TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

A Survey on Reliable Source Discovery in Social Media for Big Data

[1]J. Lysa Eben, [2]R. Renuga Devi,
[1]Research Scholar, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai. India.
[2]Assistant Professor, Vels Institute of Science, Technology & Advanced Studies (VISTAS), Chennai. India.
[1]lysaeben@rediffmail.com, [2]nicrdevi@gmail.com

Article Info
Volume 82
Page Number: 16573 - 16584
Publication Issue:
January-February 2020

Article History
Article Received*: 18 May 2019*
Revised*: 14 July 2019*
Accepted*: 22 December 2019*
Publication*: 28 February 2020*

Abstract:
The increasing growth and usage of online social network has set a spark to the big data era. Massive amount of data from various online social media sources are being generated by the humans now-a-days. Human act as social sensors through their observations and claims. The crucial task is to identify the reliability of these observations made, which is referred as truth discovery in the presences of noisy data. As the reliability of information must be identified where the observation made by the sensor may be true or false leading to binary claims. The truth discovery problem is to infer the correctness in the claimed observations. This problem is addressed by finding the reliability of the source. This survey concentrates on the existing truth discovery solutions available for social media sensing applications that are used to identify true information from among the conflicting data and an appropriate method can be chosen grounded on the comparison of methods and type of data.

1. INTRODUCTION

In this digital era, huge amount of information transfer happens as part of our life. In our daily routine we continuously generate data of different forms using different medium like Social networks, Twitters, Facebook, Blogs, Discussion forum,Smart phones, Crowd sourcing platforms,etc. This massive amount of data generated is called Big data. This data happens to be in huge Volume, Variety, Velocity and Veracity. Collecting, Storing, Describing, Modeling and Processing this data is a crucial task.Analyzinghuge amount of data is of great interest for individuals to gain knowledge of information, by the government for security checks and decision making, in business for aggregation of opinions and by researches for discovering knowledge.

The situation prevalent here is the data describing an event or matter which come from different sources, these data might conflict each other on the same scenario or matter due to error in the information or data,types, missing records, updated or outdated entries etc. This raises the question of reliability or trustworthiness of the data. An example to describe this fact is, when searched for the height of "Mount Everest" using the common search engine google,heights like 29,035', another like 29,002' and one like 29,029' were posted by different site [1]. This is one situation and similar problems exist. It is necessary to find the correct or reliable or true data from among the noisy information. To do so, noisy information from various sources for the same event must be aggregated to get a fact.

Majority voting/ averaging is one straight forward approach to reduce the conflicts. This conflict among multi- source data can be eliminated by conducting majority voting or averaging which is one straight forward approach. This method works on the principle of assumption. And the assumption is, every source

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

is equally reliable. The drawback of such approach is that assumptions may not always hold good or true or trustworthy. As with the example mentioned about the height of "Mount Everest" majority voting resulted in "29,035 feet" as its height, which was not true. The search result "29,029 feet"[1] given by Wikipedia was true. The inference from the above example reveals the fact that information standard varies among different sources and the accuracyanalysis of aggregated resultscan be enhanced by considering the reliability of the sources.

Identifying the source reliability which is unknown in priori must be inferred from the data which is a challenge. There emerges the topic of "Truthdiscovery"[4, 5, 6, 7] which is gaining its fame in recent times due to its ability to assesssource reliabilitydegree [1] and infer true information.This method used for truth discovery work without supervision, hence the reliability or credibility of the source can then be derived based on the data. Thus, the existing truth detection approach provides higher score for sources that often provide true information and the data provided by these sources are considered as truths. In general, truth discovery approach fit several scenarios. Approaches make different assumptions about the data that is input, relations that exist among sources, identified facts or truths etc... This diversification in approaches make the selection choice difficult for commoner. This survey paper gives on overall idea of the approaches and summarizes them from different aspects.

Discovering the accurate information play a vital role in taking (or) making critical decisions in various application fieldson the basis reliable data extracted from diverse source. The major benefiting sectors are Healthcare industry [9], Crowd Social sensing [2, 8, 11,14,15],Crowd sourcing [3, 16, 17], Information extraction [8,18], Knowledge graph construction [12,13].

This paper is organized as follows:Section – 1introduction of truth discovery. Section – 2 explains the literature review. Section – 3formally define the truth discovering task. Section – 4discusses popular truth discovery methods, Section – 5 components involved in truth discovery are examined and covers different truth discovery approaches. Section – 6 includes comparative study of truth discovery approaches, Section – 7 concludes the survey and suggest the future work.

2. LITERATURE REVIEW

Xiaoxin Yin, Wenzhao Tan [21] a semi – supervised approach was proposed to find true values using small set of ground truth data as well as identification of trustworthy data sources. For a given problem a confidence score for each unlabelled fact was assigned. And the confidence score took real values between the range -1 and 1 were a score close to 1 indicates high confident, that a fact is true. A score value close to -1 indicates the opposite and 0 indicates a state of dilemma if fact is true or false. The purpose of the approach is to help make predictions consistent for both labelled data and the graph structure. Even small amount of ground truth data greatly helps identify trust worthiness of data source. To arrive at an optimal solution for a problem an iterative algorithm was proposed to converge to it.

Dong Wang, Lance Kaplan, Tarek F. Abdelzaher [22]in maximum likelihood analysis of conflicting observations in social sensing addresses, the challenge to identify truth from noisy social sensing data. Growth of internet leads to new paradigm of social sensing where humans act as social sensors. The question of trustworthiness in human sensor data arises. Maximum likelihood solution to truth discovery from corroborating observation was proposed. EM scheme was proposed to efficiently solve truth discovery problem in social sensing to an optimal problem.

Xiaoxin Yin, Jiawei Han, Philip S.YU [23] the author addresses a new problem called veracity which studies the ways to trace true facts from a large amount of conflicting information on

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

different subjects provided by various web sites. An algorithm called truth finder for identifying true facts from among conflicting information using iterative methods was proposed. The concept of implication between facts was proposed to represent relationships *imp (f1->f2)* is f1's influence on f2's confidence and it takes value between -1 and 1, a positive value is an indication of f1 is correct and f2 likely to be correct, a negative value indicates f1 is correct and f2 likely to be wrong. Truth finder achieves high accuracy in finding true facts as well as identifying web sites that provide accurate information.

XiaolongXu,et al [24] proposes a truth finder algorithm which is based on entity attributes (TFAEA). It is an iterative method. It goes through four stages of processing – pattern matching, collision detecting, truth finder and data fusion. The first stage of TFAEA initial values for all reliabilities of data source are set the mutual support degree between facts and the dependent relationship between data sources are calculated, which improves the computational efficiency of the algorithm. TFAEA is not only stable but also more accurate than other truth finder algorithms.

Daniel (Yue) Zhang, Dong Wang, Yang Zhang [25] addresses "Physical Constraint-Awareness" and "Noisy incomplete data" problem. It developed constraint – aware Hidden Markov model to productively judge the evolving truth of measured variables by including physical constraint. To deal with the noisy and incomplete data challenge CA-DTD fuses sensing observations from both online social media and traditional news media using a principled approach CA-HMM.

HoupingXiao,et al [26]proposed ETCIBOOT a confidence interval estimate to identify truths were bootstrapping technique are integrated into the truth discovery procedure. Instead of providing a point estimator for inferring the objects truth, the more desirable confidence – interval estimates to identify truths is addressed. ETCIBOOT obtain a better truth estimate.

Surender Reddy Yerv, et al [37]explores the possibility of fusing social and sensor data in the cloud. For the study people mood information from tweeter associated with weather data is extracted. A numerical score is assigned for words occurring frequently and represented in ANEW list. The words in the list is assigned values in 3dimensions: Valence, Arousal and Dominance, were the value in each dimension range from 1 to 9. For overall computation conditional probabilities is adopted. To demonstrate the effectiveness of the proposal, experiments were conducted on 12 million tweets and proved efficient.

Jermaine Marshall, Dong Wang, [47]presented an analytical model to solve the mood sensitivity to discovery problem in social sensing. The new approaches solve a multidimensional estimation problem by developing a new expectation maximization (EM) based algorithm. The mood sensitivity expectation maximization scheme jointly estimates the mood sensitivity of each claim and mood sensitivity of each source. To test result four real world data sets were collected form tweeter the result demonstrated significant performance gain.

Xin Luna Dong, et al [2]considers integrating data from the web truth discovery, examined on static information from different data sources. For precise truth discovery results it considers possible dependency between sources. Bayes model is adopted to compute the probability of two data sources being dependent. To derive truth discovery, source dependency and discovering truth from conflicting information are iteratively solved. The proposed algorithm on testing with synthetic and real time datasets showed significance.

TEST Engineering & Management

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

3. EXPLORING TRUTH DISCOVERY

The concentration of the survey paper is to identify and resolve conflicts from among multi-source noisy data. Since a general observation commonly identifies conflicts in the information generated from the web [38], crowd souring data [47] etc. A common practice to find the truth among the same group of mixed information, voting or averaging approaches are followed, but the drawback of this approach is all source information are treated equally which might not hold good always.In contrast truth discovery aim to identify truth by considering the source reliability by assigning weights, based on the credibility of the information a source provides. The section clearly describes the problem and illustrates Truth discovery approach with an example.

3.1. Problem definition

The following are the definition and notation used**:** A data or information of interest is referred as object o. [1] various web that provide the information provided by sources on the search topic is referred as value v_o^s. The most reliable object from among the information is referred with the value v_o^*. The source which provides true information often is assigned weights to reflect the probability is referred as source weight w_s.

3.2. Illustration with example

The aim is to identify the birthplace of six famous personalities. Information is provided by three sources which have conflicting information. Truth is identified by majority voting and truth discovery approaches respectively [1].

Table 1: Illustrative Example [1]

	George Washington	Abraham Lincoln	Mahatma Gandhi	John Kennedy	Barack Obama	Franklin Roosevelt
Source 1	Virginia	Illinois	Delhi	Texas	Kenya	Georgia
Source 2	Virginia	Kentucky	Porbander	Massachusetts	Hawaii	New York
Source 3	Maryland	Kentucky	Mumbai	Massachusetts	Kenya	New York
Majority Voting	Virginia	Kentucky	Delhi	Massachusetts	Kenya	New York
Truth Discovery	Virginia	Kentucky	Porbander	Massachusetts	Hawaii	New York

The output shows that truth discovery approaches is more acceptable compared to majority voting. Since even in case of a tie, minority value result is predicted reliable with weighted aggregation [20].Next, a brief on three popular methods that incorporates the above principles in truth discovery is discussed.

4. POPULAR TRUTH DISCOVERY METHODS

In general truth discovery approaches follow this procedure. To identify the true data, weighted aggregation of the multisource data is performed based on the estimated source reliabilities. If a source often provides true information, then its reliability is high. So higher score is assigned to sources that have provided correct data. The majority voting method will have no chances to select the correct datawhen minority claim is true. On the other hand, truth discovery method identifiesreliable source based on source reliability score from that of unreliable sources. So, reliable information can be obtained by carrying out weighted aggregation.

4.1 Iterative method

The iterative method for truth discovery evaluates the source reliability in two steps. The first step computes the objects truth, the second step estimates the source weight. The steps are iterated until both object truth and source weight value converges. Weighted voting is used to

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

identify the truths $\{v_o^*\}_{o\in O}$ derived using weighted aggregation, such as weighted voting. Each candidate value v receives the votes form sources in the following ways.

$$Vote\ v = \left(\sum_{s\in S_v} \frac{w_s}{|v_s|}\right)^{1.2} \quad -1$$

Where S_V is the set of sources that provide this candidate value and $|V_s|$ is the number of claims made by source s. The truths are inferred by ranking the votes received the end results $\{v_o^*\}_{o\in O}$ relay more on sources with high weights.

For truth detection in computation step the source assigns equal reliability among claimed values. In the secondsource weight computation step, they discount values back from the truths as follows [1].

$$w_s = \sum_{v\in v_s}\left(vote\ (v).\frac{\frac{w_s}{|v_s|}}{\sum_{s\in S_v}\frac{w_s}{|v_s|}}\right) \quad -2$$

4.2 Optimization based methods

The main objective of optimization-based method is to measure the weighted distance between the information provider and the identified truth $\{v_o^*\}$. This method is based on the general principle of truth discovery[3, 27, 36, 46]. The optimization formulation that is used to capture the truth is as follows,

$$\underset{\{w_s\}\{v_o^*\}}{arg\,min}\sum_{o\in O}\sum_{s\in S} w_s\ .\ d(v_o^s, v_o^*) \quad -3$$

The distance function denoted by d (.) is the measurement of difference between the given input data by any valid source denoted by (s) and the identified truths $\{v_o^*\}$.By further factorising this function the resulted aggregate is narrowed down to minimized. The data provided through the source will be of high weights. Likewise, if a data provided by a source is distanced from the resulted aggregate,a low weightwill be assigned. These concepts exactly shadow the general principle followed in truth discovery.

4.3 Probabilistic graphical model-based method

Based on Probabilistic Graphical Model there are few truth discovery methods available. Bayesian probabilistic method help address the problems in truth finding on numerical data. This model characterises numerical data to infer the true value and source quality without supervision [34, 32,31]. The general procedure of PGM method is as follows,

$$\prod_{s\in S} p(w_s|\beta)\prod_{o\in O}(p(v_o^*|\alpha)\prod_{s\in S} p(v_o^s|v_o^*, w_s)) \quad -4$$

Here α, β – represents prior knowledge about truth distribution. v_o^* – denote the truth and it is set as the distribution mean.w_s – denote source weight, the reciprocal of variance.v_o^s - denote claimed value that is sampled form the parameters (v_o^*, w_s) from a distribution.$p(v_o^s|v_o^*, w_s)$is a function that links them together.Expectation maximization (EM) technique is adopted to infer values for latent variables $\{w_s\}$ and $\{v_o^*\}$ in order to maximize the likelihood. If source weight (w_s) is high, then claimed value v_o^s is close to truthv_o^*.

The three above mentioned methods are framed to capture the truth in truth discovery.Iterative method is simple, easy to understand and interpret. Optimization method and probabilistic method derived from inference is slight complex and require more explanation. All three methods are popular and widely used in Truth Discovery.

5. OTHER TRUTH DISCOVERY METHODS

A brief description of the factors namely the data being input,reliability of the source, the event or object, the value claimedand the outputneeds to be considered during Truth

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

Discovery.Other available approaches are presented. The approaches are compared under different scenarios and the result is shown in Table 2 & 3. This comparison guide users and developers in selecting the appropriate truth discovery methods and use it for specific scenarios based on the application.

5.1 Factors of Truth Discovery

Factors like the data that is input, the nature of source, the object, inferred truth, application from various domain, etc... differ based on the situation. Hence different truth discovery approaches are proposed for different scenarios.

5.1.1 Input data

A basic analysis on the features of input data is necessary before proceeding to the truth discovery step. Pre-processing of input data improves the effectiveness of truth discovery methods [1]. Several features of input data are.

Consider the time stamp for observation to identify duplicate input data [45].Identify the objects with conflict from among the information [32].Data that mean the same may be available in different formats. Identify and format it to an identical value [28].Consider the uncertainty in information extractors [44].Focus on unstructured data, together with structured data as the above provides useful information for estimating source reliability.Choose truth discovery approach that could also support streaming data apart from static data.Select a truth discovery approach that could work on unsupervised situation that is discover truth even in the absence of additional labelled information.

5.1.2 Source reliability

Estimating the reliability of the information provider (source) is an important feature of truth discovery [1]. An assumption about source consistency can be followed as, a source is likely to provide true information (more often) with the same probability for all the events. [4, 6, 36,29, 43, 30, 32]A source independency assumption help identify if the observation made by the source is independent, which infer the fact that true information or data is more likely to be identical and false information is more likely to be different. [6, 36, 42, 43, 32]A source dependency analysis help detect copied information, which helps evaluate the quality of source [10, 39, 40, 41, 42] and the correlation between sources.

5.1.3 Object

This module brief about object difficulty and relation among events or objects, that affect truth discovery [1].By considering the object difficulty 3-estimate algorithm is proposed. The errors introduced by objects and sources are separated and the sources reliability can be better estimated. Capturing the relation among objects and prior knowledge about the objects could improve performance of truth discovery [38, 29].

5.1.4 Claimed value

This section discusses about the claimed values to be considered for truth discovery approaches [1].The complementary vote technique is used to infer additional information from the claimed value. If a source provides a value about an object [6, 31] it is assumed that this source votes against other candidate values of this object. An implication of values can be used to capture the distance of continuous values. Truth discovery method also focuses on the data type of claimed values.

5.1.5 Output

The following points are to be analysed for consideration of the outputs of truth discovery [1].To identify an assumption, like single, multiple or unknown truthfor identified truths.Motivate the use of scoring techniques to assign a score to each claimed value usually in the form of probability.Adopt a performance metric to evaluate the effectiveness of the output derived from truth discovery methods.

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

5.2 Other Truth Discovery Methods

The above section discussed several truth discovery methods from five aspects namely the data being input, reliability of the source, the event or object, the value claimed and the output.This section briefly describesvarious other truth discovery methods.

Truth finder – it is based on Bayesian analysis technique. Trust worthiness of information is derived by iteratively estimating source reliability [27].AccSim – the technique followed is Bayesian analysis. An implication function is proposed to infer the similarity of claimed values [4, 28].AccuCopy – this is an improvement AccSim and detects the copying relation among sources. The weight of source provider is reduced if found that the information is a copier of other sources [4, 28].2- Estimates – this method follows "single truth assumption" based on "there is one and only one true value for each object" and this method adopts complimentary vote [6].3-Estimate – this method estimates the truth, source reliability and the difficulty factor by introducing the trust worthiness of claim values for getting the truth of each objects [6].Investment – in this method the source equally invest reliability to all claimed values, the confidence develops on a linear function defined on the sum of invested reliabilities from its providers [29].

SSTF – it follows semi-supervised truth discovery method, a label are score is assigned to each claimed value and to infer the relationship among claims mutual exclusion property are used [30].LTM – it is a probabilistic graphical model LTM, it helps infer multiple truths. Two factors precision and recall are included for determining multiple truths. LTM considers both false positive and false negative claims in determining multiple truth simultaneously [31].GTM – it follows Bayesian probabilistic approach it is proposed for working on continues data type, here mutual supportive relations are modelled [32].Regular EM –a maximum likelihood estimation problem proposed for solving crowd social sensing applications. Humans act as sensors and the problem is solved using EM algorithm [14]. LCA– a set of latent parameters α, β provide additional information on source reliability to end users [34].

Apollo – social– this method best suits the detection of source dependencies as claims can also be a re-tweet information [35].CRH – it is designed to handle heterogonies data. A distance function helps calculate the nature of varied data types and estimation of source reliability on all data types together [36].CATD – for certain claim, source may provide very few observations. CATD is best proposed for handling this phenomenon [7].

6. RESULT AND DISCUSSION

A comparative study on several Truth Discovery approaches under the above-mentioned factors are presented in table 2 and table 3 [1]. For categorical data most of the Truth Discovery work good. GTM, CATD, TF, ACCSUM & ACCUCOPY work good for continuous data by using implication function. SSTF and CRH are designed to deal with heterogeneous data. [16, 20, 33, 41] LTM, FM, LCA and CATD work about source dependency analysis. 3-esimate (during analysis) considers object difficulty. Most Truth Discovery method [19, 24, 31] to strengthen the assumption adopt complementary voting in case of unknown Truth Discovery LTM is proposed to handle the scenarios.

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

Table 2: Comparison of Truth Discovery Methods: Part 1 [1]

	Input Data				Source Reliability	
	Categorical	Continuous	Heterogeneous	Label Truth	Source Dependency	Enriched Meaning
Truth Finder	Yes	Yes				
AccuSim	Yes	Yes				
AccuCopy	Yes	Yes			Yes	
2-Estimates	Yes					
3-Estimates	Yes					
Investment	Yes					
SSTF	Yes	Yes	Yes	Yes		
LTM	Yes					Yes
CTM		Yes				
Regular EM	Yes					Yes
LCA	Yes					Yes
Apollo-Social	Yes				Yes	Yes
CRH	Yes	Yes	Yes			
CATD		Yes				Yes

Table 3: Comparison of Truth Discovery Methods: Part 2 [1]

	Object		Claimed Values	Output	
	Object Difficulty	Object Relation	Complementary Vote	Multiple Truth	Unknown Truth
Truth Finder					
AccuSim			Yes		
AccuCopy			Yes		
2-Estimates			Yes		
3-Estimates	Yes		Yes		
Investment		Yes			Yes
SSTF			Yes		
LTM				Yes	
CTM					
Regular EM					
LCA					
Apollo-Social					
CRH					
CATD					

The table clearly projects that SSTF can handle categorical, continuous and heterogeneous data. The difficulty level of an object also plays a key role in the detection of truth, among the truth discovery methods in the table, 3-estimate is the only approach to consider the objects difficulty

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

level, an advanced feature of the object. From the table it is clear that most of the truth discovery methods make use of complementary voting concept for truth assumptions of the claimed value.

7. CONCLUSION& FUTURE WORK

Increased use of social media network makes it necessary to identify truth from among the conflicts from multi-source data. This survey described the general principle followed in truth discovery approaches, stated an overall view of the improvements on this research topic, the five elementary aspects that need to be considered on examining the truth, a brief of the existing truth discovery approaches which have different assumptions on the input data, limitations and the output are clearly summarized, a few related research work carried on this topic of truth finding by different authors on certain similar related issues are also clearly stated .Additional efforts are greatly in demand to discover the relation among objects which will significantly benefit the real world application such as knowledge graph constructions. Furthermore, efficiency issue relating to the identification of truth discovery on large scale data is critical. Also validating the identified truths is a big challenge due to the fact that only limited ground truth values are available in practice.

For Truth Discovery task various methods have been proposed but still exists problem that needs to the explored. The challenges in source weight estimate for constructed data needs consideration. Due to involvement of scale of objects automatic detection of relationship among objects need to be addressed as in general it is assumed that objects are independent which is not so always. Assigning of uniform weight during initialization is observed to be disadvantages which need more investigation. For performance evaluation Truth Discovery approaches assume ground truth value is available but it is good to construct the ground truth instead of assumption, is an area that needs more attention. The credibility level of Truth Discovery approaches improves on addressing the challenges.

REFERENCES

[1] Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., ... & Han, J. (2016). A survey on truth discovery. *ACM Sigkdd Explorations Newsletter*, *17*(2), 1-16.

[2] Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, *2*(1), 550-561.

[3] Aydin, B. I., Yilmaz, Y. S., Li, Y., Li, Q., Gao, J., &Demirbas, M. (2014, June). Crowdsourcing for multiple-choice question answering. In *Twenty-Sixth IAAI Conference*.

[4] Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Integrating conflicting data: the role of source dependence. *Proceedings of the VLDB Endowment*, *2*(1), 550-561.

[5] Dong, X. L., & Srivastava, D. (2013, May). Compact explanation of data fusion decisions. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 379-390).

[6] Galland, A., Abiteboul, S., Marian, A., &Senellart, P. (2010, February). Corroborating information from disagreeing views. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 131-140).

[7] Li, Q., Li, Y., Gao, J., Su, L., Zhao, B., Demirbas, M., ... & Han, J. (2014). A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment*, *8*(4), 425-436.

[8] Le, H., Wang, D., Ahmadi, H., Uddin, Y. S., Szymanski, B., Ganti, R., &Abdelzaher, T. (2011, November). Distilling likely truth from noisy streaming data with apollo. In *Proceedings of the 9th ACM Conference on Embedded Networked Sensor Systems* (pp. 417-418).

[9] Mukherjee, S., Weikum, G., &Danescu-Niculescu-Mizil, C. (2014, August). People on drugs: credibility of user statements in health communities.

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 65-74).

[10] Dong, X. L., Berti-Equille, L., Hu, Y., & Srivastava, D. (2010). Global detection of complex copying relationships between sources. *Proceedings of the VLDB Endowment*, *3*(1-2), 1358-1369.

[11] Miao, C., Jiang, W., Su, L., Li, Y., Guo, S., Qin, Z., ... & Ren, K. (2015, November). Cloud-enabled privacy-preserving truth discovery in crowd sensing systems. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems* (pp. 183-196).

[12] Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., ... & Zhang, W. (2014, August). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 601-610).

[13] Dong, X. L., Gabrilovich, E., Heitz, G., Horn, W., Murphy, K., Sun, S., & Zhang, W. (2015). From data fusion to knowledge fusion. *arXiv preprint arXiv:1503.00302*.

[14] Wang, D., Kaplan, L., Le, H., &Abdelzaher, T. (2012, April). On truth discovery in social sensing: A maximum likelihood estimation approach. In *Proceedings of the 11th international conference on Information Processing in Sensor Networks* (pp. 233-244).

[15] Wang, S., Su, L., Li, S., Hu, S., Amin, T., Wang, H., ... &Abdelzaher, T. (2015, April). Scalable social sensing of interdependent phenomena. In *Proceedings of the 14th International Conference on Information Processing in Sensor Networks* (pp. 202-213).

[16] Li, H., Zhao, B., &Fuxman, A. (2014, April). The wisdom of minority: Discovering and targeting the right group of workers for crowdsourcing. In *Proceedings of the 23rd international conference on World wide web* (pp. 165-176).

[17] Whitehill, J., Wu, T. F., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in neural information processing systems* (pp. 2035-2043).

[18] Dong, X. L., Saha, B., & Srivastava, D. (2012). Less is more: Selecting sources wisely for integration. *Proceedings of the VLDB Endowment*, *6*(2), 37-48.

[19] Dong, X. L., & Naumann, F. (2009). Data fusion: resolving data conflicts for integration. *Proceedings of the VLDB Endowment*, *2*(2), 1654-1655.

[20] Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001, April). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web* (pp. 613-622).

[21] Yin, X., & Tan, W. (2011, March). Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web* (pp. 217-226).

[22] Wang, D., Kaplan, L., &Abdelzaher, T. F. (2014). Maximum likelihood analysis of conflicting observations in social sensing. *ACM Transactions on Sensor Networks (ToSN)*, *10*(2), 1-27.

[23] Yin, X., Han, J., & Philip, S. Y. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796-808.

[24] Xu, X., Liu, X., Liu, X., & Sun, Y. (2017). Truth finder algorithm based on entity attributes for data conflict solution. *Journal of Systems Engineering and Electronics*, *28*(3), 617-626.

[25] Zhang, D. Y., Wang, D., & Zhang, Y. (2017, December). Constraint-aware dynamic truth discovery in big data social media sensing. In *2017 IEEE International*

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

Conference on Big Data (Big Data) (pp. 57-66). IEEE.

[26] Xiao, H., Gao, J., Li, Q., Ma, F., Su, L., Feng, Y., & Zhang, A. (2016, August). Towards confidence in the truth: A bootstrapping-based truth discovery approach. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1935-1944).

[27] Yin, X., Han, J., & Yu, P. S. (2007). Truth discovery with multiple conflicting information providers on the web: Proc. of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'07).

[28] Li, X., Dong, X. L., Lyons, K., Meng, W., & Srivastava, D. (2015). Truth finding on the deep web: Is the problem solved?. *arXiv preprint arXiv:1503.00303*.

[29] Pasternack, J., & Roth, D. (2010, August). Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics* (pp. 877-885). Association for Computational Linguistics.

[30] Yin, X., & Tan, W. (2011, March). Semi-supervised truth discovery. In *Proceedings of the 20th international conference on World wide web* (pp. 217-226).

[31] Zhao, B., Rubinstein, B. I., Gemmell, J., & Han, J. (2012). A bayesian approach to discovering truth from conflicting sources for data integration. *arXiv preprint arXiv:1203.0058*.

[32] Zhao, B., & Han, J. (2012). A probabilistic model for estimating real-valued truth from conflicting sources. *Proc. of QDB*.

[33] Dong, X. L., Gabrilovich, E., Murphy, K., Dang, V., Horn, W., Lugaresi, C., ... & Zhang, W. (2015). Knowledge-based trust: Estimating the trustworthiness of web sources. *arXiv preprint arXiv:1502.03519*.

[34] Pasternack, J., & Roth, D. (2013, May). Latent credibility analysis. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1009-1020).

[35] Wang, D., Amin, M. T., Li, S., Abdelzaher, T., Kaplan, L., Gu, S., ... & Wang, X. (2014, April). Using humans as sensors: an estimation-theoretic perspective. In *IPSN-14 Proceedings of the 13th International Symposium on Information Processing in Sensor Networks* (pp. 35-46). IEEE.

[36] Li, Q., Li, Y., Gao, J., Zhao, B., Fan, W., & Han, J. (2014, June). Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 1187-1198).

[37] Yerva, S. R., Jeung, H., &Aberer, K. (2012, July). Cloud based social and sensor data fusion. In *2012 15th International Conference on Information Fusion* (pp. 2494-2501). IEEE.

[38] Meng, C., Jiang, W., Li, Y., Gao, J., Su, L., Ding, H., & Cheng, Y. (2015, November). Truth discovery on crowd sensing of correlated entities. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems* (pp. 169-182).

[39] Pochampally, R., Das Sarma, A., Dong, X. L., Meliou, A., & Srivastava, D. (2014, June). Fusing data with correlations. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 433-444).

[40] Qi, G. J., Aggarwal, C. C., Han, J., & Huang, T. (2013, May). Mining collective intelligence in diverse groups. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 1041-1052).

[41] Sarma, A. D., Dong, X. L., & Halevy, A. (2011, March). Data integration with dependent sources. In *Proceedings of the 14th International Conference on*

January - February 2020
ISSN: 0193 - 4120 Page No. 16573 - 16584

Extending Database Technology (pp. 401-412).

[42] Dong, X. L., Berti-Equille, L., & Srivastava, D. (2009). Truth discovery and copying detection in a dynamic world. *Proceedings of the VLDB Endowment*, *2*(1), 562-573.

[43] Yin, X., Han, J., & Philip, S. Y. (2008). Truth discovery with multiple conflicting information providers on the web. *IEEE Transactions on Knowledge and Data Engineering*, *20*(6), 796-808.

[44] Pasternack, J., & Roth, D. (2011, June). Making better informed trust decisions with generalized fact-finding. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[45] Rekatsinas, T., Dong, X. L., & Srivastava, D. (2014, June). Characterizing and selecting fresh data sources. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 919-930).

[46] Li, Y., Li, Q., Gao, J., Su, L., Zhao, B., Fan, W., & Han, J. (2015, August). On the discovery of evolving truth. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 675-684).

[47] Marshall, J., & Wang, D. (2016, September). Mood-sensitive truth discovery for reliable recommendation systems in social sensing. In *Proceedings of the 10th ACM Conference on Recommender Systems* (pp. 167-174).

[48] Li, F., Lee, M. L., & Hsu, W. (2014, August). Entity profiling with varying source reliabilities. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1146-1155).