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. 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.
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 HTCondor 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. 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.

SYSTEM REQUIREMENTS:

HARDWARE REQUIREMENTS:

• System : Pentium IV 2.4 GHz.
• Hard Disk : 40 GB.
• Floppy Drive : 1.44 Mb.
• Monitor : 15 VGA Colour.
• Mouse : Logitech.
• Ram : 512 Mb

SOFTWARE REQUIREMENTS:

• Operating system : Windows 7/UBUNTU.
• Coding Language : Java 1.7 ,Hadoop 0.8.1
• IDE : Eclipse
• Database : MYSQL

REFERENCES

