On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications

ABSTRACT

- Identifying trustworthy information in the presence of noisy data contributed by numerous unvetted sources from online social media (e.g., Twitter, Facebook, and Instagram) has been a crucial task in the era of big data. This task, referred to as truth discovery, targets at identifying the reliability of the sources and the truthfulness of claims they make without knowing either a priori. In this work, we identified three important challenges that have not been well addressed in the current truth discovery literature.
- The 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 Twitter, rumors, scams, and influence bots are common examples of sources colluding, either intentionally or unintentionally, to spread misinformation and obscure the truth

Contd..

- The second 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% of sources 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. In this paper, we develop a Scalable and Robust Truth Discovery (SRTD) scheme to address the above three challenges.
- In particular, the SRTD scheme jointly quantifies both the reliability of sources and the credibility of claims using a principled approach. We further develop a distributed framework to implement the proposed truth discovery scheme using Work Queue in an HTCondor system. The evaluation results on three real-world datasets show that the SRTD scheme significantly outperforms the state-of-the-art truth discovery methods in terms of both effectiveness and efficiency.

EXISTING SYSTEM

- This task, referred to as truth discovery, targets at identifying the reliability of the sources and the truthfulness of claims they make without knowing either a priori. In this work, we identified three important challenges that have not been well addressed in the current truth discovery literature.
- The first one is “mis information spread” where a significant number of sources are contributing to false claims, making the identification of truthful claims difficult.

DISADVANTAGES

- the widely spread false information appears much more prominently than the truthful information, making truth discovery a challenging task.
- Our evaluation results on three real-world events demonstrate that current truth discovery solutions perform poorly in identifying truth when misinformation is widely spread.
- Second, many current truth discovery algorithms depend heavily on the accurate estimation of the reliability of sources, which often requires a reasonably dense dataset

PROPOSED SYSTEM

- we develop a Scalable and Robust Truth Discovery (SRTD) scheme to address the above three challenges. In particular, the SRTD scheme jointly quantifies both the reliability of sources and the
- credibility of claims using a principled approach. We further develop a distributed framework to implement the proposed truth discovery scheme using Work Queue in an HTCondor system. The evaluation results on three real-world datasets show that the SRTD scheme significantly outperforms the state-of-the-art truth discovery methods in terms of both effectiveness and efficiency.

ADVANTAGES

- 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., smartphones) and the massive data dissemination opportunities enabled by online social media.
- social media sensing include real-time situation awareness services in disaster or emergency response [, intelligent transportation system applications using location-based social network services [35], and urban sensing applications using common citizens .

HARDWARE REQUIREMENTS

- Processor :Intel Pentium IV 1GHz
- RAM :256MB (Min)
- Hard Drive :5GB free space
- Monitor :1024 * 768, High Color inch
- Mouse :Scroll Mouse(Logitech)
- Keyboard :104 keys

SOFTWARE REQUIREMENTS

- OS : Windows XP/7/8
- Front End : Visual Studio 2010/ netbeans 7.1
- Back End : SQL Server 2005/ heidisql 3.2
- Browser : Any Web Browser

CONCLUSION

- In this paper, 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 stateof- the-art baselines.
- The results of this paper are important because they provide a scalable and robust approach to solve the truth discovery problem in big data social media sensing applications where data is noisy, unvetted, and sparse.

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, 9(Mar):485–516, 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, pages 583–591, Feb. 2014.

[3] J. Bian, Y. Yang, H. Zhang, and T.-S. Chua. Multimedia summarization for social events in microblog stream. IEEE Transactions on multimedia, 17(2):216–228, 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 at sc11, 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, pages 80–89. IEEE, 2014.