

Truth Discovery in Big Data Social Media Application

M. Nigade, M. Raut, P. Mane, S. Phadatare

Student

Sanjay Ghodawat Institute

Corresponding author's email id: mohini.raut14@gmail.com*

DOI: <http://doi.org/10.5281/zenodo.2653576>

Abstract

In this system first one is “misinformation spread” where a significant number of sources are contributing to false claims, making the identification of truthful claims difficult. For example, on, Instagram, rumors, Twitter scams, and influence bots are common examples of sources colluding, either intentionally or unintentionally, to spread misinformation and obscure the truth. The challenge is “data sparsity” or the “long-tail phenomenon” where a majority of sources only contribute a small number of claims, providing insufficient evidence to determine those sources’ trustworthiness. For example, in the Twitter datasets that we collected during real-world events, more than 90 only contributed to a single claim. Third, many current solutions are not scalable to large-scale social sensing events because of the centralized nature of their truth discovery algorithms. We are going to develop a Scalable and Robust Truth Discovery (SRTD) scheme to address the above all challenges. In this, the SRTD scheme jointly quantifies both the reliability of sources and the credibility of claims using a principled approach.

Keywords: Database Management, Query Processing, Scalable and Robust Truth Discovery (SRTD), Truth Discovery

INTRODUCTION

We are going to represent the Scalable Robust Truth Discovery (SRTD) scheme to solve the truth discovery problem in big

data social media sensing applications. We first outline a few observations relevant to our model. We then discuss the design intuition and present the SRTD scheme.

In this system presents a new scalable and robust approach to solve the truth discovery problem in big data social media sensing applications. Online social media (e.g., Twitter, Facebook, and Instagram) provides a new sensing paradigm in the big data era where people act as ubiquitous, inexpensive, and versatile sensors to spontaneously report their observations (often called claims) about the physical world. This paradigm is motivated by the increasing popularity of portable data collection devices (e.g., Smartphone's) and the massive data dissemination opportunities enabled by online social media. The first challenge is "misinformation spread" where a significant number of sources are contributing to false claims, making the identification of truthful claims difficult. For example, on Twitter, rumors, scams, and influence bots are common examples of sources colluding, either intentionally or unintentionally, to spread misinformation and obscure the truth. In this scheme, we work on these three important challenges that have not been well addressed in the current truth discovery literature.

GOALS AND OBJECTIVES

- To develop a novel Scalable Robust Truth Discovery (SRTD) scheme that explicitly considers various source

behaviors, content analysis of claims, and historical contributions of sources in a holistic truth discovery solution.

- To develop a light-weight distributed framework based on Work Queue to implement the SRTD scheme and improve computational efficiency.
- To achieve significant performance gain in terms of both effectiveness and efficiency compared to the baselines.

LITERATURE SURVEY

[1]Dong Wang Yang Zhang, "On Scalable and Robust Truth Discovery In Big Data Social Media Sensing Application"

In this paper we develop scalable and robust truth discovery (SRTD) scheme to address three challenges.

In this paper we can present a new scalable & robust approach to solve the truth discovery problem in big data. For example of social media sensing include real-time situation consciousness services in disaster or emergency response.

To solve the truth discovery problem rich set of principled approaches have been proposed in ML DM and Network sensing communication.

[2] Sagar Bhuta, AvitDoshi, Uehit Doshi, Meera Narvekar, “A Review of Techniques for Sentiment Analysis of Twitter Data”

Thus, it has been found out that a number of techniques can be used to perform sentiment analysis of text. But the methods are domain specific. Moreover the techniques need to be adapted to the source from which the data is extracted. If the source is a social networking website, the language use and specific conventions need to be addressed.

[3] Peter Bui, Dinesh Rajan, Badi Abdul-Wahid, Jesus Izaguirre, Douglas Thain, “Work Queue + Python: A Framework for Scalable Scientific Ensemble Applications”

Work Queue is a flexible and powerful framework for constructing scalable scientific ensemble applications. It provides attractive features such as fault-tolerance, data management, multiple scheduling algorithms, fast abort, and support for multiple distributed systems. With the introduction of the Python-Work Queue module, this functionality is now available to research scientists in a user-friendly programming language. In this paper, we present an overview of the Python- Work Queue module and then examine two real world scientific

applications that were built using Python Work Queue. The first is Rep ExWQ, which demonstrates the ability of Work Queue applications to span multiple distributed environments such as a local scampus cluster, a private grid, and a public cloud service provider. The second application is the Folding@ work data processing pipeline, which manifests Work Queue’s ability scale to over a thousand concurrent workers spread across different distributed systems. These applications are evidence of Work Queue’s effectiveness in constructing scientific ensemble applications that scale across multiple distributed systems.

[4] Jingwen Bian, Yang Yang, Hanwang Zhang, Tat-Seng Chua, “Multimedia Summarization for Social Events in Microblog Stream”, IEEE Transactions on Multimedia, 2014

In this paper, authors present a multimedia social event summarization framework which automatically generates holistic visualized summary from the microblogs of various media types. The presented framework features the exploration of the intrinsic correlations among different media types for enhancing the summarization performance. In particular, authors developed three major stages to accomplish the summarization. First,

devise an effective approach for eliminating the potentially noisy images from raw microblogs image collection.

Then, a novel Cross-Media-LDA (CMLDA) model, to discover sub-events from micro-blogs of difference media types, finally, generated multimedia summary for social events utilizing the cross-media distribution knowledge of the discovered sub-events.

PROPOSE SYSTEM

The architecture of the implemented SRTD system is shown in below. A key component is the Dynamic Task Manager (DTM), which is implemented as a master

Work Queue process that initializes a Worker Pool and dynamically spawns new tasks into the Task Pool. The DTM first divides the original TSC matrix into sub matrices as described in the previous section. Then, it spawns a set of tasks to process all sub matrices in parallel on the HT Condor system. A feedback control system is integrated with the SRTD scheme to monitor the current execution speed of each Truth Discovery (TD) task and estimate its expected finish time. The feedback control system informs the DTM of control signals based on system performance, and it dynamically adjusts the task priority and resource allocation to optimize the overall system performance.

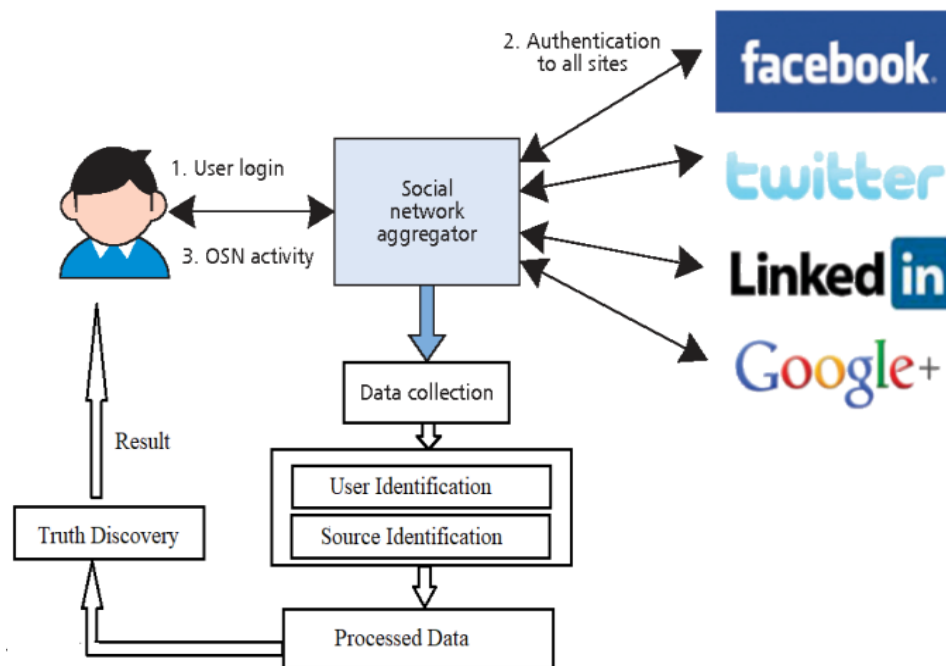


Figure: 1 System Architecture

CONCLUSION AND FUTURE SCOPE

In this system we proposed a Scalable Robust Truth Discovery (SRTD) framework to address the data veracity challenge in big data social media sensing applications. In our solution, we explicitly considered the source reliability, report credibility, and a source's historical behaviors to effectively address the misinformation spread and data sparsity challenges in the truth discovery problem. We also designed and implemented a distributed framework using Work Queue and the HT Condor system to address the scalability challenge of the problem. We evaluated the SRTD scheme using three real-world data traces collected from Twitter. The empirical results showed our solution achieved significant performance gains on both truth discovery accuracy and computational efficiency compared to other state-of-the-art baselines

REFERENCES

1. O. Banerjee, L. E. Ghaoui, and A. daspremont, "Model Selection Through Sparse Maximum Likelihood Estimation For Multivariate Gaussian Or Binary Data", Journal of Machine Learning Research, 2008.
2. S. Bhuta and U. Doshi. "A Review of Techniques for Sentiment Analysis of Twitter Data", In Proc. Int Issues and Challenges in Intelligent Computing Techniques (ICICT) Conf, 2014.
3. J. Bian, Y. Yang, H. Zhang, and T.-S. Chua, "Multimedia Summarization for Social Events in Microblog Stream", IEEE Transactions on multimedia, 2015.
4. P. Bui, D. Rajan, B. Abdul-Wahid, J. Izaguirre, and D. Thain, "Work Queue Python: A Framework For Scalable Scientific Ensemble Applications", In Workshop On Python For High Performance And Scientific Computing, 2011.
5. P.T. Chen, F. Chen, and Z. Qian, Road Traffic Congestion Monitoring In Social Media With Hinge-Loss Markov Random Fields, In Data Mining (ICDM), 2014 IEEE International Conference on, 2014.
6. X. L. Dong, L. Berti-Equille, and D. Srivastava, "Integrating Conflicting Data: The Role of Source Dependence", In Proceedings of the VLDB Endowment, 2009.

Cite this Article

M. Nigade, M. Raut, P. Mane, S. Phadatare (2019) **Truth Discovery in Big Data Social Media Application** Journal of Data Mining and Knowledge Engineering, 4(1) 40- 44

<http://doi.org/10.5281/zenodo.2653576>