11-2017

SourceVote: Fusing multi-valued data via inter-source agreements

Xiu Susie FANG

Quan Z. SHENG

Xianzhi WANG
Singapore Management University, xzwang@smu.edu.sg

Mahmoud BARHAMGI

Lina YAO

See next page for additional authors

SourceVote: Fusing Multi-valued Data via Inter-source Agreements

Xiu Susie Fang[1(✉)], Quan Z. Sheng[1], Xianzhi Wang[2], Mahmoud Barhamgi[3], Lina Yao[4], and Anne H.H. Ngu[5]

[1] Department of Computing, Macquarie University, Sydney, Australia
xiu.fang@students.mq.edu.au, michael.sheng@mq.edu.au
[2] School of Information Systems, Singapore Management University, Singapore, Singapore
sandyawang@gmail.com
[3] LIRIS Laboratory, Claude Bernard Lyon1 University, Villeurbanne, France
mahmoud.barhamgi@liris.cnrs.fr
[4] School of Computer Science and Engineering, UNSW, Sydney, Australia
lina.yao@unsw.edu.au
[5] Department of Computer Science, Texas State University, San Marcos, USA
angu@txstate.edu

Abstract. *Data fusion* is a fundamental research problem of identifying true values of data items of interest from conflicting multi-sourced data. Although considerable research efforts have been conducted on this topic, existing approaches generally assume every data item has exactly one true value, which fails to reflect the real world where data items with multiple true values widely exist. In this paper, we propose a novel approach, *SourceVote*, to estimate value veracity for multi-valued data items. SourceVote models the endorsement relations among sources by quantifying their two-sided inter-source agreements. In particular, two graphs are constructed to model inter-source relations. Then two aspects of source reliability are derived from these graphs and are used for estimating value veracity and initializing existing data fusion methods. Empirical studies on two large real-world datasets demonstrate the effectiveness of our approach.

Keywords: Data integration · Data fusion · Multi-valued data items · Inter-source agreements

1 Introduction

Last few years have witnessed a sheer amount of data produced and communicated among numerous sources over the Web. Unfortunately, these sources possess varying qualities and in many cases provide conflicting information on the same data items. This poses great challenges to data integration research on discovering true values from multi-sourced data, or the *data fusion* problem [7]. Considerable research efforts have been conducted to resolve this issue [8]. However, most of them assume that every data item has exactly one true value, i.e.,

single-valued assumption. This assumption fails to reflect the reality where many data items have multiple true values [15], e.g., the authors of a book. Given a data item, although we can simply concatenate and regard the values provided by the same source as a single joint value, like what previous methods do, the sets of values provided by different sources may overlap and implicitly support one another, making the concatenation unreasonable. For example, a source may claim "Charlie Booty, Lily James, Tim Roth" while another source may claim "Charlie Booty, Lily James" as the cast of the film "Broken". Apparently, the latter set is covered by the former set and therefore partially supports the former set. Since neglecting this hint may greatly impair the data fusion accuracy on multi-valued data items, we define and conduct focused study on a new topic called the *multi-valued data fusion* problem.

To the best of our knowledge, few research efforts have been devoted to the multi-valued issue in the field of truth discovery. We identify the challenges of multi-valued data fusion and the disadvantages of existing approaches as follows. Firstly, all existing methods require initializing source reliability, and for many of them, source reliability initialization impacts their performance in terms of convergence rate and accuracy. Secondly, there are implicit endorsement relations among sources when they provide some values in common. Intuitively, a source endorsed by more sources is regarded more authoritative and its provided values can be more trusted. Unlike other widely studied source relations such as copying relations, endorsement relations among sources are neglected by the previous work. Thirdly, sources may exhibit different behavioral features on multi-valued data items: some sources may provide erroneous values, leading to false positives, while some other sources may provide partial true values without making mistakes, leading to false negatives. While these two types of errors are equivalent for single-valued data items, for multi-valued data items, differentiating these errors is crucial for identifying the complete true value set. In a nutshell, our work makes three main contributions: (i) we propose a graph-based model, called *SourceVote*, as a solution to the multi-valued data fusion problem. It uses two graphs, i.e., $\pm$*Agreement Graph*, to model the two-sided endorsement relations among sources. Random walk computations are applied on both graphs to derive two-sided vote counts of sources and to finally estimate value veracity; (ii) we further derive two-sided source reliability from the two graphs to better estimate sources' quality and initialize existing data fusion methods; (iii) we conduct extensive experiments on two large real-world datasets. The results show that SourceVote consistently outperforms the baselines.

2 Related Work

Except uniformly initializing source reliability as 0.8 [10], most previous work helps data fusion methods to initialize source reliability based on prior knowledge, which is obtained by either semi-supervised methods [2] or leveraging an external trustful information source [3]. In comparison, our approach automatically derives source reliability by capturing source endorsement relations without

using any prior knowledge. The Web-link based data fusion methods [6,9] are the closest to our method. They compute the trustworthiness of sources and the truthfulness of values by using PageRank, where each link between a source and a value represents the source provides that value. However, they make single-valued assumption. To the best of our knowledge, *multi-valued data fusion* is rarely studied by the previous work. LTM (Latent Truth Model) [15] and the method proposed by Wang et al. [13] are two probabilistic models that take multi-valued objects into consideration. Waguih and Berti-Equille [10] conclude with extensive experiments that this type of models make strong assumptions on the prior distributions of latent variables, which render the modeled problem intractable and inhibitive to incorporating various considerations, and cannot scale well. Wang et al. [11] analyze the unique features of MTD and propose an MBM (Multi-truth Bayesian Model). However, they make strong assumptions on the copying of false information among sources and the independent provisioning of correct information by sources. It also requires initialization of several parameters including source reliability and copy probabilities of copiers. Recently, Wang et al. [12] design three models for enhancing existing truth discovery methods. Their experiments show that those models are effective in improving the accuracy of multi-truth discovery using existing truth discovery methods. However, LTM and MBM still performed better than those enhanced methods. None of the above methods takes the endorsement relations among sources into consideration. Different from them, our approach assumes no prior distribution or source dependency and requires no initialization of source reliability. Therefore, it is robust to various problem scenarios and insensitive to initial parameters.

3 The SourceVote Approach

The multi-valued data fusion problem involves three explicit inputs: (i) a set of *multi-valued data items*, denoted as O. Each $o \in O$ may have multiple true values; (ii) a set of *data sources*, denoted as S. Each $s \in S$ provides potential values on a subset of O; (iii) *claimed values*, denoted as V. Each $v \in V$ represents a value claimed by a source on a data item. Given a data item o, we denote the set of sources that provide values on it as S_o, and the set of all claimed values on it as V_o. In addition, we can derive several implicit inputs from the explicit inputs. Suppose the source s provides some specific values on item o (i.e., *positive claims*), denoted as ${V_{s_o}}^+$. By incorporating the *mutual exclusion assumption*, we believe s at the same time disclaims all the other values of o (i.e., *negative claims*), denoted as ${V_{s_o}}^-$, satisfying $V_o - {V_{s_o}}^+$.

To differentiate false positives and false negatives made by sources and to model source quality more precisely in multi-valued scenarios, our model focuses on two aspects of source reliability: *positive (resp., negative) precision*, the probability of the positive (resp., negative) claims of a source being true (resp., false). Note that the truth and source reliability are closely related. We formally define the multi-valued data fusion problem as follows:

Definition 1. ***Multi-Valued Data Fusion Problem.*** *Given a set of data items (O) and the conflicting values (V) claimed by a set of sources (S), the goal is to identify a set of true values (V_o^*) from V for each data item o, satisfying that V_o^* is as close to the ground truth as possible.* □

For multi-valued data items, sources may provide the same, overlapping, or totally different sets of values from one another. Generally, values agreed by the majority of sources are more trustworthy. Therefore, if the positive claims of a source are agreed by the majority of other sources, this source is likely to have high positive precision; likewise, if the negative claims of a source are disclaimed by the majority of sources, this source would be of high negative precision. That means the agreements among sources indicate endorsement, which further motivates us to model the quality of a source by quantifying the agreements and endorsement relations among data sources.

Given a data item o, we formally define the common values claimed by two sources as *inter-source agreement*. We consider two-sided inter-source agreements based on mutual exclusion. In particular, $+$*agreement*, the agreement between two sources (e.g., s_1 and s_2) on their positive claims of o, (denoted by ${A_o}^+(s_1, s_2)$) is calculated as:

$$ {A_o}^+(s_1, s_2) = {V_{s_1}}_o{}^+ \cap {V_{s_2}}_o{}^+ \tag{1} $$

Similarly, $-$*agreement*, the agreement between two sources on their negative claims of o (denoted by ${A_o}^-(s_1, s_2)$) is calculated as:

$$ {A_o}^-(s_1, s_2) = {V_{s_1}}_o{}^- \cap {V_{s_2}}_o{}^- = V_o - ({V_{s_1}}_o{}^+ \cup {V_{s_2}}_o{}^+) \tag{2} $$

The positive (resp., negative) precision of a source is endorsed by the +agreement (resp., –agreement) between this source and the other sources.

In this section, we present a graph-based approach, called *SourceVote*, as a solution to multi-valued data fusion, which is a two-step process: (i) creating two graphs based on agreements among sources (Sect. 3.1), and (ii) assessing two-sided source quality based on the graphs and further use the assessment results to estimate value veracity or initialize data fusion methods (Sect. 3.2).

3.1 Creating Agreement Graphs

By quantifying the two-sided inter-source agreements, we can construct two fully connected weighted graphs, namely $\pm$*agreement graph*s. In each graph, vertices represent sources, each directed edge depicts that one source agrees with/endorses another source, and the weight on each edge denotes the endorsement degree between the two sources. In particular, +agreement (resp., –agreement) graph models the +agreement (resp., –agreement) among the sources.

To construct the +agreement graph, we first formalize the endorsement from one source to another (e.g., $s_1 \rightarrow s_2$) on their positive claims. Specifically, for each data item that they both cover, we calculate the endorsement based on the

+agreement between the two sources. Then, we sum up the endorsement on all their overlapping data items as follows,

$$E^+(s_1, s_2) = \sum_{o \in O_{s_1} \cap O_{s_2}} \frac{|{A_o}^+(s_1, s_2)|}{|{V_{s_2}}_o{}^+|} \tag{3}$$

where O_s denotes the set of data items covered by s. Then, we calculate the weight on the edge from s_1 to s_2 as:

$$W^+(s_1 \rightarrow s_2) = \beta + (1 - \beta) \cdot \frac{E^+(s_1, s_2)}{|O_{s_1} \cap O_{s_2}|} \tag{4}$$

In Eq. (4), we add a "*smoothing link*" with a small weight between every pair of vertices, where β is the smoothing factor to guarantee the graph's full connectivity and the convergence of random walk computations. For our experiments, we simply set $\beta = 0.1$ (empirical studies [5] show this setting generally yields more accurate estimation). Finally, we normalize the weights of all out-going links of each vertex by dividing each weight by the sum of weights on all out-going links from this vertex. This normalization allows us to interpret the edge weights as the transition probabilities in random walk computations. We construct the –agreement graph in a similar way.

3.2 Estimating Value Veracity and Source Reliability

To derive two-sided source reliability (positive and negative precision) from the two graphs, the measurements should capture two features: (i) vertices with more input edges are assigned higher precision because those sources are endorsed by a large number of sources and should be more trustworthy[1]; (ii) endorsement from a source with more input edges should be more trusted because both the authoritative sources and the sources endorsed by authoritative sources are more likely to be trustworthy. We adopt *Fixed Point Computation Model* (FPC) to capture the transitive propagation of source trustworthiness through agreement links based on the ±agreement graphs [1].

By applying FPC, we obtain the ranking scores of the two-sided precision of each source among all the sources. Specifically, we refer to each agreement graph as a Markov chain, where vertices serve as the states and the weights on edges as transition probabilities between the states. We calculate the asymptotic stationary visiting probabilities of the Markov random walk, where for each graph, all visiting probabilities sum up to 1. Although, in this way, the visiting probabilities may not reflect the sources' real positive and negative precision, such feature renders the visiting probabilities of each source in the two graphs comparable. For this reason, we can count the visiting probability of each source in the +agreement (resp., –agreement) graph as the vote for its positive (resp., negative) claims being true (resp., false). We denote the corresponding vote count

[1] Here we neglect the smoothing links, i.e., no link would be there between two sources in the graphs if no common value exists between the two sources.

of each source as $\mathcal{V}^+(s)$ (resp., $\mathcal{V}^-(s)$) and further estimate the veracity of each claimed value as follows:

$$Veracity(v) = \begin{cases} True; & \text{if } \sum_{s \in S_v{}^+} \mathcal{V}^+(s) > \alpha \cdot \sum_{s \in S_v{}^-} \mathcal{V}^-(s) \\ False; & otherwise \end{cases} \tag{5}$$

where α is the source confidence factor, $S_v{}^+$ (resp., $S_v{}^-$) represents the set of sources that claim (resp., disclaim) v regarding o. Given a single-valued data item, if a source claims a value, the source certainly disclaims all the other potential values. However, sources may not know the number of true values on the data items and thus do not necessarily reject negative claims on multi-valued data items. Therefore, we adopt a new mutual exclusion definition [11] and further add a source confidence factor, $\alpha \in (0, 1)$, to differentiate the confidence of each source on its positive claims and negative claims.

To further quantify the two-sided source reliability based on the calculated visiting probabilities, we apply a two-step normalization process: (i) given the source which has the highest visiting probability in the +agreement graph (resp., –agreement graph), we first manually evaluate the positive precision (resp., negative precision) of the source, and then divide the evaluated positive precision (resp., negative precision) by the visiting probability to derive the *normalization rate*; (ii) normalizing the visiting probabilities of all sources as positive precision or negative precision, by multiplying the corresponding normalization rates.

Note that most existing methods start with initializing source reliability as a default value, e.g., set source reliability as 0.8 [10]. Such initialization may fundamentally impact the convergence rate and precision of methods. According to Li et al. [7], "*knowing the precise trustworthiness of sources can fix nearly half of the mistakes in the best fusion results*". As constructing and computing our agreement graphs can be easily realized and require no initialization of source reliability, our approach can be applied to existing methods for more precise source reliability initialization.

4 Experiments

We used two real-world datasets, including the *Parent-Children Dataset* [9] and the *Book-Author Dataset* [14]. To compare our method with traditional data fusion algorithms, we investigated the existing approaches that can be modified to tackle the multi-valued data fusion problem. As a result, we identified six methods as baselines: Voting, Sums (Hubs and Authorities) [6], Average-Log [9], TruthFinder [14], 2-Estimates [4], LTM [15], and MBM [11].

4.1 Comparison of Data Fusion Methods

Table 1 shows the performance of different approaches on the two datasets. The results show that our approach consistently achieved the best recall and F_1 score among the methods. Compared with the two existing multi-valued data fusion

Table 1. Comparison of different methods: the best and second best performance values are in bold.

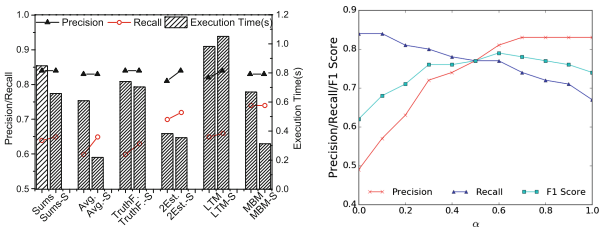
Method	Book-Author dataset				Parent-Children dataset			
	Precision	Recall	F_1 score	Time(s)	Precision	Recall	F_1 score	Time(s)
Voting	**0.84**	0.63	0.72	**0.07**	**0.90**	0.74	0.81	**0.56**
Sums	**0.84**	0.64	0.73	0.85	**0.90**	0.88	0.89	1.13
Avg-Log	**0.83**	0.60	0.70	0.61	**0.90**	0.88	0.89	**0.75**
TruthFinder	**0.84**	0.60	0.70	0.74	**0.90**	0.88	0.89	1.24
2-Estimates	0.81	0.70	0.75	**0.38**	**0.91**	0.88	0.89	1.34
LTM	0.82	0.65	0.73	0.98	0.88	**0.90**	0.89	0.99
MBM	**0.83**	**0.74**	**0.78**	0.67	**0.91**	0.89	**0.90**	2.17
SourceVote	0.81	**0.77**	**0.79**	0.63	**0.90**	**0.92**	**0.91**	0.91

methods (LTM and MBM), SourceVote had the lowest execution time. This is because LTM conducted complicated Bayesian inference over a probabilistic graphical model, and MBM includes time-consuming copy detection. Moreover, Both LTM and MBM are iterative approaches; in contrast, our approach is based on a simpler graph-based model. Although our approach achieved no significantly superior precision, the recall was improved drastically. For F_1 score, SourceVote consistently achieved the highest values for both datasets. The results reveal that our approach performs the best overall among all these baseline methods, which is consistent with our expectation because it makes no prior assumption and considers the endorsement relations among sources by combining with the graph-based method.

4.2 Empirical Studies of Different Concerns

We conducted experiments on the aforementioned baselines[2], to validate the feasibility of modeling source reliability by quantifying two-sided inter-source agreements and the feasibility of using SourceVote to initialize the existing data fusion methods. Figure 1(a) describes the performance comparison of the SourceVote initialized methods with their original versions on the Book-Author dataset. The results show that initializing source reliability by applying *SourceVote* almost led to better performance of all methods, indicated by higher precision and recall, and lower execution time. This reflects that the source reliability evaluated by *SourceVote* is more accurate than the widely applied default value of 0.8. With precise initialization, all methods achieved faster convergence speed. We also investigated the performance of SourceVote by tuning the values of the source confidence factor α from 0 to 1 on both datasets. Figure 1(b) shows the impact of α on the performance of SourceVote on the Book-Author dataset. The overall

[2] Note that we did not apply *SourceVote* to *Voting*, because *Voting* assumes all sources are equally reliable.



(a) (b)

Fig. 1. (a) Comparison between the original versions of representative existing data fusion methods and the versions that apply SourceVote for precise source reliability initialization. The latter versions are marked by suffix "-s". (b) Performance of SourceVote under varying source confidence factor, i.e., α.

performance of SourceVote peaked at the point of $\alpha = 0.6$ with an F_1 score of 0.79, which is consistent with our intuition that source confidence on positive claims should be more respected. For $\alpha \in [0.3, 0.9]$, the lowest F_1 score of SourceVote is 0.76, which is still higher than the other baseline methods. The experimental results on Parent-Children dataset showed the similar results.

5 Conclusion

In this paper, we have proposed a novel approach, *SourceVote*, to address the multi-valued data fusion problem. Our approach models the endorsement relations among sources by quantifying the agreements among sources on their positive and negative claims. Two aspects of sources reliability are derived from the modelled relations. Due to the compact feature of SourceVote, it can be leveraged to initialize and improve the existing data fusion methods. Experimental results on two large real-world datasets show that our approach outperforms the state-of-the-art data fusion methods.

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
2. Dong, X.L., et al.: Less is more: selecting sources wisely for integration. VLDB Endow. (PVLDB) **6**(2), 37–48 (2013)
3. Dong, X.L., et al.: 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 (KDD 2014), New York, USA (2014)
4. Galland, A., et al.: Corroborating information from disagreeing views. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010), New York, USA (2010)
5. Gleich, D.F., et al.: Tracking the random surfer: empirically measured teleportation parameters in pagerank. In: Proceedings of the 19th International World Wide Web Conference (WWW 2010), Raleigh, NC, USA (2010)

6. Kleinberg, J.: Authoritative sources in a hyper-linked environment. J. ACM **46**(5), 604–632 (1999)
7. Li, X., et al.: Truth finding on the deep web: is the problem solved? VLDB Endow. (PVLDB) **6**(2), 97–108 (2013)
8. Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., Han, J.: A survey on truth discovery. ACM SIGKDD Explor. Newsl. **17**(2), 1–16 (2015)
9. Pasternack, J., Roth, D.: Knowing what to believe (when you already know something). In: Proceedings of the 23th International Conference on Computational Linguistics (COLING 2010), Stroudsburg, PA, USA (2010)
10. Waguih, D.A., Berti-Equille, L.: Truth discovery algorithms: an experimental evaluation. arXiv preprint (2014). arXiv:1409.6428
11. Wang, X., et al.: An integrated Bayesian approach for effective multi-truth discovery. In: Proceedings of the 24th ACM International Conference on Information and Knowledge Management (CIKM 2015), Melbourne, Australia (2015)
12. Wang, X., et al.: Empowering truth discovery with multi-truth prediction. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 2016), pp. 881–890 (2016)
13. Wang, X., et al.: Truth discovery via exploiting implications from multi-source data. In: Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM 2016), pp. 861–870 (2016)
14. Yin, X., et al.: Truth discovery with multiple conflicting information providers on the web. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2007), San Jose, California, USA (2007)
15. Zhao, B., et al.: A Bayesian approach to discovering truth from conflicting sources for data integration. The VLDB Endow. (PVLDB) **5**(6), 550–561 (2012)