ISSN: 2278 – 7798
International Journal of Science, Engineering and Technology Research (IJSETR)
Volume 5, Issue 8, August 2016

REVIEWING TRUTH DISCOVERY APPROACHES AND METHODS FOR BIG DATA INTEGRATION

Arunima Kumari
Dept. of computer Science and Engineering
Mtech Student
DCRUST Murthal
Sonepat, India

Dr. Dinesh Singh
Dept. of computer Science and Engineering
Assistant Professo r
DCRUST Murthal
Sonepat, India

Abstract—
Truth-finding is the fundamental technique for affirming reports from multiple sources in both data integration and corporative intelligent applications. Traditional truth finding methods presume a single true value for each data item and hence cannot deal will multiple true values i.e., the multi-truth-finding problems. So far, the existing approaches handle all the problems whether it is multi-truth-finding problem or the single-truth-finding problem in the same way. The multi-truth-finding problem is having its unique features, such as the involvement of sets of qualities in cases, diverse ramifications of between worth mutual exclusion, and larger source profiles. With consideration of these features we could provide new opportunities for obtaining more accurate truth finding results. Based on this insight, integrating data from several origins has been increasingly becoming a commonplace in both Web requests to support collective intellect and cooperative decision making. Unfortunately, it is not infrequent that the data concerning a solitary item comes from different origins that could be loud, out-of-date, or even erroneous. It is consequently of paramount significance to ascertain such fights amid the data and to find out that piece of data which is extra reliable. While the single-truth-finding setback (STF)—which aims at discovering the solitary real worth for an item—has been extensively learned, an extra finished case, whereas several real benefits (or multi-truth) could continue for a solitary item, is scarcely explored. In this paper we survey various Multi Truth Discovery for data integration processes

Keywords: *Truth Discovery, Data Integration, Single Truth Discovery (STF), Multi Truth Discovery (MTF).*

I. INTRODUCTION

Data integration is a permeate challenge faced in applications that need to interrogate across multiple autonomous and heterogeneous data sources. Data integration is all-important in large enterprises that possess a large number of data sources, for furtherance in large-scale scientific projects, where multiple researchers produce data sets independently, for better and improved practice among various government agencies, each with their own data sources, and in providing good search quality throughout the millions of data sources that are structured on the World-Wide Web basis.

II. DATA INTEGRATION CHALLENGES

Several fundamental factors ensure that data integration challenges will continue to engross our community for a long time to come. The first factor is social. Data integration is basically about getting people to cooperate and share data. It involves retrieving the appropriate data, convincing people to share it and proposing them an incentive to do so (either in terms of ease of sharing or benefits from the resulting applications), and convincing data owners about their concerns for data sharing (e.g., privacy, effects on the performance of their systems) will be addressed.

The second factor, which needs to be highlighted, is complexity of integration. When we talk in terms of many application contexts it is not even clear what it actually means to integrate data or how the combined sets of data can be operated together. Consider an example; the merger of two companies and therefore to handle their different stock option packages it's a need for a single system. What do stock options in one company even mean in the context of a merged company? While this example seems like a business question (and it is), it illustrates the demands that may be imposed on the data management systems to accommodate such unexpected complexity.

Because of all these reasons, data integration has been looked up to as a problem as Artificial Intelligence, maybe even harder than that! When we work as a community, our goal should focus on to create tools that facilitate data integration in different scenarios. By addressing the following specific challenges could lead towards that goal.

DATASPACES: PAY-AS-YOU-GO DATA MANAGEMENT

Long setup time required is one of the fundamental shortcomings of both the database and data integration

systems. In a database system, before we receive any services or get any benefit, one must be needed to first create a schema and then populate the database with tuples. In a data integration system, one needs to create the semantic mappings to retrieve any visibility into the various data sources. The management of dataspaces accentuates the idea of pay-as-you-go data management: offering some services right away without any setup time required, and meliorates the services as more investment is being made into creating semantic relationships. In other terms we can say, a dataspace should not only offer keyword search over any data in any source but also with no setup time required. Progressing further, we can extract associations between heterogeneous data items in a dataspace with the use of a set of heuristic extractors, and further querying those associations with path queries. Finally, when we have decided that we actually needed a tighter integration between a pair of data sources, we needed to create a mapping automatically and ask the human to further modify and validate it.

Uncertainty and lineage:

Research on finagling uncertain data and data lineage has a long run story in our community. As in traditional database management that manages uncertainty and lineage looks like a skillful feature, in data integration it becomes a requirement. It is obvious in nature that data from multiple sources will be uncertain and even inconsistent with each other. The systems must be able to ponder about the certainty of the data, and when there are chances that they cannot automatically determine its certainty, reference of the user to the lineage of the data so they could determine for themselves which source is more reliable. Pervading data integration systems introspection abilities will extend their applicability and their ability for dealing with diverse data integration settings. A recent line of work in the community is about to start to address these issues.

Reusing human attention:

The ability to reuse human attention is one of the principles for achieving tighter semantic integration among different data sources. In simple terms, every time whenever a human interacts with a dataspace, they are indirectly being given by a semantic clue about the data or about relationships among data sources. Examples of such clues are obtained whenever a user query about data sources (even in group), whenever users create semantic mappings or when they cut and paste some sort of data from one place to another. We can obtain semantic integration much faster if we can build systems that leverage these semantic clues. We already have confronted with few examples where reusing human attention has been very successful, but this is an area that is very advanced for additional research and development in it. In some cases we can take advantage of work that users are doing as a part of their job, in others we can come up to for some help by asking some well-chosen questions, and in others we just simply exploit structure that already exists such as schemas in large number or web service descriptions.

III. TRUTH DISCOVERY

In the era of information explosion, data have been wandered into every aspect of our lives, and we are continuously generating data through a variety of channels, such as social networks, blogs, discussion forums, crowdsourcing platforms, etc. These data are analyzed at both individual and population levels, by business for aggregating opinions and recommending valuable products, by governments for decision making and security checking, and by researchers for discovering new knowledgeable ideas . In these scenarios, data, even describing the same object or event, can come from a variety of sources. However, the gathered information about the same object from various sources may conflict with each other due to errors, missing records, typos, faults, misprinted data, out-of-date data, etc. For example, the top search results given by Google for the query like “the height of Mount Everest" include “29; 035 feet", “29; 002 feet" and “29; 029 feet". Among these assembles of noisy information, which one is more trustworthy, or which represents the true fact? In this and many more similar problems, it is important to combine and collect noisy information about the same set of objects or events gathered from various sources to get true and accurate facts.

One straightforward approach for eliminating conflicts among multi-source data is to conduct majority voting or averaging. The biggest shortcoming of such approaches is that they assume all the sources as equally reliable. Unfortunately, this supposal may not hold in most of the cases. With the generalization of the aforementioned \Mount Everest" example: Using majority voting, the result “29; 035 feet", which has the highest number of occurrences, will be considered as the truth. However, in the search results obtained, the information “29; 029 feet" from Wikipedia source is the truth. This example reveals about information quality that varies a lot among different sources, and aggregated results in respect of accuracy can be further improved by capturing the reliabilities of various related sources. The challenge is that source reliability is commonly unknown a priori in practice and has to be defined from the data.

With the view of this challenge, the topic of truth discovery has gained much popularity in the recent days due to its ability to estimate source reliability degrees and infer true information. As the truth discovery methods usually work without any supervision, the source reliability can only be generalized based on the data given. Thus in existing work, the source reliability estimation and truthfinding steps are combined through the following principle: The sources providing true information more often will be assigned with higher reliability degrees, and the information which is supported by reliable sources will be considered as truths.

With this well-known common principle, truth discovery approaches which have been proposed to various scenarios nowadays, and these approaches make different assumptions about input data, source relations, identified truths, etc. Due to this vast diversity, it may not be easy for people to compare and choose an appropriate approach among various approaches one for their tasks.

Truth discovery has an outstanding part in information age. On one hand we need accurate and exact information more than ever needed, but on the other hand inconsistent information is ineluctable due to the most common feature of big data – "variety". The development of truth discovery can benefit many applications in different fields where critical decisions are being to be made based on the reliable information extracted from various sources. Examples include healthcare, crowd/social sensing, crowdsourcing information extraction, knowledge graph construction and so on. These in addition with other applications demonstrate the broader impact of on multi-source information integration through the truth discovery.

IV. TRUTH DISCOVERY CHALLENGES

Duplicate input data

It is possible that one source may make several observations about the same object. For example, a Wikipedia contributor may edit the information about the same entry several times, or a crowdsourcing worker can submit his output for a specific task multiple attempts. However, most of the truth discovery methods assume that each source makes at most one observation about an object. If the timestamp for each observation is available, a possible approach is to consider the data freshness and select the up-to-date observation. Otherwise, some pre-defined rules can be adopted to choose one from multiple observations.

Objects without conflict

For some objects, all the observations made by sources have the same claimed value. In this case, most of the truth discovery methods should give the same results which is the claimed value (one exception is the method that considers "unknown" as an output candidate). So it might be safe to remove these trivial records. Furthermore, the authors report claims that this pre-processing improves the effectiveness of truth discovery methods. This is because of the fact that if all the sources agree with each other, these observations may not contribute (too much) to the source reliability estimation. It should be pointed out that these trivial records do affect the estimation of source reliability, and thus this pre-processing step should be carefully examined before performing it.

Input data format

As the information is collected from various sources, they may have different formats. For example, when the object is "the height of Mount Everest", some sources have claimed values as "29; 029 feet", while others have claimed values as "8848 meters". Another case, for example, "John Smith" and "Smith, John", is commonly observed in text data. In fact, these claimed values are the same one and they should be formatted to an identical value.

Input uncertainty

When the observations are extracted from textual data (for example, in the knowledge fusion task) or the sources provide observations with their confidence indicators (for example, in the question-answering system), it is necessary to consider the uncertainty of these observations. The authors propose a way to generalize truth discovery methods, which considers multi-dimensional uncertainty, such as the uncertainty in information extractors.

Structured vs unstructured data

Previously, most work considers the inputs from structured databases. Recently, increasingly more work focuses on unstructured input, such as texts. These unstructured data provide more information such as corpus evidence, URL, confidence, and question text which are useful for source reliability estimation. However, at the same time, this extra information introduces more noise and uncertainty.

Streaming data

In many real-world applications, data continue to arrive over time. Most of the existing truth discovery methods are batch algorithms and work on static data. These methods are inefficient to process streaming data as they need to re-run the batch algorithms when new data are available. To tackle this challenge, some truth discovery algorithms have been designed for different types of streaming data.

Labeled truths

Besides the input data, truth discovery methods might assume some additional labeled information. As labeled truths are usually difficult to collect, most truth discovery methods are unsupervised, i.e., estimating the truths without any labeled information. While in, the authors assume that a small set of truths are available and thus the proposed algorithms work in semi-supervised settings. Therefore, a few available truths can be used to guide source reliability estimation and truth computation. The results show that even a small set of labeled truths could improve the performance.

V. TRUTH DISCOVERY METHODS

We summarize various truth discovery methods from five aspects, namely, input data, source reliability, object, claimed value, and the output. Now, we briefly describe several representative truth discovery methods, and compare them under different features. By providing such a comparison, we hope to give some guidelines so that users and developers can choose an appropriate truth discovery method and apply it to the specific application scenarios. Due to space limitation, here we only describe each truth discovery algorithm briefly. For more details about these methods, the readers may refer to the reference papers.

- **TruthFinder:** In TruthFinder, Bayesian analysis is adopted to iteratively estimate source reliabilities and identify truths. The authors also propose the source consistency assumption and the concept of “implication", which are widely adopted in other truth discovery methods.
- **AccuSim:** AccuSim also applies Bayesian analysis. In order to capture the similarity of claimed values, the implication function is adopted.
- **AccuCopy:** This method improves AccuSim, and considers the copying relations among sources. The proposed method reduces the weight of a source if it is detected as a copier of other sources.
- **2-Estimates:** In this approach, the single truth assumption is explored. By assuming that there is one and only one true value for each object, this approach adopts complementary vote.
- **3-Estimates:** 3-Estimates augments 2-Estimates by considering the difficulty of getting the truth for each object.
- **Investment:** In this approach, a source uniformly “invests" its reliability among its claimed values, and the confidence of a claimed value grows according to a non-linear function defined on the sum of invested reliabilities from its providers. Then the sources collect credits back from the confidence of their claimed values.
- **SSTF:** In this semi-supervised truth discovery approach, a small set of labeled truths are incorporated to guide the source reliability estimation. Mean- while, both mutual exclusivity and mutual supports are adopted to capture the relations among claimed values.
- **LTM:** LTM is a probabilistic graphical model which considers two types of errors under the scenario of multiple truths: false positive and false negative. This enables LTM to break source reliability into two parameters, one for false positive error and the other for false negative error.
- **GTM:** GTM is a Bayesian probabilistic approach especially designed for solving truth discovery problems on continuous data.
- **Regular EM:** Regular EM is proposed for crowd/social sensing applications, in which the observations provided by humans can be modeled as binary variables. The truth discovery task is formulated as a maximum likelihood estimation problem, and solved by EM algorithm.
- **LCA:** In LCA approach, source reliability is modeled by a set of latent parameters, which can give more informative source reliabilities to end-users.
- **Apollo-social**: Apollo-social fuses the information from users on social media platforms such as Twitter. In social network, a claim made by a user can either be originally published by him or be re-tweeted from other users. Apollo-social models this phenomenon as source dependencies and incorporates such dependency information into the truth discovery procedure.
- **CRH:** CRH is a framework that deals with the het-erogeneity of data. In this framework, different types of distance functions can be plugged in, to capture the characteristics of different data types, and the estimation of source reliability is jointly performed across all the data types together.
- **CATD:** CATD is motivated by the phenomenon that many sources only provide very few observations. It is not reasonable to give a point estimator for source reliability. Thus in this confidence-aware truth discovery approach, the authors derive the confidence interval for the source reliability estimation.

VI. RELATED WORK

M Lamine Ba, et al [1]In this paper, they considered the case where there is an inherent structure in the arguments made by the sources about real-world objects, that involve different quality levels of a given source on different groups of attributes of an object. They do not assume this structuring which is being given, but instead of finding it automatically, they have found it by researching and weighting the

segmentation of the sets of attributes of an object, and then applying a reference truth finding algorithm on each subset of the optimal partition. Our experimental results on synthetic and real-world datasets show that they obtain better precision at truth finding than baselines in cases where data has an inherent structure. There are many interesting challenges in this problem for further development. First, they are experimenting with new scoring strategies and different greedy algorithms that construct an optimal partition starting from the set of singletons. The initial results show that they are more efficient in terms of total execution time with a result of near-optimal solution. Second, they aim at combining our partitioning approach with source selection methods in order to further leverage both the inherent structure of data and knowledge from domain experts.

F Zhang, et al [2]In this paper, they propose a modified method to find the most trustable source and identify the true information. Our goal is to minimize the distance between the true information and the overall observed descriptions through considering the accuracy and the coverage of all the data sources at the same time. The experiments on the real dataset demonstrate the efficacy of our method. In future work, they want to take other factors into consideration and these factors includes value similarity (how close of two values about same claim), time (nothing is absolutely right and truth is changing with time) and property correlation. Meanwhile, they want to implement the parallelization of the algorithm and run it on Big Data platform such as Spark to improve efficiency.

C Meng, et al [3] this assignment, alluded to as truth discovery, has as of late pulled in much consideration. Existing work typically presumes independence amongst entities. However, correlations among entities are usually observed in many applications. Such kind of correlation information is essential in truth discovery task. It is impossible to get true in formation when entities are not observed by adequate reliable users. In such cases, it is significantly important to disseminate trustworthy information from related correlated entities that have been discovered by reliable users. They devised the job of truth discovery on corresponded entities as an optimization issue in which both truths and user unwavering quality are mimicked as variables. The correlation among related entities contributes to the difficulty of figuring out this problem. In light of the challenge, they suggest both sequential and parallel results. In the sequential solution, they partition the entities into disjoint independent and autonomous sets and deduce iterative approaches grounded on block coordinate descent. In the parallel solution, they conform the solution to MapReduce programming model that can be accomplished on Hadoop clusters. Experiments and tests on real-world crowd sensing applications depict the advantages of the suggested method on discovering truths from convicting information described on correlated entities.

Y Li, et al [4]To address this problem, they investigate the temporal relations amongst both the object truths and source unwavering quality, and suggest an incremental truth disclosure framework that can dynamically modify object truths and source weights upon the entry of new information. Hypothetical analysis is provided to show that the proposed method is ensured to focalize at a quick rate. The tests on three real world applications and a set of synthetic data show the benefits of the proposed strategy over best in class truth discovery methods. We propose to discover truths from dynamic data, where the collected information comes sequentially, and both truths and the source reliability evolve over time. This is a challenging task since they have to come up with an efficient way to capture the temporal relations among the identified trustworthy information and source reliability. To address the efficiency issue, they propose an incremental method by studying the equivalence between optimization-based solution and MAP estimation.

S Zhi, et al [5] When incorporating information from multiple and heterogeneous sources, it is common to run into conflicting answers to the same question. Truth discovery is to deduce the most accurate and complete integrated answers from multiple conflicting sources. There exist some cases, where questions for which the true answers are omitted from the candidate answers given by all sources. Without any prior knowledge, these questions being named as no-truth questions are difficult to be discerned from the questions that are having true answers being named as has truth questions. Particularly, these questions which are named as no-truth questions degrade the exactness of the answer of integration system. To address such a challenge they have introduced source quality, which is being made up of three fine-grained measures: silent rate, false spoken rate and true spoken rate. By joining these three measures, they propose a probabilistic graphical model, which at the same time surmises truth and also source quality with no earlier preparing ground truth answers. Also, since deriving this graphical model requires parameter tuning of the prior of truth, they propose an introduction plan based upon an amount named truth existence score, which synthesizes two indicators, namely, interest rate and consistency rate. Contrasted and existing methods, our method can effectively filter out no-truth questions, which results in more exact source quality estimation. Consequently, our method provides more accurate also, finish answers to both has-truth and no-truth questions.

Experiments on three real-world datasets illustrate the also, finish answers to both has-truth and no-truth questions. truth discovery methods.

XL Dong, et al [6]The degree of excellence of web sources has been traditionally evaluated using exogenous signals such as the hyperlink structure of the graph. They propose a new approach that relies on endogenous signals, namely, the correctness of questions, which results in more exact source quality has few false facts is considered to be trustworthy. The realities are naturally removed from every source by data extraction methods commonly used to construct knowledge bases. They propose a way to differentiate errors made in the extraction procedure from factual errors in the web source in essence, by utilizing joint deduction as a part of a novel multi-layer probabilistic model. This paper proposes a new metric for evaluating web-source quality– knowledge-based trust. They proposed a sophisticated probabilistic model that jointly gauges the accuracy of extractions and source information, and the trustworthiness of sources. In addition, they presented a calculation that progressively chooses the level of granularity for each source. Experimental results have shown both promise in assessing web source quality and change over existing techniques for knowledge fusion.

Fenglong Ma, et al [7]The most significant challenge for this task is to figure out source reliability and select answers that are given by superb sources. Existing work figure out this problem by estimating source's unwavering quality and collecing inquiry's actual answers (i.e., the truths). However, these methods presume that a source has the same unwavering quality degree on all the inquiries, however ignore the fact that sources' reliability may vary significantly among various themes. To catch different aptitude levels on different topics, they propose FaitCrowd, a fine grained truth disclosure model for the assignment of collecting clashing data collected from multiple users/sources. FaitCrowd jointly models the procedure of producing inquiry substance and sources' provided answers in a probabilistic model to estimate both topical skill and genuine answers simultaneously. In this paper, they propose a new probabilistic Bayesian model to deal with the challenge of deducing fine grained source reliability. By collectively modeling question content and collected answers, the proposed model learns the topics of questions, topic-specific expertise of sources, and the true answers simultaneously. Experimental results on two real crowdsourced datasets show the potency of the proposed FaitCrowd model. They demonstrate that FaitCrowd can successfully detect the true answers from the expert sources on the corresponding topics even when their answers are minority in the answer set. Analysis shows that the learned topical expertise for sources is con-sistent with the genuine topical aptitude.

Chenglin Miao, et al [8]In this paper, they propose a novel cloud-empowered security protecting truth revelation (PPTD) framework for crowd sensing systems, which can achieve the insurance of clients' tangible information as well as their reliability scores derived by the truth discovery approaches. The key thought of the proposed structure is to perform weighted aggregation on users' encrypted data using homomorphic cryptosystem. In order to deal with large-scale data, they also propose to parallelize PPTD with MapReduce framework. Through extensive probes engineered information as well as certifiable crowd sensing systems, they justify the guarantee of strong security and high exactness of our proposed outline work. In this paper, they design a cloud-enabled privacy- preserving truth discovery (PPTD) framework to tackle the issue of privacy protection in crowd sensing systems. The key idea of PPTD is to perform weighted aggregation on the encrypted data of users using homomorphic cryptosystem, and iteratively conduct two phases (i.e., secure weight up- date and secure truth estimation) until convergence. During this procedure, both user's observation values and his reliability score are protected. In order to process large-scale data efficiently, a parallelized extension of PPTD is also pro- posed based on the MapReduce framework.

Xin Luna Dong, et al [9] BDI contrasts from customary information incorporation in numerous measurements: (i) the number of data sources, even for a single domain, has become in the several thousands, (ii) a number of the data sources are very dynamic, as a huge amount of newly gathered information are persistently made accessible, (iii) the data sources are extremely heterogeneous in their structure, with significant assortment notwithstanding for generously comparable entities, and (iv) the data sources are of widely disagreeing qualities, with huge contrasts in the scope, accuracy and timeliness of data provided. This tutorial investigates the advancement that has been made by the information joining community on the topics of schema mapping, record linkage what's more, information combination in tending to these novel difficulties confronted by big data integration, and identifies a range of open problems for the community. This tutorial reviews state-of-the-art methods for data integration in addressing the challenges raised by Big Data: volume and number of sources, speed, variety, and veracity. They discuss how close they are

to meeting these difficulties and recognize numerous open issues for future research.

X Wang, et al [10] regrettably, the multi-truth-finding problem has its unique features, such as the involvement of sets of qualities in cases, diverse ramifications between worth mutual exclusion, and larger source profiles. Considering these elements could give new chances to acquiring more accurate truthfinding results. Based on this insight, they suggest an integrated Bayesian approach to the multi-truth-finding problem, by taking these features into consideration. To improve the truth-finding ratio, they reformulate the multitruthfinding problem model based on the mappings between sources and (different sets of) qualities. New shared selective relations are characterized to reflect the possible co-existence of multiple true values. A better grained duplicate identification strategy is likewise proposed to deal with sources with large profiles. The experimental results on three certifiable datasets demonstrate the adequacy of our approach. We propose an integrated Bayesian approach, which comprehensively incorporates novel methods on three key aspects that characterize the multi-truth-finding problem (MTF), namely source value mapping, mutual exclusive relation, and source dependency, to better solve the problem.

Y Li, et al [11]Y Li Various truth discovery methods have been proposed for various scenarios, and they have been successfully connected in different application spaces. In this review, they focus on providing a comprehensive overview of truth revelation strategies, and compressing them from various angles. They also discuss some future directions of truth discovery research. They trust that this overview will professional bit a superior understanding of the current progress on truth discovery, and order some road maps on how to apply these approaches in application domains. As the existing truth discovery approaches have different assumptions about input data, constraints, and the output, they have been clearly compared. When choosing a truth discovery approach for a particular task, users and developers can refer to this comparison as guidelines. They have also discussed some future directions of truth discovery research. More efforts are highly in demand to explore the relations among objects, which will greatly benefit the real-world applications such as knowledge graph construction. Furthermore, efficiency issue becomes a bottle- neck for the deployment of truth discovery on large-scale data. Besides, how to evaluate the performance or validate the identified truths is a big challenge because of the way that restricted groundtruth is accessible in practice.

DA Waguih, et al [12] They provide reference implementations and an in-depth evaluation of the methods based on extensive experiments on synthetic and real-world data. They analyze aspects of the problem that have not been explicitly studied some time recently, for example, the effect of introduction and parameter setting, convergence, and scalability. They provide an experimental structure for broadly contrasting the strategies in a wide range of truth discovery scenarios where source coverage, numbers and appropriations of contentions, and genuine positive claims can be controlled and used to evaluate the quality and performance of the algorithms. Finally, they report comprehensive findings obtained from the experiments and provide new insights for future research. Future work consists of extending this work in a number of fronts. Firstly, they hope that our synthetic data set generation framework can be used and extended for parameter setting, testing and inside and out assessment of other existing or new calculations in a variety of truth discovery scenarios (e.g., with controlling source dependence and value similarity). The main advantage of the framework proposed is to control a complete ground truth (usually difficult to get with genuine information sets) and copy true truth discovery scenarios. Secondly, they can see many challenging research parkways for the up and coming era of truth revelation methods: (1) To improve scalability on the number of sources to be applicable to data from social networks and social media, (2) To improve the algorithm's precision for pessimistic situations when a large portion of sources are not dependable and have few conflicting values,
(3) To improve the usability and repeatability of the calculations, either by rearranging the parameterization or combining multiple methods to find optimal parameter setting.

D Yu. Et al [13] Information Extraction using various information sources and systems is beneficial because of multisource/framework union and testing due to the resulting inconsistency and redundancy. They coordinate IE and truth-discovering research and present a new unsupervised multi-dimensional truth finding framework which fuses signals from various sources, different frameworks and multiple pieces of evidence by knowledge graph construction through multi-layer profound etymological examination. Tests on the case study of Slot Filling Validation demonstrate that our methodology can discover truths precisely (9.4% higher F-score than supervised methods) and efficiently (finding 90% truths with one and only a large portion of the expense of a benchmark without credibility estimation).In this paper they leverage the strengths of these two distinct, but complementary research paradigms and propose a novel unsupervised multi-dimensional truthfinding framework

incorporating signals both from multiple sources, multiple systems and multiple evidences based on knowledge graph construction with multi-layer linguistic analysis. Experiments on a challenging SFV task demonstrated that this framework can find high-quality truths efficiently. In the future they will focus on exploring more inter-dependencies among reactions such as temporal and causal relations.

VII. Conclusion and Future Scope

Coordinating information from various sources has been progressively becoming a commonplace in both Web applications to support aggregate knowledge and collective basic leadership. Unfortunately, it is not unusual that the information about a solitary thing originates from various sources, which may be noisy, out-of-date, or even erroneous. It is therefore of foremost significance to determine such clashes among the data and to find out which piece of information is more reliable. While the single-truth-finding problem (STF)— which aims at finding the single true value for an item—has been widely studied, a more universal case, where multiple true values (or multi-truth) might exist for a single item, is rarely explored. In fact, multi-truth scenarios commonly exist in our real lives. For example, a book is usually authored by several people; a conference may have several deadlines; and the presidents of the United States involve a long list of names. For future, we want to investigate the truth existence problem of truth discovery and propose a novel probabilistic model to incorporate these measures as sources generating the answer set given true answers. Also, we proposed effective search approaches for truth finding using Bayesian networks. Extensive experiments on three real-world datasets will be used to clearly show that proposed model outperforms state-of-the-art truth discovery approaches. Interesting future work includes solving the truth existence problem when the independence assumption between sources does not hold. When two sources are dependent, the answers they agree with should be discounted. The source dependence will affect the initialization of truth prior as well. In Future Work, we propose an integrated Bayesian approach to address the above challenges. We will utilize Tabu Search for improving the search quality of the Bayesian methods. We will work an effective Bayesian Network Combined with Tabu Search for the problem model for multitruth discovery based on the relations between sources and values, and present corresponding methods for grouping sources and values to enable the truth discovery.

VIII. References

[1]. Lamine Ba, M., Roxana Horincar, Pierre Senellart, and Huayu Wu. "Truth Finding with Attribute Partitioning." In Proceedings of the 18th International Workshop on Web and Databases, pp. 27-33. ACM, 2015.

[2]. Zhang, Fan, Li Yu, Xiangrui Cai, Ying Zhang, and Haiwei Zhang. "Truth Finding from Multiple Data Sources by Source Confidence Estimation." In 2015 12th Web Information System and Application Conference (WISA), pp. 153-156. IEEE, 2015.

[3]. Meng, Chuishi, Wenjun Jiang, Yaliang Li, Jing Gao, Lu Su, Hu Ding, and Yun Cheng. "Truth discovery on crowd sensing of correlated entities." In Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems, pp. 169-182. ACM, 2015.

[4]. Li, Yaliang, Qi Li, Jing Gao, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. "On the discovery of evolving truth." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 675-684. ACM, 2015.

[5]. Zhi, Shi, Bo Zhao, Wenzhu Tong, Jing Gao, Dian Yu, Heng Ji, and Jiawei Han. "Modeling truth existence in truth discovery." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1543-1552. ACM, 2015.

[6]. Dong, Xin Luna, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. "Knowledge-based trust: Estimating the trustworthiness of web sources." Proceedings of the VLDB Endowment 8, no. 9 (2015): 938-949.

[7]. Ma, Fenglong, Yaliang Li, Qi Li, Minghui Qiu, Jing Gao, Shi Zhi, Lu Su, Bo Zhao, Heng Ji, and Jiawei Han. "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 745-754. ACM, 2015.

[8]. Miao, Chenglin, Wenjun Jiang, Lu Su, Yaliang Li, Suxin Guo, Zhan Qin, Houping Xiao, Jing Gao, and Kui Ren. "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. ACM, 2015.

[9]. Dong, Xin Luna, and Divesh Srivastava. "Big data integration." In Data Engineering (ICDE), 2013 IEEE 29th International Conference on, pp. 1245-1248. IEEE, 2013.

[10]. Wang, Xianzhi, Quan Z. Sheng, Xiu Susie Fang, Lina Yao, Xiaofei Xu, and Xue Li. "An integrated Bayesian approach for effective multi-truth discovery." CIKM 2015 (2015).

[11]. Li, Yaliang, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. "A Survey on Truth Discovery." arXiv preprint arXiv:1505.02463 (2015).

[12]. Waguih, Dalia Attia, and Laure Berti-Equille. "Truth discovery algorithms: An experimental evaluation." arXiv preprint arXiv:1409.6428 (2014).

[13]. Yu, Dian, Hongzhao Huang, Taylor Cassidy, Heng Ji, Chi Wang, Shi Zhi, Jiawei Han, Clare R. Voss, and Malik Magdon-Ismail. "The Wisdom of Minority: Unsupervised Slot Filling Validation based on Multi-dimensional Truth-Finding." In COLING, pp. 1567-1578. 2014.