Data fusion with source authority and multiple truth
(Discussion Paper)

Fabio Azzalini, Davide Piantella and Letizia Tanca

Politecnico di Milano, Italy {name.surname}@polimi.it

Abstract. The abundance of data available on the Web makes more and more probable the case of finding that different sources contain (partially or completely) different values for the same item. Data Fusion is the relevant problem of discovering the true values of a data item when two entities representing it have been found and their values are different. Recent studies have shown that when, for finding the true value of an object, we rely only on majority voting, results may be wrong for up to 30% of the data items, since false values are spread very easily because data sources frequently copy from one another. Therefore, the problem must be solved by assessing the quality of the sources and giving more importance to the values coming from trusted sources. State-of-the-art Data Fusion systems define source trustworthiness on the basis of the accuracy of the provided values and on the dependence on other sources. In this paper we propose an improved algorithm for Data Fusion, that extends existing methods based on accuracy and correlation between sources by taking into account also source authority, defined on the basis of the knowledge of which sources copy from which ones. Our method has been designed to work well also in the multi-truth case, that is, when a data item can also have multiple true values. Preliminary experimental results on a multi-truth real-world dataset show that our algorithm outperforms previous state-of-the-art approaches.

1 Introduction

The massive use of user-generated content, the Internet of Things and the tendency to transform every real-world interaction into digital data have lead to the problem of how to make sense of the huge mass of data available nowadays. In this context, not only a source can store a previously unimaginable amount of data, but also the number of sources that can provide information relevant for a query increases dramatically, even in very specific contexts.

With all these conflicting data available on the web, discovering their true values is of primary importance. The solution of this problem is Data Fusion, where the true value of each data item is decided. Redundancy per se is not enough, since it has been shown in [3] that, if we rely only on majority vote, we could get wrong results even in 30% of the times. In order to get more accurate results we propose a Bayesian approach able to evaluate *source quality*.

Data fusion algorithms can be divided into two sub-classes: single-truth and multi-truth, the latter denoting the case when a data item may have multiple

true values. Such scenarios are common in everyday life, where many actors can play in a movie or a book can have many authors, like Alice's book "Foundations of Databases"[10] by Serge Abiteboul, Rick Hull and Victor Vianu. We decided to design our model to work also in the multi-truth case.

Currently, many single-truth data fusion algorithms exist in literature, and a few of them exploit Bayesian inference to estimate the veracity of each value and the trustworthiness of sources. TruthFinder [8] applies Bayesian analysis to compute the probability of a value being true, conditioned to the observation of values provided by the sources. Accu[4] applies a Bayesian iterative approach to compute the veracity of values, assuming uniform distribution of false values for each data item and source independence. These two assumptions have been relaxed by PopAccu[9] and AccuCopy[1] respectively.

Less attention has been devoted to studying the problem of multi-truth finding: to our knowledge, only three algorithms try to solve it. MBM[6] approaches multi-truth data fusion with a model that focuses on mappings and relations between sources and sets of provided values, introducing also a copy-detection phase to discover dependencies between sources. DART[5] computes, for each source, a *domain expertise* score relative to the domains of input data. This score is used in a Bayesian inference process to model source trustworthiness and value confidence; sources are assumed to be independent. LTM[7] exploits probabilistic graphical models to find all the true values claimed for each data item.

State of the art systems on Data Fusion define source trustworthiness based on the accuracy of the provided values and on the dependence on other sources. In this paper we propose an improved algorithm for Data Fusion. Our method extends existing methods based on accuracy and correlation between sources taking also into consideration the authority of sources. Authoritative sources are defined as the ones that have been copied by many sources: the key idea is that, when source administrators decide to copy data, they will choose the sources that they perceive as most trustworthy.

To summarize, in this paper we make the following contributions:

- We present a new formula for *domain-aware copy detection* with the goal of determine the probability that source S_i copies, from source S_j, data items belonging to a specific domain. Our copy detection process exploits the *domain expertise* of the sources and can also assign different probabilities to the two directions of copying.
- An urgent need of the truth discovery process is to determine *what sources we can trust*. We present a *fully unsupervised* algorithm that can assign an authority score to each source for each domain. This process is based on the natural habit of choosing, to copy a missing value, the source that provides the correct value with the highest probability - in other words, the most authoritative one.
- We present an improved algorithm for assessing values' veracity in a multi-truth discovery process, exploiting *source authority* in copy detection, positively rewarding sources according to their authority.

In Section 2 we present preliminary information, Section 3 provides the details about our approach and in Section 4 we show the experimental results.

2 Background and Preliminaries

We now present more in details two methods that have been of great importance for our work.

DART. This algorithm exploits an iterative domain-aware Bayesian approach to do multi-truth discovery over a dataset composed starting from different sources. Its key intuition is that, in general, a source may have different quality of data for different domains. For each source, they define the *domain expertise score* $e_{d_i}(s)$, measuring the source's experience in a given domain, and assign a confidence $c_s^o(v)$ to each value v provided by a source s, reflecting how much s is convinced that the value v is (part of) the correct value(s) for object o.

The *veracity* $\sigma_o(v)$ of value v for object o is the probability that v is a true value of o, which is better estimated at each iteration of the discovery process. The goal of the DART algorithm is to evaluate the probability that a value v is true given the observation of the claimed data $\psi(o)$ (i.e. $P(v|\psi(o))$). Being $P(\psi(o)|v)$ and $P(\psi(o)|\bar{v})$ the probabilities of having the observation $\psi(o)$ when v is true or false respectively, Bayesian inference can be used to express $P(v|\psi(o))$ as shown below:

$$P(v|\psi(o)) = \frac{P(\psi(o)|v)P(v)}{P(\psi(o))} = \frac{P(\psi(o)|v)\ \sigma_o(v)}{P(\psi(o)|v)\ \sigma_o(v) \quad + \quad P(\psi(o)|\bar{v})(1-\sigma_0(v))} \tag{1}$$

Our main criticism to DART is the assumption that sources are independent, which is a clear oversimplification of the real world. We will explain how we have relaxed this assumption in the following section.

MBM is a Bayesian algorithm for multi-truth finding that takes into consideration also the problem of source dependence. It computes, for each source and set of values, an *independence score* based on the values provided by all the sources. The *independence score* is then used to discredit, in the voting phase, sources that don't provide their values independently.

Our criticisms to MBM are the assumption that there is no mutual copying between sources in the whole dataset and the fact that the algorithm is not able to distinguish the direction of copying. In the following section we will describe how we have relaxed these assumptions.

Table 1 describes the notation that will be used in the following sections.

3 Methodology

We now present ADAM (Authority Domain Aware Multi-truth data fusion), a method based on Bayesian inference and source authority that iteratively refines the probability that a provided value for a data item is true.

3.1 Copy detection

Starting from [6], we have devised a *domain-aware* copy detection algorithm to assign different probabilities to the two directions of copy. This model works at *domain granularity*, therefore it can more accurately approximate the real world behaviour of correlated copying [2].

Notation	Description
$O(s)$	Set of all objects provided by source s
$O^d(s)$	Set of objects in domain d provided by source s
$V_s(o)$	Set of all values claimed for object o by source s
$\overline{V_s(o)}$	Set of all values claimed for object o by sources $\neq s$
$S_o^d(v)$	Sources that provide value v for object o in domain d
$S_o^d(\bar{v})$	Sources that don't provide value v for object o in domain d
$e_d(s)$	Expertise of source s in domain d
$\sigma_o(v)$	Veracity of value v for object o
$\tau_d^{rec}(s)$	Recall of source s in domain d
$\tau_d^{sp}(s)$	Specificity of source s in domain d
$c_s^o(v)$	Confidence score of value v of object o related to source s
$s_i \rightarrow s_j$	Source i is copying at object level *from* source j
$s_i \perp s_j$	Sources i and j are independent at object level
$s_i \overset{d}{\rightarrow} s_j$	Sources i is copying *from* source j for domain d
Θ_{ij}^d	Set of common objects in domain d between sources i and j
$c_{ij}(o) =: c$	Values provided by both sources i and j for object o
$\psi_{ij}^o =: \psi_c$	Observation of c
$\psi(o)$	Observation of the values provided by object o
$A_d(s)$	Authority of source s in domain d

Table 1. Notation

Scope. Given an object o and two sources s_i and s_j, we denote by ψ_{ij}^o the observation of the common values $c_{ij}(o)$ for a common object $o \in \Theta_{ij}^d$ in domain d provided by two source s_i and s_j.

Assumptions. In our copy detection algorithm we assume that there is no mutual copying *at domain level*, i.e., if source s_1 copies from source s_2 regarding domain $\bar{d}$, then s_2 can copy from s_1 only values for objects in domains $\tilde{d} \neq \bar{d}$; we also assume that two sources can only be either independent or copiers.

Object copying. For each pair of sources i, j, after we have defined the truth probability of the group of values in c as the probability that all the values are correct (Eq. 2), we can compute the likelihood of ψ_c in different cases of source dependence and truthfulness of c. Similarly to [6], we state that if s_i has copied from s_j, or the other way round, then they provide the same common values c, no matter the veracity of c (Eq. 3).

$$\sigma(c) = \prod_{v \in c} \sigma(v) \tag{2}$$

$$\begin{cases} P(\psi_c | s_i \rightarrow s_j,\ c \text{ true}) = P(\psi_c | s_j \rightarrow s_i,\ c \text{ true}) = 1 \\ P(\psi_c | s_i \rightarrow s_j,\ c \text{ false}) = P(\psi_c | s_j \rightarrow s_i,\ c \text{ false}) = 1 \end{cases} \tag{3}$$

Eq.s 4 and 5 define the probabilities that both sources provide the same group of values c independently of each other, in the two cases that c is true and false.

$$P(\psi_c | s_1 \perp s_2,\ c \text{ true}) = \tau^{rec}(s_1)\ \tau^{rec}(s_2)\ [1 - \tau^{sp}(s_1)]\ [1 - \tau^{sp}(s_2)] \tag{4}$$

$$P(\psi_c | s_1 \perp s_2,\ c \text{ false}) = \tau^{sp}(s_1)\ \tau^{sp}(s_2)\ [1 - \tau^{rec}(s_1)]\ [1 - \tau^{rec}(s_2)] \tag{5}$$

Bayesian model. If we apply a Bayesian inference approach we can now compute the probability of two sources being dependent or independent, and in the first case we can also define which of the two is the copier.

With $Y = \{s_i \to s_j;\ s_j \to s_i;\ s_i \perp s_j\}$ we define the three possible outcomes.

$$\begin{aligned} P(y|\psi_c) &= \frac{P(\psi_c|y)\ P(y)}{\sum_{y' \in Y} P(\psi_c|y')\ P(y')} \\ &= \frac{P(y)\left[P(\psi_c|y,\ c \text{ true})\ \sigma(c) + P(\psi_c|y,\ c \text{ false})\ (1-\sigma(c))\right]}{\sum_{y' \in Y} P(y')\ \left[P(\psi_c|y',\ c \text{ true})\ \sigma(c) + P(\psi_c|y',\ c \text{ false})\ (1-\sigma(c))\right]} \end{aligned} \tag{6}$$

We now have to find a way to estimate the *prior probability* of the Bayesian model: $P(s_i \to s_j)$, $P(s_j \to s_i)$ and $P(s_j \perp s_i)$, that are all the different configurations of *object copying* between sources s_i and s_j. We define this as the probability of the two sources being independent or copiers in the domain of the object we are considering, defined in Eq. 11. For ease of notation we apply the following definition, recalling that d is the same domain of $\Theta_{ij}^d \ni o$ where o is the object of ψ_c that we are analyzing.

$$\begin{cases} P(s_i \to s_j) & =: \rho_{ij}^d \\ P(s_j \to s_i) & =: \rho_{ji}^d \\ P(s_j \perp s_i) & = 1 - \rho_{ij}^d - \rho_{ji}^d \end{cases} \tag{7}$$

and replace Eq.s 7, 2, 3, 4 and 5 into Eq. 6, with the following result:

$$P(s_i \to s_j|\psi_c) = \frac{\rho_{ij}^d}{\rho_{ij}^d + \rho_{ji}^d + \left(1 - \rho_{ij}^d - \rho_{ji}^d\right)\ P_u} \tag{8}$$

where

$$\begin{aligned} P_u\ :=\ & \sigma(c)\ \left[\tau^{rec}(s_i) \cdot \tau^{rec}(s_j) \cdot (1 - \tau^{sp}(s_i)) \cdot (1 - \tau^{sp}(s_j))\right] + \\ & + (1 - \sigma(v))\ \left[\tau^{sp}(s_i) \cdot \tau^{sp}(s_j) \cdot (1 - \tau^{rec}(s_i)) \cdot (1 - \tau^{rec}(s_j))\right] \end{aligned} \tag{9}$$

Non-shared values. With Eq. 8 we have expressed the probability that a source s_i has copied from another source s_j their common values c for object o. We now have to take into consideration other possible non-in-common values to opportunely compute the probability that c were really copied. We have chosen to scale the copy probability by the Jaccard similarity of the two sets of values of o claimed by the two sources s_i and s_j, as shown in Eq. 10.

$$J_{ij}(o) = J_{ji}(o) = \frac{\left|V_{s_i}(o) \cap V_{s_j}(o)\right|}{\left|V_{s_i}(o) \cup V_{s_j}(o)\right|} \tag{10}$$

Domain-level copying. We can use the concept of copying an object o to define the act of copying with respect to a domain d as defined in Eq. 11.

$$P\left(s_i \xrightarrow{\mathrm{d}} s_j \middle| \Theta_{ij}^d\right) := \frac{\sum_{o \in \Theta_{ij}^d} P(s_i \to s_j|\psi_c) \cdot J_{ij}(o)}{\left|\Theta_{ij}^d\right|} \tag{11}$$

Initialization. Since in the initialization phase we have no prior knowledge of ρ_{ij}^d, we decided to exploit the fact that sources with high expertise in domain d are less likely to be copiers for domain d and that sources with low expertise in d tend to copy from sources with higher expertise in d. These ideas can be summarized in the initialization expressed in Eq. 12.

$$\rho_{ij}^d = \left[1 - e^d(s_i)\right] e^d(s_j) \qquad \forall s_i,\ s_j \in \mathcal{S}\ \wedge\ s_i \neq s_j \tag{12}$$

3.2 Source authority

The key idea to define the authority of a source in a specific domain with respect to the outcomes of the copy detection process is that, if many sources copy some values from the same source s_a, it is because s_a is considered authoritative and more trustworthy. For each source $s_j \in \mathcal{S}$, we define $C_d(s_j)$ in Eq. 13 as the set of all the sources that copy *from* source s_j with probability above a given threshold Γ:

$$C_d(s_j) := \left\{ s_i \in \mathcal{S} \,\middle|\, P(s_i \xrightarrow{\mathrm{d}} s_j | \Theta^d_{ij}) > \Gamma \right\} \tag{13}$$

Qualitatively, the *unadjusted authority score* of source s in domain d is *how much source s is copied in d w.r.t. how much all sources are copied in d* (Eq. 14).

$$a_d(s_j) := \frac{\sum_{s_i \in C_d(s_j)} P(s_i \xrightarrow{\mathrm{d}} s_j | \Theta^d_{ij})}{\sum_{s_k \in S} \sum_{s_l \in C_d(s_k)} P(s_l \xrightarrow{\mathrm{d}} s_k | \Theta^d_{kl})} \tag{14}$$

Note that in general the cardinality of $\mathcal{S}$ (i.e. the number of sources) is high and the parameter Γ should not be set too close to 1 to better exploit the variety of outcomes of the copy detection process. This configuration leads to $a_d(s) \ll 1$. We can accordingly apply a linear conversion to $a_d(s)$ in order to map it on the interval $[0; 1]$. We denote this new score as $A_d(s)$ or *authority of source s* in domain d, computed as:

$$A_d(s) := \frac{a_d(s) - a^{min}_d}{a^{max}_d - a^{min}_d} \tag{15}$$

3.3 Veracity

We have extended the `DART` Bayesian inference model in order to exploit the authority score of each source. Our key idea is to positively reward sources according to their authority, which can be achieved with Eq.s 16 and 17, respectively.

$$P(\psi(o)|v) = \prod_{s \in S^d_o(v)} \tau^{rec}_d(s)^{e_d(s)c_s(v)+A_d(s)} \prod_{s \in S^d_o(\bar{v})} (1 - \tau^{sp}_d(s))^{e_d(s)c_s(v)+A_d(s)} \tag{16}$$

$$P(\psi(o)|\bar{v}) = \prod_{s \in S^d_o(\bar{v})} \tau^{sp}_d(s)^{e_d(s)c_s(v)+A_d(s)} \prod_{s \in S^d_o(v)} (1 - \tau^{rec}_d(s))^{e_d(s)c_s(v)+A_d(s)} \tag{17}$$

In a multi-truth context, *precision* cannot be the only metrics for source trustworthiness [7], but we should use *recall* and *specificity*: *source recall* is the probability that true values are claimed as true (Eq. 18), while *source specificity* is the probability that false values are claimed as false (Eq. 19).

$$\tau^{rec}_d(s) = \frac{\sum_{o \in O^d(s)} \sum_{v \in V_s(o)} \sigma_o(v)}{\sum_{o \in O^d(s)} |V_s(o)|} \tag{18}$$

$$\tau^{sp}_d(s) = \frac{\sum_{o \in O^d(s)} \sum_{v' \in \overline{V_s(o)}} (1 - \sigma_o(v'))}{\sum_{o \in O^d(s)} \left|\overline{V_s(o)}\right|} \tag{19}$$

At each iteration of the algorithm veracity scores of values are refined, this leads to a better estimation of copy detection and source authority, that in turn will improve again values' veracity in the next iteration. The algorithm stops iterating when the updates of all veracities are less then a given threshold δ. The output of the algorithm is, for each object o in the dataset, a set of values whose veracities are greater or equal to a given threshold θ.

4 Experimental Results

We now present the result of an experimental comparison between our algorithm, `ADAM` (Authority Domain Aware Multi-truth data fusion), and the original `DART` in different configurations of the input data.

4.1 Dataset

We have used as input data a subset of the same *book dataset* that has been used for the evaluation of the `DART` algorithm, kindly made available by Xueling Lin, one of the authors of [5]. Our goal was to discover the correct values of the multi-truth parameter `authors` using the `category` attribute to clusterize books into domains.

For our experiments, we have been able to use a subset of this dataset matching another validated and trustworthy dataset considered as golden truth for the book-authors binding. The dataset used in our experiments is composed by 90,867 tuples from 2,680 sources and 1,958 books, spanning all the 18 domains (i.e. categories of book genres) of the original dataset.

Parameter	Value
α	1.5
Γ	0
δ	0.1
η	0.2
θ	0.5
$\bar{\sigma}$	0.5
$\bar{\tau}^{rec}$	0.8
$\bar{\tau}^{sp}$	0.9

Table 2. Parameters

Our algorithm depends on several parameters; Table 2 reports the value used for each of them. When an indication was present in [5], we used the same provided value to ensure optimal comparability between the two algorithms.

4.2 Results

We have developed in Python 3.7 both an implementation of `DART` (following as precisely as possible the guidelines expressed in [5]), and our extension `ADAM`. Even though our interest was in determine the impact of our extensions on `DART` performances, we have also developed a simple version of `MajorityVote` as baseline comparison.

`ADAM` has its F-1 score higher than `DART` in the 76% of the times. Moreover in our experiments `ADAM` has required strictly less iterations before convergence in the 65% of the times with respect to `DART`, in some cases the number of iterations required was less than a half. At first sight this faster convergence might seem to be due only to the increment of $A(s)$ in the exponent in Eq.s 16 and 17 but with a more precise analysis we discover that $A(s) \not\sim 0$ only for a small fraction of the sources, modeling in a correct manner the desired meaning of authority which by definition should be related to only a small subset of objects.

We have run 37 comparison between `DART`, `ADAM` and `MajorityVote` using the same input data for the three algorithms at each run, focusing on both input regarding single and multiple domains. We particularly focus in this section on a subset of 10 runs, reporting in Table 3 the metrics of `DART` and `ADAM` of those runs and finally in Table 4 we aggregate the results of all 37 runs reporting the averaged metrics of `MajorityVote`, `DART` and `ADAM`.

Domain	Records	|D|	|O|	|S|	DART			ADAM		
					Prec.	**Rec.**	**F-1**	**Prec.**	**Rec.**	**F-1**
Travels	221	1	44	120	0.9167	1	0.9565	0.9767	0.9545	0.9655
Reference	497	1	19	279	0.6875	0.8148	0.7458	0.84	0.7778	0.8077
History	1639	1	114	565	0.9416	0.8487	0.8927	0.9918	0.8176	0.8963
Arts	1114	1	22	505	0.5161	0.8889	0.6531	0.9545	0.6176	0.75
Random	2764	14	50	615	0.8852	0.9474	0.9153	0.98	0.875	0.9245
Random	2408	13	50	811	0.9016	0.873	0.8871	0.9808	0.85	0.9107
Random	5010	16	100	880	0.8252	0.9147	0.8676	0.9903	0.8361	0.9067
Random	4772	16	100	866	0.837	0.8433	0.8401	0.9706	0.7615	0.8534
Random*	4276	17	100	976	0.5359	0.7366	0.6205	0.9225	0.5265	0.6704
Random	1998	14	50	572	0.9	0.9643	0.931	1	0.9273	0.9623

Table 3. Experimental results (*contains only books with at least 2 authors)

Method	Precision	Recall	F-1
`MajorityVote`	0.9354	0.6958	0.7820
`DART`	0.7953	0.8621	0.8273
`ADAM`	0.9182	0.7727	0.8392

Table 4. Average results of 37 runs

5 Conclusions

We presented `ADAM`, an improved algorithm for multi-truth data fusion. A quicker termination and better results confirm that our idea to reward authoritative sources has led to an increase in the algorithm performance and accuracy.

References

1. Dong, Xin Luna and Berti-Equille, Laure and Srivastava, Divesh: Truth Discovery and Copying Detection in a Dynamic World. VLDB (2009)
2. Blanco, Lorenzo and Crescenzi, Valter and Merialdo, Paolo and Papotti, Paolo: Probabilistic Models to Reconcile Complex Data from Inaccurate Data Sources. Advanced Information Systems Eng (2010)
3. Li, Xian and Dong, Xin Luna and Lyons, Kenneth and Meng, Weiyi and Srivastava, Divesh: Truth Finding on the Deep Web: Is the Problem Solved? CoRR (2015)
4. Dong, Xin Luna and Berti-Equille, Laure and Srivastava, Divesh: Integrating Conflicting Data: The Role of Source Dependence. VLDB (2009)
5. Lin, Xueling and Chen, Lei: Domain-aware Multi-truth Discovery from Conflicting Sources. VLDB (2018)
6. Wang, Xianzhi and Sheng, Quan Z. and Fang, Xiu Susie and Yao, Lina and Xu, Xiaofei and Li, Xue: An Integrated Bayesian Approach for Effective Multi-Truth Discovery. CIKM (2015)
7. Bo, Zhao and Benjamin, Rubinstein and Jim, Gemmell and Jiawei, Han: A Bayesian Approach to Discovering Truth from Conflicting Sources for Data Integration. CoRR (2012)
8. Xiaoxin, Yin and Jiawei, Han and Philip, Yu: Truth Discovery with Multiple Conflicting Information Providers on the Web. TKDE (2007)
9. Dong, Xin Luna and Saha, Barna and Srivastava, Divesh. Less is more: Selecting sources wisely for integration. VLDB (2012)
10. Abiteboul, Serge and Hull, Richard and Vianu, Victor: Foundations of databases: the logical level, Addison-Wesley (1995)