processed separately in mobile cloud applications, and the potential has not been investigated sufficiently.

Social-network services, like Twitter, facilitate users to share their ideas, opinions, pictures, videos, news, and other various forms of contents in the Web. Such social data typically contains highly valuable information, aiding a wide range of applications; for example, allowing social scientists to understand human behaviors, companies to figure out their customers' preferences, and news agencies to identify significant news, etc. The works in [1], [2], [3], [4], [5] demonstrate the successful extraction of useful information from micro-blogs, showing that these analysis can act as social sensors resulting in social measurements.

In this paper, we explore the possibility of fusing social and sensor data in the cloud, while dealing with massive data streams. To this end, we present a travel recommendation system as a practice of the fusion, which offers the information of people's moods regarding the predicted weather on where and when users wish to travel. The recommendation system gears various components towards effective, large-scale social and sensor data fusion.

II. The Travel Recommendation System

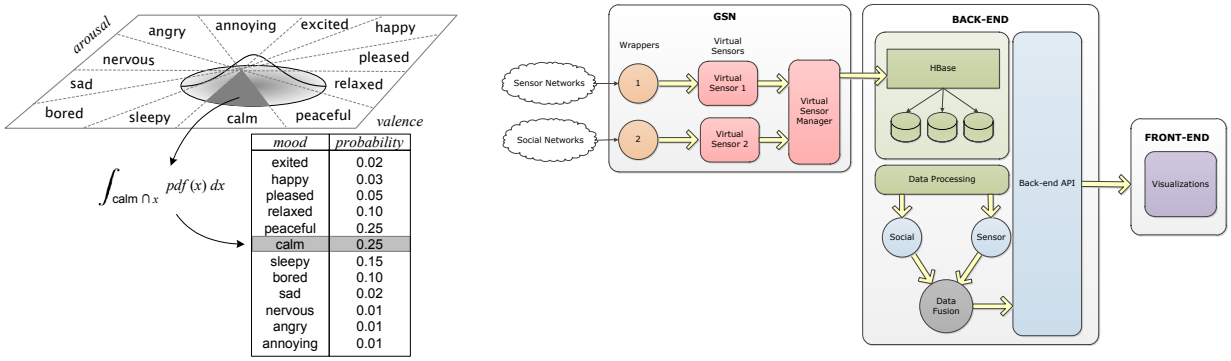
In our recommendation system, users provide their travel intentions (place and approximate date of travel), and then the system provides the information of how enjoyable the place would be on the day for travel, in addition to the typical information offered by ordinary travel recommender systems. This recommendation process is comprised of the following steps:

1) A user first offers the details for travel to the system, e.g., going to London next Friday.
2) The system obtains the information of predicted weather on London next Friday, from a real-time weather prediction service (e.g., WeatherUnderground).
3) The system looks up the mood information of people associated London and Friday, which is continuously mined and updated from raw social and sensor data.
4) The system offers the information of how enjoyable the trip to London on next Friday would be, according to the mood probability estimation.

In order to extract people's mood information from social (Twitter) feeds associated with sensor (weather) data, we propose a data space, called *mood space*, which serves as a conceptual base-ground where social and sensor data can be mapped.

Given a tweet message, our framework associates a mood label with the message, by computing the valence and arousal scores of the tweet, which is represented as a data point in the mood space, shown in Fig. 1(a). When a stream of tweets are represented as data points in mood space, the framework computes a set of mood probabilities, according to a day, weather label and location. This compact representation of mood probabilities, will be later used by the framework to suggest recommendations.

We consider weather information at a particular location as *sensor data* in our data fusion setting. Specifically, we



(a) An example of mood probability computation.

(b) Architecture of the fusion system.

Fig. 1. Travel Recommendation System: A practice of Social and Sensor Data Fusion in the Cloud

make use of services provided by WeatherUnderground[1], in order to periodically query the weather of a particular location. WeatherUnderground is a service that provides real-time weather information from nearly 32,000 weather stations around the world.

In order to store and process massively streaming social and sensor data in the cloud, we propose a system established upon a combination of state-of-the-art cloud systems, including Hadoop[6], HBase[7], and GSN (Global Sensor Networ) [8].

Fig. 1(b) shows an overview of the system. The Fusion framework consists of three primary components: (a) *GSN* is a stream processing engine that supports a flexible integration of data streams. In our travel recommendation system, GSN serves as a wrapper that receives streaming social as well as sensor data from twitter and weather data sources. Processed data is then stored in the back-end. (b) *Back-End* contains both Hadoop and HBase, serving as a storage-and-computing platform. Hadoop (MapReduce) is a popular framework for data-intensive distributed computing of batch jobs. In our system, Hadoop is used to parse continuously streaming tweets as well as weather data delivered in an XML format, based on a cluster that is built on 16 machines. The parsed data are then stored in HBase, which is commonly used as a "Hadoop storage". (c) *Front-End* implements a user interface of the recommendation system. Specifically, this component takes user inputs for querying, and delivers the inputs to the back-end. The query results returned from the back-end are then visualized through the front-end.

III. Experiments

We evaluated our data-fusion approach on comprehensive experimental results, obtained from using 12 million tweets as well as meteorological sensor readings collected over four months for London. We extracted the different social metrics (like mood information), and performed the fusion of social and sensor data using various configurable Hadoop(Map-Reduce) jobs.The results demonstrate various interesting findings, including the degree of variation in happiness according to a particular weather type, day, and location. We observed that people in general are more happier on the weekends $\{Fri, Sat, Sun\}$ compared to the weekdays $\{Mon, Tue, Wed, Thu\}$, and $Monday$ had the least fraction of happier tweets. For London tweets, we see that people are happiest on sunnier days, followed by cloudy days and least happy when it is raining. This trend was strong for weather-related tweets in comparison to all the tweets. Furthermore through our simple prediction models, we statistically prove that our mood estimation based on the fusion is effective and accurate. Relying on scalable cloud components, our fusion process can be readily expanded to many more locations with little effort.

[1] http://www.wunderground.com

IV. Conclusions

In this paper, we explored the potential of the data fusion by proposing a theoretical framework that enables to analyze tweet messages for extracting people's moods depending on day, weather, and location. We implemented the framework as a travel recommendation system that facilitates the fusion process over massively streaming data. The system is established upon several well-known cloud systems, allowing scalable data-fusion processing and being effective.

V. Acknowledgements

This work was partly funded by the Swiss Nano-Tera OpenSense project (Nano-Tera ref. 839 401) and the European Commission in the PlanetData NoE (contract nr. 257641).

References

[1] J. Bollen, H. Mao, and X.-J. Zeng, "Twitter mood predicts the stock market," *ArXiv e-prints*, Oct. 2010.
[2] "Pulse of the nation," http://www.ccs.neu.edu/home/amislove/twittermood/.
[3] S. R. Yerva, Z. Miklos, and K. Aberer, "Entity-based classification of twitter messages," *IJCSA*, vol. 9, no. 2, pp. 88–115, 2012.
[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," in *WWW '10*. New York, NY, USA: ACM, Apr. 2010, pp. 851–860.
[5] M. Nagarajan, A. Sheth, and S. Velmurugan, "Citizen sensor data mining, social media analytics and development centric web applications," in *WWW '11*. New York, NY, USA: ACM, 2011, pp. 289–290.
[6] T. White, *Hadoop: The Definitive Guide*, O. Media, Ed. O'Reilly Media, 2009.
[7] "Hbase," http://hbase.apache.org.
[8] K. Aberer, M. Hauswirth, and A. Salehi, "A middleware for fast and flexible sensor network deployment," in *VLDB*, 2006, pp. 1199–1202.