International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Detection of Truth Discovery in Big Data Social Media Sensing Applications

A Hemadri Naidu, J Naga Muneiah

Abstract: *With the rapid growth of online social media and ubiquitous Internet connectivity, social sensing has emerged as a new crowd sourcing application paradigm of collecting observations (often called claims) about the physical environment from humans or devices on their behalf. A fundamental problem in social sensing applications lies in effectively ascertaining the correctness of claims and the reliability of data sources without knowing either of them a priori, which is referred to as truth discovery. While significant progress has been made to solve the truth discovery problem, some important challenges have not been well addressed yet. First, existing truth discovery solutions did not fully solve the dynamic truth discovery problem where the ground truth of claims changes over time. Second, many current solutions are not scalable to large-scale social sensing events because of the centralized nature of their truth discovery algorithms. Third, the heterogeneity and unpredictability of the social sensing data traffic pose additional challenges to the resource allocation and system responsiveness. 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. The evaluation results on three real-world data traces (i.e., Boston Bombing, Paris Shooting and College Football) show that the SSTD scheme is scalable and outperforms the state-of-the- art truth discovery methods in terms of both effectiveness and efficiency.*

Keywords: *Big Data, SRTD, Data Sparsity, Robust, Social Media Sensing*

I. INTRODUCTION

This paper presents a scalable streaming truth discovery scheme for social sensing applications. Social sensing has emerged as a new paradigm of crowdsourcing applications where humans are used as ubiquitous, versatile and inexpen- sive sensors to report their observations (often called claims) about the physical world [35]. This paradigm is motivated by the proliferation of portable data collection devices (e.g., smartphones), the wide adaptation of online social media (e.g., Twitter, Facebook) and the ubiquitous Internet connectivity (e.g.,WiFi, 4G/5G). Examples of social sensing include ob- taining real-time situation awareness in disaster and emergency response scenarios [34], intelligent transportation system ap- plications using location based social network services [1], geotagging and urban sensing applications using inputs from common citizens [40]. A critical challenge in social sensing is referred to as *truth discovery* where the goal is to identify the reliability of the sources and the truthfulness of claims they make without the prior knowledge on either of them.

Consider a campus attack scenario (e.g., OSU attack in Nov. 2016) as an example. A significant amount of reports about the current situation of the attack (e.g., the number of casualties, the escape path of suspects and safety alerts) are available from the social sensors (e.g., news reporters and common citizens on social media). However, those social sensors may not always generate reliable claims and some of their claims may even contradict with each other. Table I shows some example tweets collected in the OSU attack. We observe the first two tweets report that there was a shooting happening at OSU campus while the third one report it was false news. In general, it is very challenging to identify the truthfulness of the claims without knowing the reliability of the individual sources who make them *a priori*. Additionally, sources could also intentionally or unintentionally propagate the misinformation through their social networks [38]. All these complexities make the truth discovery in social sensing a non-trivial task to accomplish.

Table-I: Example Tweets on Contradicting Claims in OSU Campus Attack, November, 2016.

Tweet	Timestamp
OSU POSSIBLE SHOOTING: I am on campus near @OSUengineering TONS of police.	28 Nov 2016, 7:23 AM
There was a shooting at Ohio state please pray for people's safety #osu	28 Nov 2016, 7:47 AM EST
Liberals putting out fake claims about the ter- rorist attack. 1st not a shooting, 2nd not an American, 3rd not nazi but Islamic #osushooting	28 Nov 2016, 11:37 AM EST

A rich set of solutions have been proposed in network sensing, data mining, machine learning communities to solve the truth discovery problem [7], [14], [17], [25], [37], [39], [41]. However, several significant challenges have not been well addressed by the state-of-the-art solutions. First, existing solutions did not fully solve the dynamic truth discovery problem where the ground truth of claims changes over time.

Revised Manuscript Received on December 30, 2019.

* Correspondence Author
Mr.A Hemadri Naidu*, M.Tech Student, Dept. of CSE, Chadalawada Ramanamma Engineering College, Tirupati, India. hemadrinaidu3@gmail.com
Prof. J Naga Muneiah, Head Department, of CSE, Chadalawada Ramanamma Engineering College, Tirupati, India. nagamuni513@gmail.com

There are two critical tasks in addressing the dynamic truth challenge.

The first is to capture the transition of truth in a timely manner and the second one is to be robust against noisy data that may lead to the incorrect detection of the truth transition. Only a small number of schemes been proposed to solve the dynamic truth discovery problem. For example, Pal et. al considered the evolving information of objects and esti- mated the truth of variables in current time interval based on sources' historical claims [24]. Li et al. proposed a Maximum A Posterior based real-time algorithm to solve the dynamic truth discovery problem [9]. However, these approaches could be unresponsive when the amount of social sensing data is large or the amount of resources on the deployed system is limited. Moreover, their solutions are not robust in noisy and sparse social sensing scenarios since their truth discovery accuracy is sensitive to the quality and quantity of thesensing data.

Second, existing truth discovery solutions did not fully explore the scalability aspect of the truth discovery problem. Social sensing applications often generate large amounts of data during some important events (e.g., disasters, sports, unrests) [27]. For example, 3.8 million people have generated a total of 16.9 million tweets with tweet per minute peaked at a rate of over 152,000 in Super Bowl 2016 [22]. However, current centralized truth discovery solutions are incapable of handling such large volume of social sensing data due to the resource limitation on a single computing device. A limited number of distributed solutions have been developed to address the scalability issue of the truth discovery problem. Both Ouyang et al. [23] proposed a distributed solution based on Hadoop system, but there are several non-trivial drawbacks. First, Hadoop is a heavy-weight solution in the sense that it requires a long start up time. Second, Hadoop is designed as a batch processing system that is most suitable for data of very large volume (e.g., Petabytes of data) and may not be the best solution for the size of datasets collected in many social sensing events (e.g., GB to TB). Third, Hadoop assumes homogeneity of the underlying computing nodes [29], which ignores the heterogeneity of the computational resources we have in real distributed systems.

The third challenge lies in the heterogeneity and unpre- dictability of the streaming data traffic. First, different topics partitions can be processed in a synchronized manner [23]. However, such strong homogeneity assumption on the data streams barely holds in real world social sensing applications. In this paper, we develop a Scalable Streaming Truth Discovery (SSTD) scheme to address the above challenges. To address the dynamic truth discovery challenge, we develop a Hidden Markov Model based solution to dynamically es- timate the true value of claims based on the observations reported by social sensors. To address the scalability challenge, we developed a light-weight distributed framework that is both *scalable* and *efficient* to solve the truth discovery problem using Work Queue and HTCondor system. To address the data heterogeneity challenge, we integrated the Scalable and Robust Truth Discovery (SRTD) scheme with an optimal workload allocation mechanism using feedback control (i.e., Proportional Integral Derivative (PID) controller) to dynamically allocate the resources (e.g., cores, memories) to the truth discovery tasks. We evaluated the Scalable and Robust Truth Discovery (SRTD) scheme in comparison with the state-of-the-art truth discovery baselines using three real-world social sensing data traces (i.e., Boston Bombing, Paris Shooting and College Football) collected from Twitter. The evaluation results show that our Scalable and Robust Truth Discovery (SRTD) scheme significantly outperforms the compared baselines in terms of truth discovery accuracy and computational efficiency.

In summary, the contributions of this paper are as follows:

- This paper addresses three fundamental challenges in truth discovery problem in social sensing: *dynamic truth*, *scalability* and *heterogeneity of streaming data*.
- We develop the Scalable and Robust Truth Discovery (SRTD) scheme that incorporates the Hidden Markov Model (HMM) to effectively address the dynamic truth discovery challenge.
- We develop a light-weight distributed framework based on Work Queue and HTCondor system to address the scalability challenge.
- We integrate the Scalable and Robust Truth Discovery (SRTD) scheme with an optimal workload allocation mechanism to address the heterogeneity of the streaming social sensing data.
- We evaluate the performance of the Scalable and Robust Truth Discovery (SRTD) scheme and compare it with the state-of-the-art truth discovery solutions through real-world case studies. The evaluation results demonstrate the effectiveness and significant per- formance gains achieved by our scheme.

II. PROBLEM FORMULATION

In this section, we formulate our robust truth discovery problem in big data social media sensing. In particular, consider a social media sensing application where a group of M sources $S = (S_1, S_2, \ldots, S_M)$ reports a set of N claims, namely, $C = (C_1, C_2, \ldots, C_N)$. Let S_i denote the ith source and C_j denote the jth claim. We define $RP^t_{i,j}$ to be the report made by source S_i on claim C_j at time t. Take Twitter as an example; a source refers to a user ac- count and a claim is a statement of an event, object, or topic that is derived from the source's tweet. For example, a tweet "Not much of the comment about the Dallas shooting has focused on the fact the sniper was a veteran." is associated with a claim "Dallas shooting sniper was a veteran". The tweet itself is considered as the report. We observe that the social media sensing data is often sparse (i.e., the majority of sources only contribute to a limited number of claims in an event). We further define $C_j = T$ and $C_j = F$ to represent that a claim is true or false, respectively. Each claim is also associated with a ground truth label x^*_j such that $x_j = 1$ when C_j is true and $x_j = 0$ otherwise.

Retrieval Number: B3557129219/2019©BEIESP
DOI: 10.35940/ijeat.B3557.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

The goal of the truth discovery task is to jointly estimate the truthfulness of each claim and the reliability of each source, which is defined as follows:

***DEFINITION 1.* Claim Truthfulness D_j for claim C_j:** The likelihood of a claim to be true. The higher D_j is, the more likely the claim C_j is true.

Formally we define D_j to estimate:

$Pr(C_j = T)$

***DEFINITION 2.* Source Reliability R_i for source S_i:** A score represents how trustworthy a source is. The higher R_i is, the more likely the source S_i will provide credible and trustworthy information. Formally we define R_i to estimate:

$$Pr(C_j = T|SC_{i,j} = T)$$

where $SC_{i,j} = T$ denotes that source S_i reports claim C_j to be true.

Since sources are often unvetted in social media sensing applications and may not always report truthful claims, we need to explicitly model the reliability of data sources in our problem formulation. However, it is challenging to accurately estimate the reliability of sources when the social media sensing data is sparse [34]. Fortunately, the reports themselves often contain extra evidence and information to infer the truthfulness of a claim. In the Twitter example, the text, pictures, URL links, and geotags contained in the tweet can all be considered as extra evidence of the report. To leverage such evidence in our model, we define a *credibility score* for each report to represent how much the report contributes to the truthfulness of a claim.

We first define the following terms related to the credibility score of a report made by source S_i on claim C_j at time k.

DEFINITION 3. Attitude Score (ρk) : Whether a source believes the claim is true, false or does not provide any report. We use 1, -1 and 0 to represent these attitudes respectively.

DEFINITION 4: Hidden States of Truth: the true value for the claim at a given time instant that is not directly observable.

DEFINITION 5. Independent Score: (ηk): A score in the range of (0,1) that measures whether the report Ri,u is made independently or copied from other sources. A higher score is assigned to a report that is more likely to be made independently.

III. PROPOSED ALGORITHMS

This Paper describes the three challenges nothing but the Misinformation, data sparsity and trustworthiness using the SRTD(Scalable and robust trust discovery scheme). Certain observations are made which are relevant to our model as follows

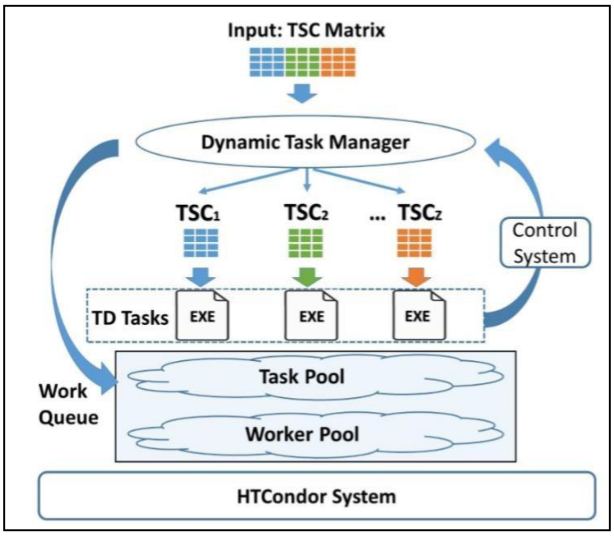
 Observation 1: Sources often spread false information from others without independent verification by simply copying or forwarding information (e.g., retweets on Twitter).

 Observation 2: False claims have intensive debates on those claims and often controversial and sources tend to disagree with each other.

 Observation 3: If a source debunk sits previous claim, it's very likely the previous claim is false because people are generally prone to be self-consistent.

More specifically, the contribution score of source Si on claim Cj is denoted as CSij and it is formally calculated as: CSij = sgn(SLSK i,j)K X k=1RK+1−k i |SLSki,j|

The following diagram explains about the SRTD architecture



A. Architecture of SRTD

Usually from the above diagram Dynamic Task Manager which implemented as a master work Queue process that initializes a work pool, dynamically spawns new task into the task pool and it considered as a key component. DTM divides the TSC matrix into sub matrices then it spawns set4 of tasks to process all submatrices on the HTCondor system. SRTD is always integrated with the feedback control system to monitor the current execution speed of each Truth Discovery task and also to estimate its expected finish time. DTM was informed by the feedback control system of control signals based on the performance of the system and it dynamically adjust the task priority and resource allocation to optimize the overall system performance.

B. SRTD iAlgorithm

1) Algorithm 1 Scalable Robust Truth Discovery (SRTD)

Input: TSC matrix

Output: claim truthfulness ˆ x∗ j,∀1 ≤ j ≤ N

Initialize Ri = 0.5,∀i ≤ M; set the values of credibility scores; initialize max iteration = 100

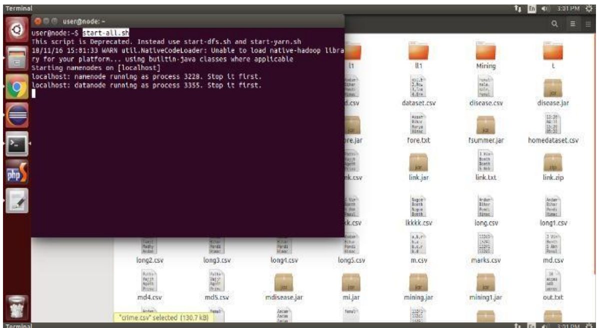
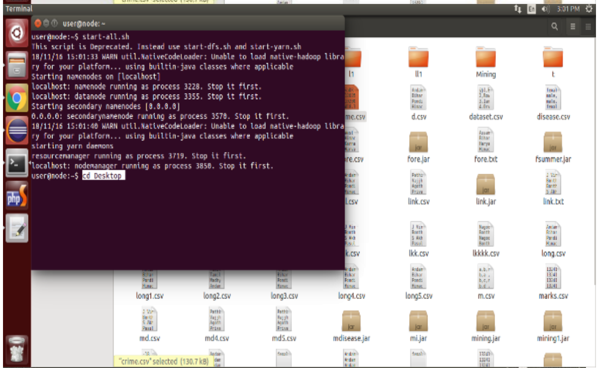
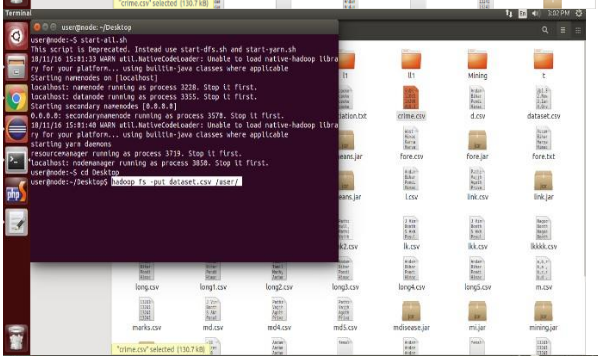
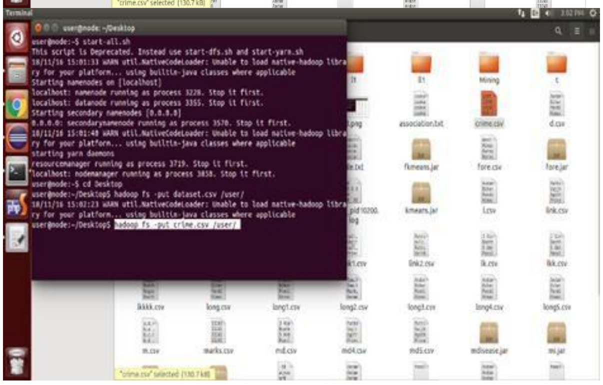
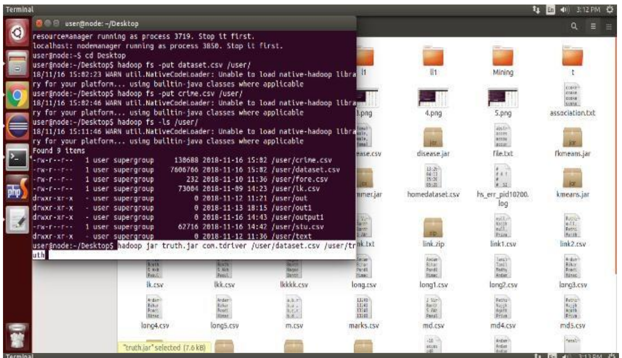
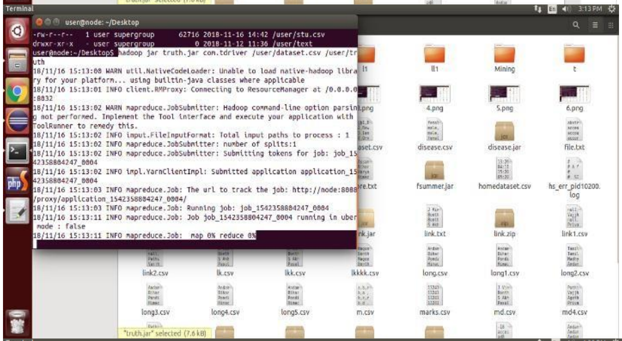
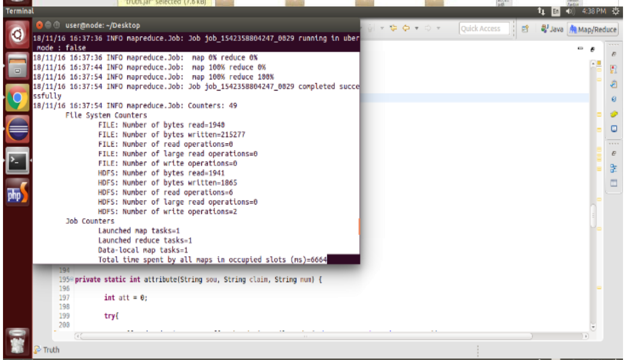
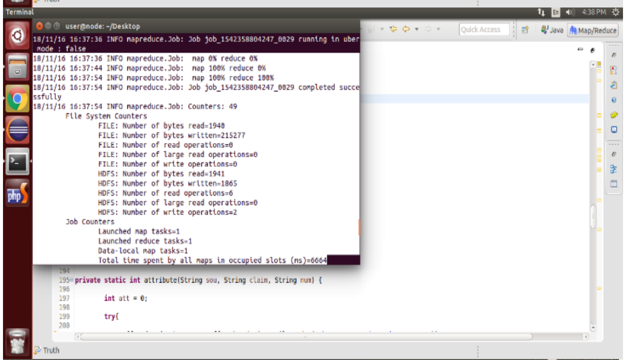
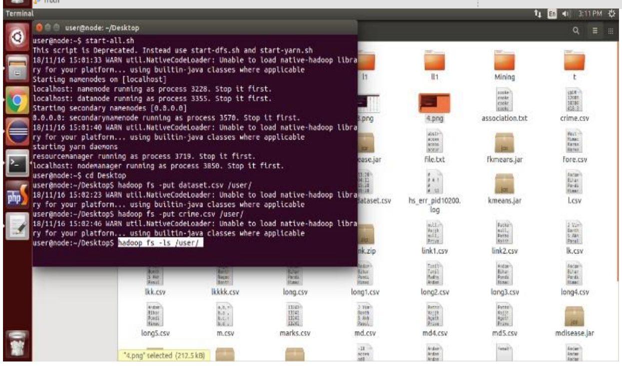
Split Original TSC matrix into Z submatrices, let S(z) denote the number of sources in the z-th submatrix while{Dj}do not converge or reach max iteration do

for all z,1 ≤ z ≤ Z do for all i,1 ≤ i ≤ S(z) do for all j,1 ≤ j ≤ N do

if TSCij exists then
compute CSij based on Equation (6)
end if
end for
end for
for all i,1 ≤ i ≤ S(z) do
estimate Ri based on Equation
(7)

end for
for all $j, 1 \leq j \leq N$ do
compute $TCz\ j$ based on Equation (10)
end for
estimate Dj based on Equations (11) and (12)
end for
end while
for all $j, 1 \leq j \leq N$ do if $Dj \geq$ threshold then output $\hat{}\ x*\ j = 1$
else
output $\hat{}\ x*\ j = 0$
end if
end for

IV. RESULTS AND ANALYSIS

Retrieval Number: B3557129219/2019©BEIESP
DOI: 10.35940/ijeat.B3557.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

V. 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 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. P. Bui, D. Rajan, B. Abdul-Wahid, J. Izaguirre, and D. Thain. Work queue+ python: A framework for scalable scientific ensemble appli- cations. In Workshop on python for high performance and scientific computing at sc11, 2011.
2. Z. Z. J. Cheng and W. Ng. Truth discovery in data streams: A single- pass probabilistic approach. In In Proc. of CIKM, pages 1589–1598, 2014.
3. 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.
4. Q. L. et al. A confidence-aware approach for truth discovery on long- tail data. In Proceedings of the VLDB Endowment, volume 8, pages 425–436, Dec. 2014.
5. Y. L. et al. On the discovery of evolving truth. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15, 2015.
6. R. Farkas, V. Vincze, G.Mora, J. Csirik, and G.Szarvas. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. In In Proceedings of the Fourteenth Conference on Computational Natural Language Learning., 2010.
7. A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In In Proc. of the ACM Interna- tional Conference on Web Search and Data Mining (WSDM'10), pages 131–140, 2010.
8. M. Ham, Y. H. Lee, and J. W. Fowler. Integer programming-based real- time scheduler in semiconductor manufacturing. In Winter Simulation Conference, pages 1657–1666. Winter Simulation Conference, 2009.
9. C. Huang and D. Wang. Spatial-temporal aware truth finding in big data social sensing applications. In Proceedings of Trust-com/BigDataSE/ISPA, volume 2, pages 72–79. IEEE, 2015.
10. C. Huang and D. Wang. Topic-aware social sensing with arbitrary source dependency graphs. In International Conference on Information Processing in Sensor Networks (IPSN), pages 1–12. ACM/IEEE, 2016.
11. C. Huang and D. Wang. Critical source selection in social sensing applications. In Distributed Computing in Sensor Systems (DCOSS), 2017 International Conference on to appear. IEEE, 2017.
12. C. Huang, D. Wang, and N. Chawla. Towards time-sensitive truth discovery in social sensing applications. In Proceedings of International Conference on Mobile Ad Hoc and Sensor Systems (MASS), pages 154–
13. 162. IEEE, 2015.
14. C. Huang, D. Wang, and N. Chawla. Scalable uncertainty-aware truth discovery in big data social sensing applications for cyber-physical systems. IEEE Transactions on Big Data, 2017.
15. M. J. Litzkow, M. Livny, and M. W. Mutka. Condor-a hunter of idle workstations. In Distributed Computing Systems, 1988., 8th Interna- tional Conference on, pages 104–111. IEEE, 1988.
16. J. Marshall, M. Syed, and D. Wang. Hardness-aware truth discovery in social sensing applications. In Distributed Computing in Sensor Systems (DCOSS), 2016 International Conference on, pages 143–152. IEEE, 2016.
17. J. Marshall and D. Wang. Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In Proceedings of Interna- tional Conference on Recommender Systems (Recsys), pages 167–174. ACM, 2016.
18. J. Marshall and D. Wang. Towards emotional-aware truth discovery in social sensing applications. In Smart Computing (SMARTCOMP), 2016 IEEE International Conference on, pages 1–8. IEEE, 2016.
19. nielson. Super bowl 50: Nielsen twitter tv ratings post-game report.
20. R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. Nor- man. Parallel and streaming truth discovery in large-scale quantitative crowdsourcing.
21. J. Qadir, A. Ali, A. Zwitter, A. Sathiaseelan, J. Crowcroft, et al. Crisis analytics: Big data driven crisis response. arXiv preprint arXiv:1602.07813, 2016.
22. L. Rabiner and B. Juang. An introduction to hidden markov models.
23. ieee assp magazine, 3(1):4–16, 1986.
24. B. T. Rao, N. Sridevi, V. K. Reddy, and L. Reddy. Performance issues of heterogeneous hadoop clusters in cloud computing. arXiv preprint arXiv:1207.0894, 2012.
25. D. E. Rivera, M. Morari, and S. Skogestad. Internal model control: Pid controller design. Industrial & engineering chemistry process design and development, 25(1):252–265, 1986.
26. T. Shelton, A. Poorthuis, M. Graham, and M. Zook. Mapping the data shadows of hurricane sandy: Uncovering the sociospatial dimensions of 'big data'. Geoforum, 52:167–179, 2014.
27. M. Y. S. Uddin, M. T. A. Amin, H. Le, T. Abdelzaher, B. Szymanski, and T. Nguyen. On diversifying source selection in social sensing. In Proc. Ninth Int Networked Sensing Systems (INSS) Conf, pages 1–8, June 2012.
28. A. Viterbi. Error bounds for convolutional codes and an asymptoti- cally optimum decoding algorithm. IEEE Transactions on Information Theory, 13(2):260–269, Apr. 1967.
29. D. Wang, T. Abdelzaher, and L. Kaplan. Surrogate mobile sensing.
30. IEEE Communications Magazine, 52(8):36–41, 2014.
31. D. Wang, T. Abdelzaher, and L. Kaplan. Social sensing: building reliable systems on unreliable data. Morgan Kaufmann, 2015.
32. D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal. Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications. In The 33rd International Conference on Distributed Computing Systems (ICDCS'13), July 2013.
33. D. Wang, M. T. Al Amin, T. Abdelzaher, D. Roth, C. R. Voss, L. M. Kaplan, S. Tratz, J. Laoudi, and D. Briesch. Provenance-assisted classification in social networks. IEEE Journal of Selected Topics in Signal Processing, 8(4):624–637, 2014.
34. D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu,
35. Pan, H. Liu, C. C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra,
36. B. Szymanski, and H. Le. Using humans as sensors: An estimation-theoretic perspective. In Proc. 13th Int Information Processing in Sensor Networks Symp. IPSN-14, pages 35–46, Apr. 2014.
37. D. Wang and C. Huang. Confidence-aware truth estimation in social sensing applications. In International Conference on Sensing, Commu- nication, and Networking (SECON), pages 336–344. IEEE, 2015.
38. D. Wang, L. Kaplan, and T. F. Abdelzaher. Maximum likelihood analysis of conflicting observations in social sensing. ACM Transactions on Sensor Networks, 10(2):1–27, Jan. 2014.
39. D. Wang, L. Kaplan, H. Le, and T. Abdelzaher. On truth discovery in social sensing: A maximum likelihood estimation approach. In Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf, pages 233–244, Apr. 2012.
40. J. Wang, D. Wang, Y. Zhao, and T. Korhonen. Fast anti-collision algorithms in rfid systems. In Mobile Ubiquitous Computing, Systems, Services and Technologies, 2007. UBICOMM'07. International Confer- ence on, pages 75–80. IEEE, 2007.
41. W. Xue, J. Shi, and B. Yang. X-rime: cloud-based large scale social net- work analysis. In Services Computing (SCC), 2010 IEEE International Conference on, pages 506–513. IEEE, 2010.
42. S. R. Yerva, H. Jeung, and K. Aberer. Cloud based social and sensor data fusion. In Information Fusion (FUSION), 2012 15th International Conference on, pages 2494–2501. IEEE, 2012.

43. X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. IEEE Transactions on Knowledge and Data Engineering, 20(6):796–808, June 2008.
44. D. Zhang, H. Rungang, and D. Wang. On robust truth discovery in sparse social media sensing. In Big Data (Big Data), 2016 IEEE International Conference on. IEEE, 2016.

AUTHORS PROFILE

Mr. Hemadri Atukuru received the B.Tech (CSE) from Sri Venkateswara college of engineering tirupati affiliated to Jawaharlal Nehru Technological University anantapur, India in 2016. He is pursuing his M.Tech in CSE from Chadalavada ramanamma engineering college Tirupati. His areas of interests include Big data, Data structures and Algorithms.

Prof. Janapati Naga Muneiah received the B.Tech (CSE) from Jawaharlal Nehru Technological University, Hyderabad, India in 2001 and M.Tech in CSE from Sri Venkateswara University, Tirupati, India in 2010. He is pursuing his Ph.D in Jawaharlal Nehru Technological University, Kakinada, India in Computer Science and Engineering faculty. He has got 17 years of teaching experience. Presently he is working as Professor and Head of Department of CSE in Chadalawada Ramanamma Engineering college, Tirupati, A.P, India. His areas of interests include Data Mining, Data Warehousing, Big Data, Data Structures and Algorithms. He has guided 11 M. Tech theses. He has published more than 10 papers in International journals and some of them are published in SCIE and SCOPUS indexed journals.

Retrieval Number: B3557129219/2019©BEIESP
DOI: 10.35940/ijeat.B3557.129219
Journal Website: www.ijeat.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication