Efficient Truth Detection in BigData Online Social Media Applications

Ch.Shravya,S. Naga Raju
Department of CSE,
KAKATIYA INSTITUTE OF TECHNOLOGY AND SCIENCE, WGL
Telangana, India.

Abstract—Detecting trustworthy information within the sight of noisy data contributed by different unvetted sources from online social media (e.g., Twitter, Facebook, and Instagram) has been a pivotal errand in the period of big data. This assignment, alluded to as truth discovery, focuses at identifying the unwavering quality of the sources and the truthfulness of cases they make without knowing either apriori. In this work, we distinguished three important difficulties that have not been all around tended to in the present truth discovery writing. The first is "misinformation spread" where an important number of sources are contributing to false claims, making the ID of truthful claims troublesome. For instance, on Twitter, gossipy tidbits, tricks, and influence bots are basic instances of sources colluding, either intentionally or unintentionally, to spread misinformation and darken the truth. The second test is "data sparsity" or the "long-tail phenomenon" where a dominant part of sources just contributes few cases, providing insufficient proof to determine those sources' trustworthiness. For instance, in the Twitter datasets that we gathered during certifiable occasions, over 90% of sources just added to a single case. Third, many current arrangements are not scalable to large-scale social sensing occasions on account of the brought together nature of their truth discovery calculations. In this paper, we build up a Scalable and Robust Truth Discovery (SRTD) plan to address the over three difficulties. Specifically, the SRTD conspire jointly measures both the unwavering quality of sources and the believability of cases using a principled methodology. We further build up a disseminated framework to actualize the proposed truth discovery conspire using Work Queue in a HTCondor framework.

Keywords:Big Data, Truth Discovery, Rumor Robust, Sparse Social Media Sensing, Scalable, Twitter.

1.Introduction

This paper introduces another scalable and robust way to deal with tackle the truth discovery issue in big data social media sensing applications. Online social media (e.g., Twitter, Facebook, and Instagram) gives another sensing worldview in the big data time where individuals go about as omnipresent, inexpensive, and adaptable sensors to spontaneously report their perceptions (frequently called cases) about the physical world. This worldview is propelled by the increasing prevalence of compact data gathering gadgets (e.g., cell phones) and the gigantic data dissemination openings empowered by online social media. Instances of social media sensing include constant circumstance mindfulness benefits in a debacle or crisis reaction, intelligent transportation framework applications using area based social network administrations, and urban sensing applications using basic residents. A basic test that exists in social media sensing is truth discovery where the objective is to distinguish solid sources and truthful cases from enormous noisy, unfiltered, and notwithstanding conflicting social media data.

The truth discovery issue remains in the core of the veracity challenge of big data social media sensing applications. To tackle the truth discovery issue, a rich arrangement of principled methodologies has been projected in machine learning, data mining, and network sensing networks [1], [2], [3]. Be that as it may, three important difficulties still can't seem to be all around tended to by existing truth discovery arrangements in social media sensing applications.

To start with, current truth discovery arrangements don't completely address the "misinformation spread" issue where an important number of sources are spreading false data on social media. For instance, a bit of misinformation on Twitter saying that a 8-year-old young lady was murdered while running during the Boston Marathon has been so generally spread that the misinformation to debunking proportion was 44:1. In models this way, the generally spread false information seems much more conspicuous than the truthful information, making truth discovery a challenging undertaking.

Indeed, in this present reality Twitter datasets we gathered, over 90% of clients just contribute a single tweet. In such a situation where a larger part of sources contributes just few cases, there exists insufficient proof for precise estimation of source unwavering quality. Li et al. and Xiao et al. have expressly examined the issue of data sparsity and exhibited that many existing truth discovery calculations neglect to give great estimations to source unwavering quality when the dataset is scanty.

For instance, in an extraordinary situation where a client just posts one tweet, current truth discovery plans might probably distinguish binary estimations of unwavering quality (either 0 or 1), resulting in poor appraisals of real source dependability. Third, existing truth discovery arrangements did not completely investigate the versatility part of the truth discovery issue [6]. Social sensing applications regularly produce large measures of data during important occasions (e.g., debacles, sports, unrests). For instance, during the 2016 Super Bowl, 3.8 million individuals created a sum of 16.9 million tweets with a pinnacle rate of more than 152,000 tweets for every minute [7]. Current brought together truth discovery arrangements are incapable of handling such a large volume of social sensing data because of the asset constraint of a single computing gadget. A couple of conveyed arrangements have been created to address the adaptability issue of the truth discovery issue [5]. Notwithstanding, they undergo the mean effects of issues, for example, long startup times and ignorance of the heterogeneity of computational assets.

2.RELATED WORK:

2.1 Sentiment analysis:There has been a quick increase in the utilization of social networking sites over the most recent couple of years.[9] Individuals most advantageously express their perspectives and opinions on a wide exhibit of subjects by means of such sites. Sentiment analysis and opinion mining mainly focuses on opinions which express or imply positive or negative sentiments.

2.2 Truth detection:

Truth detection has established a significant amount of attentioninrecentyears, and previous studies haveestablished various models to address this vital challenge in big data applications. The truth discovery problem was first officiallydefinedbyYinetal. [10], Truth Finder, was proposed. Pasternack etal. extended this model by integrating prior knowledge of constraints into truth detection solutions and proposed several solutions. In our work truth detection is done by SRTD algorithm

3.Implementation

In this paper, we develop a Scalable and Robust Truth Discovery (SRTD) scheme to report the misinformation spread, data sparsity, and scalability challenges in big data social media sensing applications. To address the misinformation spread challenge, the SRTD scheme explicitly models different behaviors that sources exhibit, for example, copying/forwarding, self-correction, and spamming. To address data sparsity, the SRTD scheme hires aniterativealgorithm [8] that estimates truthfulnessby calculating the credibility sores.

3.1 SRTD algorithm:

The SRTD algorithm is an iterative algorithm that jointly computes the case truthfulness and source reliability by explicitly considering the credibility scores of the sources. We initialize the model with uniform case truthfulness scores and uniform source reliability scores. In each iteration, we first update the reliability score of each source using the truthfulness scores of cases reported by the source just as the commitment score of the source itself

```
SRTD algorithm
Input: data in the form of chunks
Output: claim truthfulness
1.Intialize $R_i$ =0.5, max iteration=100, set values for credibility scores ($SLS_{i,j}^{k}$)
2.compute contribution score
  $CS_{ij}$=sgn($SLS_{i,j}^{k}$) $\sum_{k=1}^{K} R_i^{K+1-k}|SLS_{i,j}^{k}$
3.Estimate $R_i$ reliability
  $R_i = \frac{\sum_{j\in F(i)} |CS_{ij}|(\chi(CS_{ij})D_j + (1-\chi(CS_{ij}))(1-D_j))}{\sum_{j\in F(i)} |CS_{ij}|}$
  $\chi(a) = \begin{cases} 1, a > 0 \\ 0, a \leq 0 \end{cases} \quad (7)$
4.compute
  $TC_j = \sum_{i\in k(j)} CS_{ij}$
5.Estimate claim truthfulness
  $D_j = \frac{1}{1+\exp(TC_{j)}}$
6.Repeat steps 2,3,4,5 until max iteration
7.If $D_j$ >= Threshold output=true else output= false
```

To address the scalability test, we develop a lightweight distributed framework by Work Queue [7]. We summarize our commitments as pursues:

a. First address three important challenges (i.e., misinformation spread, data sparsity, and scalability) in solving the truth discovery problem in big data social media sensing applications.
b. Second develop a novel SRTD scheme that explicitly considers different source behaviors, content analysis of cases, and chronicled commitments of sources in an all-encompassing truth discovery arrangement.
c. Third is to improve computational efficiency by developing a light-weight framework[distributed]based on Work Queue to evaluate SRTD scheme.

3.2 Overview of SRTD Architecture

The architecture of the implemented SRTD system is appeared in bellow figure. A key component is the Dynamic Task Manager (DTM), which is implemented as a master Work Queue process that initializes a Worker Pool and powerfully produces new errands into the Task Pool. The DTM first divides the original TSC grid into structural data as described in the previous section. Then, it produces a set of undertakings to process all structural data in parallel system. A feedback control system is integrated with the SRTD scheme to screen the current execution speed of each Truth Discovery (TD) assignment and estimate its expected finish time. The feedback control system informs the DTM of control signals based on system performance, and it progressively alters the undertaking need and resource assignment to optimize the overall system performance.

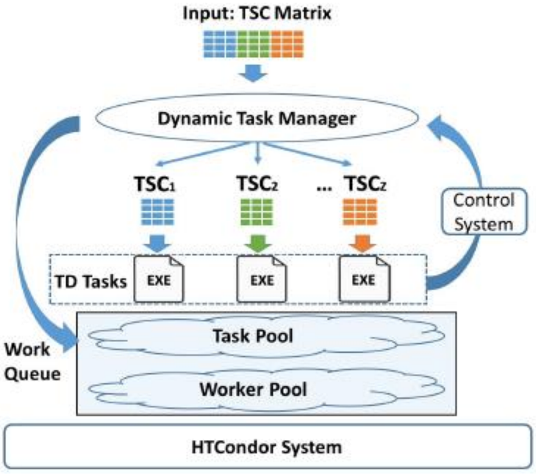


Fig: SRTD System Overview

In this section, we present a distributed implementation of the SRTD system using Work Queue. We initially introduce the Work Queue framework. Then, we present the implementation of the SRTD scheme, focusing on the allotment, management, and control of the distributed truth discovery assignments.

3.3 WorkQueue

Work Queue is a lightweight framework for implementing large-scale distributed systems [4]. This framework permits the master process to define a set of undertakings (i.e., Task Pool), submit them to the queue, and sit tight for completion. Work Queue maintains an elastic worker pool that enables users to scale the number of workers up or down as required by their applications. A worker is defined as a process that performs specific computational capacities described by the assignments. Once running, each worker gets back to the master process, arranges for data transfers, and executes the errands. We use Work Queue over the HTCondor system to take advantage of its dynamic resource portion mechanism for errand allotments.

3.4 Distributed Task Allocation

To make SRTD a scalable scheme, we divide the input data into multiple subsets and process them in parallel. Specifically, we first divide the TSC framework into Z structural data TSC1; TSC2; :::; TSCZ. The DTM then commences a TD task for each submatrix. The TD task performs the following operations:

1) Compute the source reliability.
2) Compute halfway case truthfulness TCj.
3) Wait for all other TD assignments related to Cj to finish, and compute Dj by aggregating all incomplete case truthfulness of Cj
4) Update the Contribution Scores of sources.
5) Repeat the above steps until SRTD converges.

Note that the third step requires sharing information among different TD tasks.We achieve this by sharing a typical directory between TD errands in the system. After computing the incomplete case truthfulness and source reliability scores, each TD task records the intermediate results (i.e. TC j and Ri) into a file in the shared directory. Before the end of each iteration, the DTM aggregates the results from files and updates the structural data for each TD task. We note that this sharing mechanism introduces I/O overhead to the performance of the SRTD scheme. However, we found that this I/O overhead is relatively little compared to the absolute execution time of the truth discovery algorithm, which is appeared in the evaluation results in the next section.

4. Experimental Data Processing

We developed a data crawler based on the Twitter open search API to collect these data traces by specifying query terms and the geographic regions related to the events. We noted that every one of the three datasets are very sparse. Based on our input dataset, just 1.4% of sources contribute more than two cases while 91.5% of sources contribute just a single case.

We conducted the following data preprocessing steps to prepare the datasets for the experiment: (I) cluster comparative tweets into the same cluster to generate claims; (ii) derive semantic link scores; (iii) generate the TSC Matrix; and (iv) generate ground truth labels. The details of these steps are summarized below.

Clustering: Here originally grouped comparative tweets into the same cluster using a variant of K-means algorithm that can effectively handle streaming Twitter data and the Jaccard distance to calculate the "distance" (i.e., similitude) between tweets. For each generated cluster, we picked a representative statement as the case and we take each Twitter user as the source for our model described.

Computing Credibility Score:

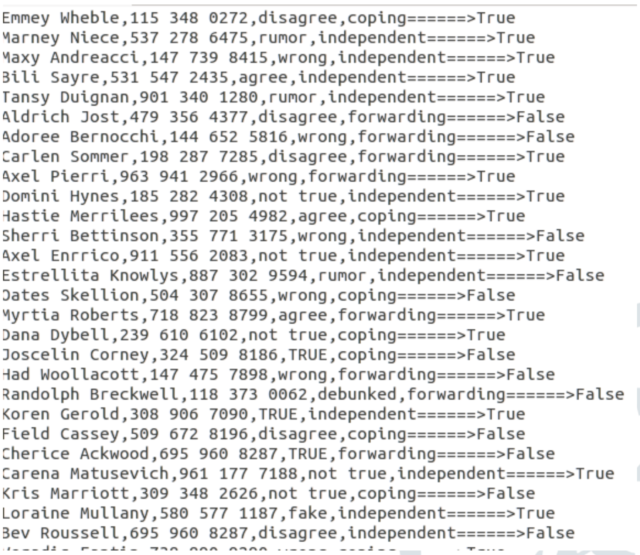
To compute the Credibility Score of a report (i.e., tweet), we originally calculated the Attitude Score of a source by performing a combination of sentiment analysis [9] and keyword matching. Specifically, we performed Polarity Analysis to detect tweets that express strong negative sentiment (with negativity value < 0.6) as "disagree" using the Python NLTK toolbox 3. We further captured disagreeing tweets based on whether it contains certain keywords, for example, "fake", "false", "debunked", "talk", "wrong", and "not true". We assigned a score of "1" and "-1" for non-disagreeing and disagreeing tweets respectively. We then calculated the Uncertainty Score by implementing a simple text classifier using drama learn and trained it with the data. To compute the Independent Score, we implemented a content to naturally label a tweet as "dependent" in the event that it is I) a retweet; or ii) significantly like other tweets (i.e., with a Jaccard distance less than 0.1) that were posted with earlier timestamps. We assigned a relatively low score to these dependent tweets.

Generating the Time-series Source-Claim Matrix:

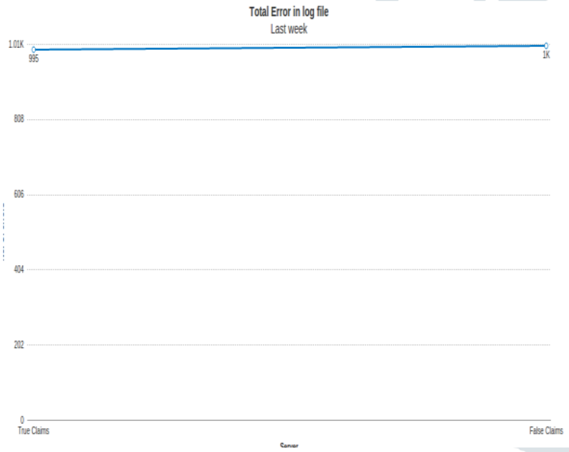
Here generated the TSC grid as pursues: for source Si, we recorded the majority of its reports (i.e., tweets) that were related to Cj based on the clustering results. We sorted the tweets in chronological order and for each tweet, we derived the credibility score based on equation (3). The resulting timeseries vector {SLS1 I,j, SLS2 i,j, … .SLSnI,j} i;jg was stored as the element TSCij in the grid.

4.1 EXPERIMENTAL RESULTS:

For experimental results in our work we took some sample data collected from social media. As discussed in the above section that is experimental data processing the datasets are being processed and given as input to SRTD algorithm. The output is in the form of a report and as graph.



```
Emmey Wheble,115 348 0272,disagree,coping======>True
Marney Niece,537 278 6475,rumor,independent======>True
Maxy Andreacci,147 739 8415,wrong,independent======>True
Bili Sayre,531 547 2435,agree,independent======>True
Tansy Duignan,901 340 1280,rumor,independent======>True
Aldrich Jost,479 356 4377,disagree,forwarding======>False
Adoree Bernocchi,144 652 5816,wrong,forwarding======>False
Carlen Sommer,198 287 7285,disagree,forwarding======>True
Axel Pierri,963 941 2966,wrong,forwarding======>True
Domini Hynes,185 282 4308,not true,independent======>True
Hastie Merrilees,997 205 4982,agree,coping======>True
Sherri Bettinson,355 771 3175,wrong,independent======>False
Axel Enrrico,911 556 2083,not true,independent======>True
Estrellita Knowlys,887 302 9594,rumor,independent======>False
Dates Skellion,504 307 8655,wrong,coping======>False
Myrtia Roberts,718 823 8799,agree,forwarding======>True
Dana Dybell,239 610 6102,not true,coping======>True
Joscelin Corney,324 509 8186,TRUE,coping======>False
Had Woollacott,147 475 7898,wrong,forwarding======>False
Randolph Breckwell,118 373 0062,debunked,forwarding======>False
Koren Gerold,308 906 7090,TRUE,independent======>True
Field Cassey,509 672 8196,disagree,coping======>False
Cherice Ackwood,695 960 8287,TRUE,forwarding======>False
Carena Matusevich,961 177 7188,not true,independent======>True
Kris Marriott,309 348 2626,not true,coping======>False
Loraine Mullany,580 577 1187,fake,independent======>True
Bev Roussell,695 960 8287,disagree,independent======>False
```



5.Conclusion

In this paper, we used SRTD algorithm to evaluate the truth discovery contributed from significant sources. The credibility scores are explicitly considered. Evaluation of the SRTD scheme is done by using real world datasets which are derived by experimental processing of sample twitter data. The empirical results showed our answer achieved significant performance gains on both truth discovery accuracy and computational efficiencycompared to other state-of-the-craftsmanship baselines. The results of this paper are important because they provide a scalable and robust solution for detecting the truthfulness where data is noisy, unvetted, and sparse contributed by social media sensing applications.

References

[1]X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: the role of source dependence. In Proceedings of the VLDB Endowment, pages 550–561, 2009.

[2]X. X. et al. Towards confidence in the truth: A bootstrapping based truth discovery approach. In Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16, 2016

[3] J. Marshall and D. Wang. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In Proceedings of the 10th ACM Conference on Recommender Systems, pages 167–174. ACM, 2016

[4]Nielson. Super bowl 50: Nielsen twitter tv ratings post-game report.

[5] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. Norman. Parallel and streaming truth discovery in large-scale quantitative crowdsourcing

[6]C. Huang, D. Wang, and N. Chawla. Scalable uncertainty-aware truth discovery in big data social sensing applications for cyberphysical systems. IEEE Transactions on Big Data, 2017.

[7]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

[8]Zhang, D., Wang, D., Vance, N., Zhang, Y., & Mike, S. (2018). *On Scalable and Robust Truth Discovery in Big Data Social Media Sensing Applications. IEEE Transactions on Big Data, 1–1.*doi:10.1109/tbdata.2018.2824812

[9]Bhuta, S., & Doshi, U. (2014). *A review of techniques for sentiment analysis Of Twitter data. 2014 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT).*doi:10.1109/icicict.2014.6781346