Probabilistic Models to Reconcile Complex Data from Inaccurate Data Sources

Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo, and Paolo Papotti

Università degli Studi Roma Tre
Via della Vasca Navale, 79 – Rome, Italy
{blanco,crescenz,merialdo,papotti}@dia.uniroma3.it

Abstract. Several techniques have been developed to extract and integrate data from web sources. However, web data are inherently imprecise and uncertain. This paper addresses the issue of characterizing the uncertainty of data extracted from a number of inaccurate sources. We develop a probabilistic model to compute a probability distribution for the extracted values, and the accuracy of the sources. Our model considers the presence of sources that copy their contents from other sources, and manages the misleading consensus produced by copiers. We extend the models previously proposed in the literature by working on several attributes at a time to better leverage all the available evidence. We also report the results of several experiments on both synthetic and real-life data to show the effectiveness of the proposed approach.

1 Introduction

As the Web is offering increasing amounts of data, important applications can be developed by integrating data provided by a large number of data sources. However, web data are inherently imprecise, and different sources can provide conflicting information. Resolving conflicts and determining what values are (likely to be) true is a crucial issue to provide trustable reconciled data.

Several proposals have been developed to discover the true value from those provided by a large number of conflicting data sources. Some solutions extend a basic vote counting strategy by recognizing that values provided by accurate sources, i.e. sources with low error rates, should be weighted more than those provided by others [10,12]. Recently, principled solutions have been introduced to consider also the presence of sources that copy from other sources [6]. As observed in [1], this is a critical issue, since copiers can cause misleading consensus on false values. However, they still suffer some limitations, as they are based on a model that considers only simple (atomic) data. Sources are seen as providers that supply data about a collection of objects, i.e. instances of a real world entity, such as a collection of stock quotes. Nevertheless, it is assumed that objects are described by just one attribute, e.g. the price of a stock quote. On the contrary, data sources usually provide complex data, i.e. collections of tuples with many attributes. For example, sources that publish stock quotes always deliver values for price, volume, max and min values, and many other attributes.

B. Pernici (Ed.): CAiSE 2010, LNCS 6051, pp. 83–97, 2010.

Existing solutions, focused on a single attribute only, turn out to be rather restrictive, as different attributes, by their very nature, may exhibit drastically different properties and evidence of dependence. This statement is validated by the observation that state-of-the-art algorithms, when executed on real datasets lead to different conclusions if applied on different attributes published by the same web sources.

Source 1

	volume	min	max
AAPL	699.9k	90	150
GOOG	**1.1m**	380	545
YHOO	125k	21	48

Source 2

	volume	min	max
AAPL	699.9k	90	150
GOOG	**1.1m**	380	545
YHOO	125k	21	48

Source 3

	volume	min	max
AAPL	699.9	90	150
GOOG	1100.0k	**381**	**541**
YHOO	125.0k	21	**44**

Source 4

	volume	min	max
AAPL	699.9k	**91**	150
GOOG	1100.0k	**381**	**541**
YHOO	125.0k	21	**44**

True values

	volume	min	max
AAPL	*699.9k*	*90*	*150*
GOOG	*1100.0k*	*380*	*545*
YHOO	*125.0k*	*21*	*48*

Fig. 1. Three sources reporting stock quotes values.

This behavior can be caused by two main reasons: lack of evidence (copiers are missed) or misleading evidence (false copiers are detected). In Figure 1 we make use of an example to illustrate the issues: four distinct sources report financial data for the same three stocks. For each stock symbol are reported three attributes: volume, min and max values of the stock. The fifth table shows the true values for the considered scenario: such information is not provided in general, in this example we consider it as given to facilitate the discussion.

Consider now the first attribute, the stock volume. It is easy to notice that Source 1 and Source 2 are reporting the same false value for the volume of GOOG (errors are in bold). Following the intuition from [6], according to which copiers can be detected as the sources share false values, they should be considered as copiers. Conversely, observe that Source 3 and Source 4 report only true values for the volume and therefore there is not any significant evidence of dependence. The scenario radically changes if we look at the other attributes. Source 3 and Source 4 are reporting the same incorrect values for the max attribute, and they also make a common error for the min attribute. Source 4 also reports independently an incorrect value for the min value of AAPL. In this scenario our approach concludes that Source 3 and Source 4 are certainly dependent, while the dependency between Source 1 and Source 2 would be very low. Using previous approaches and by looking only at the volume attribute, Source 1 and Source 2 would been reported as copiers because they share the same formatting rule for such data (i.e., false copiers detected), while Source 3 and Source 4 would been considered independent sources (i.e., real copiers missed).

In this paper, we extend previous proposals to deal with sources providing complex data, without introducing any remarkable computation efforts. We formally

describe our algorithms and give a detailed comparison with previous proposals. Finally, we show experimentally how the evidence accumulated from several attributes can significantly improve the performance of the existing approaches.

Paper Outline. The paper is organized as follows: Section 2 illustrates related work. Section 3 describes our probabilistic model to characterize the uncertainty of data in a scenario without copier sources. Section 4 illustrates our model for analyzing the copying relationships on several attributes. Section 5 presents the result of the experimental activity. Section 6 concludes the paper.

2 Related Work

Many projects have been active in the study of imprecise databases and have achieved a solid understanding of how to represent uncertain data (see [5] for a survey on the topic). The development of effective data integration solutions making use of probabilistic approaches has also been addressed by several projects in the last years. In [8] the redundancy between sources is exploited to gain knowledge, but with a different goal: given a set of text documents they assess the quality of the extraction process. Other works propose probabilistic techniques to integrate data from overlapping sources [9].

On the contrary, so far there has been little focus on how to populate such databases with sound probabilistic data. Even if this problem is strongly application-specific, there is a lack of solutions also in the popular fields of data extraction and integration. Cafarella et al. have described a system to populate a probabilistic database with data extracted from the Web [4], but they do not consider the problems of combining different probability distributions and evaluating the reliability of the sources.

TruthFinder [12] was the first project to address the issue of discovering true values in the presence of multiple sources providing conflicting information. It is based on an iterative algorithm that exploits the mutual dependency between source accuracy and consensus among sources. Other approaches, such as [11] and [10], presented fixpoint algorithms to estimate the true values of data reported by a set of sources, together with the accuracy of the sources. These approaches do not consider source dependencies and they all deal with simple data.

Some of the intuitions behind TruthFinder were formalized by Dong *et al.* [6] in a probabilistic Bayesian framework, which also takes into account the effects related to the presence of copiers among the sources. Our probabilistic model is based on such Bayesian framework and extends it to the case with sources that provide complex data. A further development by the same authors also considers the variations of truth values over time [7]. This latter investigation applies for time evolving data and can lead to identify outdated sources.

3 Probabilistic Models for Uncertain Web Data

In our setting, a source that provides the values of a set of properties for a collection of objects is modeled as a *witness* that reports an *observation*. For

example, on the Web there are several sources that report the values of price, volume, dividend for the NASDAQ stock quotes. We say that these sources are witnesses of all the cited properties for the NASDAQ stock quotes.

Different witnesses can report inconsistent observations; that is, they can provide inconsistent values for one or more properties of the same object. We aim at computing: (*i*) the probability that the observed properties of an object assume certain values, given a set of observations that refer to that object from a collection of witnesses; (*ii*) the accuracy of a witness with respect to each observed property, that is, the probability that a witness provides the correct values of each observed property for a set of objects. With respect to the running example, we aim at computing the probability distributions for volume, min and max values of each observed stock quote, given the observations of the four witnesses illustrated in Figure 1. Also, for each witness, we aim at computing its accuracies in providing a correct value for volume, min and max property.

We illustrate two models of increasing complexity. In the first model we assume that each witness provides its observations independently from all the other witnesses (*independent witnesses assumption*). Then, in Section 4, based on the framework developed in [6], we remove this assumption and consider also the presence of witnesses that provide values by copying from other witnesses. The first model is developed considering only one property at a time, as we assume that a witness can exhibit different accuracies for different properties. More properties at a time are taken into account in the second model, which considers also the copying dependencies. As we discussed in the example of Figure 1, considering more properties in this step can greatly affect the results of the other steps, and our experimental results confirmed this intuition, as we report in Section 5.

For each property, we use a discrete random variable X to model the possible values it assumes for the observed object. $\mathcal{P}(X = x)$ denotes the prior probability distribution of X on the $x_1, \ldots, x_{n+1}$ possible values, of which one is correct, denoted x_t, and the other n are incorrect. Also, let $\dot{x}$ denote the event $X = x = x_t$, i.e. the event *"X assumes the value x, which is the correct value x_t"*. The individual observation of a witness is denoted o; also, $v(o)$ is used to indicate the reported value. The *accuracy* of a witness w, denoted A, corresponds to the conditional probability that the witness reports x_t, given $\dot{x}$; that is: $A = P(o|\dot{x})$, with $v(o) = x_t$.

In the following we make several assumptions. First, we assume that the values provided by a witness for an object are independent on the values provided for the other objects (*independent values assumption*). Also, we assume that the value provided by a witness for a property of an object is independent of the values provided for the other properties of the same object, and that the accuracy of a witness for a property is independent of the property values (*independent properties assumptions*). Finally, for the sake of simplicity, we consider a uniform distribution (then $\mathcal{P}(X = x) = \frac{1}{n+1}, \forall x$), and we assume that the cardinality of the set of values provided by the witnesses is a good estimation for n (*uniform distribution assumption*).

Given an object, the larger is the number of witnesses that agree for the same value, the higher is the probability that the value is correct. However, the agreement of the witnesses' observations contributes in increasing the probability that a value is correct in a measure that depends also on the accuracy of the involved witnesses. The accuracy of a witness is evaluated by comparing its observations with the observations of other witnesses for a set of objects. A witness that frequently agrees with other witnesses is likely to be accurate.

Based on these ideas of mutual dependence between the analysis of the consensus among witnesses and the analysis of the witnesses accuracy, we have developed an algorithm [3] that computes the distribution probabilities for the properties of every observed object and the accuracies of the witnesses. Our algorithm takes as input the observations of some witnesses on multiple properties of a set of objects, and is composed of two main steps:

1. *Consensus Analysis*: based on the agreement of the witnesses among their observations on individual objects and on the current accuracy of witnesses, compute the probability distribution for the properties of every object (Section 3.1);
2. *Accuracy Analysis*: based on the current probability distributions of the observed object properties, evaluate the accuracy of the witnesses (Section 3.2).

The iterations are repeated until the accuracies of the witnesses do not significantly change anymore.

3.1 Probability Distribution of the Values

The following development refers to the computation of the probability distribution for the values of one property of an object, given the observations of several witnesses, and the accuracies of the witnesses with respect to that property. The same process can be applied for every object and property observed by the witnesses.

Given a set of witnesses $w_1, \ldots, w_k$, with accuracy $A_1, \ldots, A_k$ that report a set of observations $o_1, \ldots, o_k$ our goal is to calculate: $P\Big(\dot{x}\Big|\bigcap_{i=1}^{k} o_i\Big)$; i.e. we aim at computing the probability distribution of the values an object may assume, given the values reported by k witnesses.

First, we can express the desired probability using the Bayes' Theorem:

$$P\Big(\dot{x}\Big|\bigcap_{i=1}^{k} o_i\Big) = \frac{P\Big(\dot{x}\Big)P\Big(\bigcap_{i=1}^{k} o_i\Big|\dot{x}\Big)}{P\Big(\bigcap_{i=1}^{k} o_i\Big)} \tag{1}$$

The events $\dot{x}_j$, with $j = 1 \ldots n+1$, form a partition of the event space. Thus, according to the Law of Total Probability:

$$P\Big(\bigcap_{i=1}^{k} o_i\Big) = \sum_{j=1}^{n+1} P\Big(\dot{x}_j\Big)P\Big(\bigcap_{i=1}^{k} o_i\Big|\dot{x}_j\Big) \tag{2}$$

Assuming that the observations of all the witnesses are independent,[1] for each event $\dot{x}$ we can write:

$$P\Big(\bigcap_{i=1}^{k} o_i \Big| \dot{x}\Big) = \prod_{i=1}^{k} P\Big(o_i \Big| \dot{x}\Big) \tag{3}$$

Therefore:

$$P\Big(\dot{x} \Big| \bigcap_{i=1}^{k} o_i\Big) = \frac{P\Big(\dot{x}\Big) \prod\limits_{i=1}^{k} P\Big(o_i \Big| \dot{x}\Big)}{\sum\limits_{j=1}^{n+1} P\Big(\dot{x}_j\Big) \prod\limits_{i=1}^{k} P\Big(o_i \Big| \dot{x}_j\Big)} \tag{4}$$

$P(\dot{x})$ is the prior probability that X assumes the value x, then $P(\dot{x}) = \mathcal{P}(\dot{x}) = \frac{1}{n+1}$; $P(o_i|\dot{x})$ represents the probability distribution that the i-th witness reports a value $v(o_i)$. Observe that if $v(o_i) = x_t$ (i.e. the witness reports the correct value) the term coincides with the accuracy A_i of the witness. Otherwise, i.e. if $v(o_i) \neq x_t$, $P(o_i|\dot{x})$ corresponds to the probability that the witness reports an incorrect value. In this case, we assume that $v(o_i)$ has been selected randomly from the n incorrect values of X.

Since $P(o_i|\dot{x})$ is a probability distribution:

$$\sum_{v(o_i)\neq x_t} P(o_i|\dot{x}) = 1 - A_i.$$

Assuming that every incorrect value is selected according to the uniform prior probability distribution, we can conclude:

$$P(o_i|\dot{x}) = \begin{cases} A_i & , v(o_i) = x_t \\ \frac{1-A_i}{n} & , v(o_i) \neq x_t \end{cases}. \tag{5}$$

Combining (4) and (5), we obtain the final expression to compute $P\Big(\dot{x}\Big| \bigcap\limits_{i=1}^{k} o_i\Big)$.

3.2 Witnesses Accuracy

We now illustrate the evaluation of the accuracy of the witnesses with respect to one property, given their observations for that property on a set of objects, and the probability distributions associated with the values of each object computed as discussed in the previous section.

Our approach is based on the intuition that the accuracy of a witness can be evaluated by considering how its observations for a number of objects agree with those of other witnesses. Indeed, assuming that a number of sources independently report observations about the same property (e.g. trade value) of a shared set of objects (e.g. the NASDAQ stock quotes), these observations unlikely agree by chance. Therefore, the higher are the probabilities of the values reported by a witness, the higher is the accuracy of the witness.

[1] This assumption is a simplification of the domain that we will remove later by extending our model to deal with witnesses that may copy.

We previously defined the accuracy A_i of a witness w_i as the probability that w_i reports the correct value. Now, given the set of m objects for which the witness w_i reports its observations $o_i^j, j = 1 \dots m$, and the corresponding probability distributions $P_j(\dot{x} | \bigcap_{q=1}^{k} o_q^j)$, computed from the observations of k witnesses with the equation (4), we estimate the accuracy of w_i as the average of the probabilities associated with the values reported by w_i:

$$A_i = \frac{1}{m} \sum_{j=1}^{m} P_j\Big(X = v(o_i^j) \mid \bigcap_{q=1}^{k} o_q^j\Big) \tag{6}$$

where $v(o_i^j)$ is the value of the observation reported by w_i for the object j.

Our algorithm [3] initializes the accuracy of the witnesses to a constant value, then it starts the iteration that computes the probability distribution for the value of every object (by using equations (4) and (5)) and the accuracy of witnesses (equation (6)).

4 Witnesses Dependencies over Many Properties

We now introduce an extension of the approach developed in [6] for the analysis of dependence among witnesses that removes the *independent witnesses assumption*. The kind of dependence that we study is due to the presence of copiers: they create "artificial" consensus which might lead to misleading conclusions.

As we consider witnesses that provide several properties for each object, we model the provided values by means of tuples. We assume that a copier either copies a whole tuple from another witness or it does not copy any property at all (*no-record-linkage assumption*). In other words, we assume that a copier is not able to compose one of its tuple by taking values (of distinct properties) from different sources. Otherwise, note that a record-linkage step would be needed to perform its operation, and it would be more appropriate to consider it as an integration task rather than a copying operation.

As in [6], we assume that the dependency between a pair of witnesses is independent of the dependency between any other pair of witnesses; the copiers may provide a copied tuple with a-priori probability $0 \leq c \leq 1$, and they may provide some tuples independently from other witnesses with a-priori probability $1 - c$ (*independent copying assumption*).

Under these assumptions, the evidence of copying could greatly improve by considering several properties, since it is much less likely that multiple values provided by two witnesses for the same object coincide by chance.

4.1 Modeling Witnesses Dependence

In order to deal with the presence of copiers, in the following we exploit the approach presented in [6] to modify the equations obtained with the independency assumption.

Let $W_o(x)$ be the set of witnesses providing the value x on an object and let W_o be the set of witnesses providing any value on the same object, the equation (3) can be rewritten as follows:

$$P\Big(\bigcap_{i=1}^{k} o_i \Big| \dot{x}\Big) = \prod_{w\in W_o(x)} A_w \prod_{w\in W_0 - W_o(x)} \frac{1-A_w}{n} = \prod_{w\in W_o(x)} \frac{n\cdot A_w}{1-A_w} \prod_{w\in W_o} \frac{1-A_w}{n} \tag{7}$$

Among all the possible values $x_1, \ldots, x_{n+1}$, assuming as before a uniform a-priori probability $\frac{1}{n+1}$ for each value, the equation (2) can be rewritten as follows:

$$P\Big(\bigcap_{i=1}^{k} o_i\Big) = \sum_{j=1}^{n+1} P\Big(\bigcap_{i=1}^{k} o_i \Big| \dot{x_j}\Big) P(\dot{x_j}) = \frac{1}{n+1} \sum_{j=1}^{n+1} \prod_{w\in W_o(x_j)} \frac{n\cdot A_w}{1-A_w} \prod_{w\in W_o} \frac{1-A_w}{n}$$

The probability that a particular value is true given the observations (equation (4)), denoted $P(x)$, can be rewritten as follows:

$$P(x) = P\Big(\dot{x}\Big|\bigcap_{i=1}^{k} o_i\Big) = \frac{P\Big(\bigcap_{i=1}^{k} o_i \Big| \dot{x}\Big)\frac{1}{n+1}}{P(\bigcap_{i=1}^{k} o_i)} = \frac{\prod_{w\in W_o(x)} \frac{n\cdot A_w}{1-A_w}}{\sum_{j=1}^{n+1} \prod_{w\in W_o(x_j)} \frac{n\cdot A_w}{1-A_w}}$$

The denominator is a *normalization factor*, it is independent of $W_o(x)$ and it will be denoted ω to simplify the notation.

For taking into account the witnesses' dependency, it is convenient to rewrite $P(x) = \frac{e^{C(x)}}{\omega}$ where $C(x)$ is the *confidence* of x, which is basically the probability expressed according to a logarithmic scale:

$$C(x) = \ln P(x) + \ln \omega = \sum_{w\in W_o(x)} \ln \frac{n\cdot A_w}{1-A_w}$$

If we define the *accuracy score* of a witness w as:

$$A'_w = \ln \frac{n\cdot A_w}{1-A_w}$$

it arises that we can express the confidence of a value x as the sum of the accuracy scores of the witnesses that provide that value for independent witnesses:

$$C(x) = \sum_{w\in W_o(x)} A'_w$$

We can now drop the independent witnesses assumption; to take into account the presence of copiers the confidence is computed as the weighted sum of the accuracy scores:

$$C(x) = \sum_{w\in W_o(x)} A'_w I_w$$

where the weight I_w is a number between 0 and 1 that we call the *probability of independent opinion* of the witness w. It essentially represents which "portion" of the opinion of w is expressed independently of the other witnesses. For a perfect copier I_w equals to 0, whereas for a perfectly independent witness I_w equals to 1.

I_w can be expressed as the probability that a value provided by w is not copied by any other witness:

$$I_w = \prod_{w' \neq w} (1 - cP(w \to w'))$$

where $P(w \to w')$ is the probability that w is a copier of w', and c is the a-priori probability that a copier actually copies the value provided.

Next, we will discuss how to compute a reasonable value of $P(w \to w')$ for a pair of witnesses.

4.2 Dealing with Many Properties

In [6] it is illustrated a technique to compute the probability $P(w_1 \to w_2)$ that w_1 is copier of w_2, and the probability $P(w_1 \perp w_2)$ that w_1 is independent of w_2 starting from the observations of the values provided by the two witnesses for one given property.

Intuitively, the dependence between two witnesses w_1 and w_2 can be detected by analyzing for which objects they provide the same values, and the overall consensus on those values. Indeed, whenever two witnesses provide the same value for an object and the provided value is false, this is an evidence that the two witnesses are copying each other. Much less evidence arises when the two have a common true value for that object: those values could be shared just because both witnesses are accurate, as well as independent.

We consider three probabilities, $P(w_1 \perp w_2)$, $P(w_1 \to w_2)$, $P(w_2 \to w_1)$, corresponding to a partition of the space of events of the dependencies between two witnesses w_1 and w_2: either they are dependent or they are independent; if they are dependent, either w_1 copies from w_2 or w_2 copies from w_1. $P\Big(w_1 \perp w_2 \big| \Phi\Big) =$

$$\frac{P\Big(\Phi \big| w_1 \perp w_2\Big) P(w_1 \perp w_2)}{P\Big(\Phi \big| w_1 \perp w_2\Big) P(w_1 \perp w_2) + P\Big(\Phi \big| w_1 \to w_2\Big) P(w_1 \to w_2) + P\Big(\Phi \big| w_2 \to w_1\Big) P(w_2 \to w_1)}$$

Here Φ corresponds to $\bigcap_{i=1}^{k} o_i$, i.e. the observations of the values provided by the k witnesses, and namely, o_i corresponds to the observation of the tuples provided by the witness w_i on the object.

The a-priori knowledge of witnesses dependencies can be modeled by considering a parameter $0 < \alpha < 1$, and then setting the a-priori probability $P\big(w_1 \perp w_2\big)$ to α; $P\big(w_1 \to w_2\big)$ and $P\big(w_2 \to w_1\big)$ are both set to $1 - \frac{\alpha}{2}$.[2]

[2] A similar discussion for $P\Big(w_1 \to w_2 \big| \Phi\Big)$, and $P\Big(w_2 \to w_1 \big| \Phi\Big)$ is omitted for space reasons.

The probabilities $P\Big(\Phi\big|w_1 \perp w_2\Big)$, $P\Big(\Phi\big|w_1 \rightarrow w_2\Big)$, $P\Big(\Phi\big|w_2 \rightarrow w_1\Big)$ can be computed with the help of the *independent values* assumption: the values independently provided by a witness on different objects are independent of each other.

For the sake of simplicity, here we detail how to compute, given the assumptions above, and considering our generative model of witnesses, $P\Big(\Phi\big|w_1 \perp w_2\Big)$, i.e. the probability that two independent witnesses w_1 and w_2 provide a certain observation Φ in the case of two properties denoted A and B for which they respectively exhibit error rates[3] of ϵ_1^A, ϵ_1^B, ϵ_2^A, ϵ_2^B. A similar development would be possible in the case of witnesses providing more than two properties.

Given the set of objects O for which both w_1 and w_2 provide values for properties A and B, it is convenient to partition O in these subsets: $O_{tt} \cup O_{tf} \cup O_{ft} \cup O_{ff} \cup O_d = O$. For objects in $O_{tt} \cup O_{tf} \cup O_{ft} \cup O_{ff}$, w_1 and w_2 provide the same values of properties A and B, whereas for objects in O_d the two witnesses provide different values for at least one property. In the case of objects in O_{tt}, the witnesses agree on the true value for both properties; for objects in O_{tf} they agree on the true value of A and on the same false value of B; similarly for O_{ft} they agree on the same false value of A and on the true value of B; finally, in the case of O_{ff} they agree on the same false values for both properties.

We first consider the case of both witnesses independently providing the same values of A and B and these values are either both true or both false. According to the *independent properties* assumption, w_i provides the pair of true values for A and B with probability $(1-\epsilon_i^A)(1-\epsilon_i^B)$, and a particular pair of false values with probability $\frac{\epsilon_i^A}{n_A}\frac{\epsilon_i^B}{n_B}$, with n_A (respectively n_B) being the number of possible false values for the property A (resp. B). Given that the witnesses are independent, and there are $n_A \cdot n_B$ possible pairs of false values on which the two witnesses may agree, we can write:

$$P(o \in O_{tt}|w_1 \perp w_2) = (1-\epsilon_1^A)(1-\epsilon_2^A)(1-\epsilon_1^B)(1-\epsilon_2^B) = P_{tt}$$
$$P(o \in O_{ff}|w_1 \perp w_2) = \frac{\epsilon_1^A \epsilon_2^A}{n_A}\frac{\epsilon_1^B \epsilon_2^B}{n_B} = P_{ff}$$

A witness w_i independently provides a true value of A and a particular false value for B with probability $(1-\epsilon_i^A)\frac{\epsilon_i^B}{n_B}$ (similarly for $P(o \in O_{ft}|w_1 \perp w_2)$):

$$P(o \in O_{tf}|w_1 \perp w_2) = (1-\epsilon_1^A)(1-\epsilon_2^A)\frac{\epsilon_1^B \epsilon_2^B}{n_B} = P_{tf}$$
$$P(o \in O_{ft}|w_1 \perp w_2) = (1-\epsilon_1^B)(1-\epsilon_2^B)\frac{\epsilon_1^A \epsilon_2^A}{n_A} = P_{ft}$$

All the remaining cases are in O_d:

$$P(o \in O_d|w_1 \perp w_2) = 1 - P_{tt} - P_{tf} - P_{ft} - P_{ff} = P_d$$

[3] The error rate ϵ of a witness with respect to a property is the complement at 1 of its accuracy A with respect to the same property: $\epsilon = 1 - A$.

The *independent values assumption* allows us to obtain $P\Big(\Phi\big|w_1 \perp w_2\Big)$ by multiplying the probabilities and appropriately considering the cardinalities of the corresponding subsets of O:

$$P\Big(\Phi\big|w_1 \perp w_2\Big) = P_{tt}^{|O_{tt}|} \cdot P_{tf}^{|O_{tf}|} \cdot P_{ft}^{|O_{ft}|} \cdot P_{ff}^{|O_{ff}|} \cdot P_{d}^{|O_{d}|}.$$

Now we detail how to compute $P\Big(\Phi\big|w_1 \rightarrow w_2\Big)$, but we omit $P\Big(\Phi\big|w_2 \rightarrow w_1\Big)$ since it can be obtained similarly. Recall that, according to our model of copier witnesses, a copier with a-priori probability $1-c$ provides a tuple independently. In this case, we can reuse the probabilities P_{tt}, P_{ff}, P_{tf}, P_{ft}, P_d obtained above for independent witnesses with weight $1-c$. However, with a-priori probability c, a copier witness w_1 provides a tuple copied from the witness w_2 and hence generated according to the same probability distribution function of w_2. For instance, w_2 would generate a pair of true values with probability $(1-\epsilon_2^A)(1-\epsilon_2^B)$. Concluding:

$$\begin{aligned}
P(o \in O_{tt}|w_1 \rightarrow w_2) &= (1-\epsilon_2^A)(1-\epsilon_2^B)c + P_{tt}(1-c) \\
P(o \in O_{ff}|w_1 \rightarrow w_2) &= \epsilon_2^A \epsilon_2^B c + P_{ff}(1-c) \\
P(o \in O_{tf}|w_1 \rightarrow w_2) &= (1-\epsilon_2^A)\epsilon_2^B c + P_{tf}(1-c) \\
P(o \in O_{ft}|w_1 \rightarrow w_2) &= (1-\epsilon_2^B)\epsilon_2^A c + P_{ft}(1-c)
\end{aligned}$$

For the remaining cases, we have to consider that since the witnesses are providing different values for the same object, it cannot be the case that one is copying the other.

$$P(o \in O_d|w_1 \rightarrow w_2) = (1 - P_{tt} - P_{tf} - P_{ft} - P_{ff})(1-c)$$

Again, the *independent values assumption* allows us to obtain $P\Big(\Phi\big|w_1 \rightarrow w_2\Big)$ by multiplying these probabilities raised to the cardinality of the corresponding subset of O.

5 Experiments

We now describe the settings and the data we used for the experimental evaluation of the proposed approach. We conducted two sets of experiments. The first set of experiments was done with generated synthetic data, while the second set was performed with real world data extracted from the Web.

For the following experiments we set α=0.2 and c=0.8.

5.1 Synthetic Scenarios

The goal of the experiments with synthetic data was to analyze how the algorithms perform with sources of different quality.

We conducted two sets of experiments EXP1 and EXP2 to study the performances of the approach with different configurations as summarized in Figure 2. In the two sets there are three possible types of sources: *authorities*, which provide true values for every object and every attribute; *independents*, which make

	#authorities	#independents	#copiers	$\overline{A}$
EXP1	0	8	10	0.1 - 0.9
EXP2	1	7	10	0.1 - 0.9

Fig. 2. Configurations for the synthetic scenarios

mistakes according to the source accuracy $\overline{A}$; *copiers*, which copy according to a copying rate r from the independents, and make mistakes according to the source accuracy $\overline{A}$ when they report values independently. The experiments aim at studying the influence of the varying source accuracies and the presence of an authority source (notice that the authority is not given: the goal of the experiments is to detect it).

In all the experiments we generated sources with $N = 100$ objects, each described by a tuple with 5 attributes with values for all the objects; the copiers copy from an independent source with a frequency $r = 0.8$. In all the scenarios each copier copies from three independents, with the following probabilities: 0.3, 0.3, 0.4.

In order to evaluate the influence of complex data, for each of these configurations we varied the number of attributes given as input to the algorithm with three combinations: 1, 3, and 5 attributes. We remark that our implementation coincides with the current state of the art when only one attribute is considered [6]. To highlight the effectiveness of our algorithm, we also compared our solution with a *naive approach*, in which the probability distribution is computed with a simple voting strategy, ignoring the accuracy of the sources.

To evaluate the performance of the algorithms we report the *Precision* (P), i.e. the fraction of objects on which we select the true values, considering as candidate true values the ones with the highest probability.

Results. The results of our experiments on the synthetic scenarios are illustrated in Figure 3. For each set of experiments we randomly generated the datasets and applied the algorithms 100 times. We report a graphic with the average Precision for the naive execution and for the three different executions of our approach. We used in fact executions of MultiAtt(1) with only one attribute given as input, executions of MultiAtt(3) with three attributes, and executions of MultiAtt(5) with five.

From the two sets it is apparent that the executions with multiple attributes always outperform the naive execution and the one considering only one attribute. In the first set EXP1, MultiAtt(3) and MultiAtt(5) present some benefits compared to previous solutions, but are not able to obtain excellent precision in presence of high error rates. This is not surprising: even if MultiAtt(3) and MultiAtt(5) are able to identify perfectly what sources are copiers, there are 8 independent sources reporting true values with a very low frequency and for most of the objects the evidence needed to compute the true values is missing. The scenario radically changes in EXP2, where an authority exists and MultiAtt(5) is able to return all the correct values even for the worst case, while MultiAtt(3) and MultiAtt(1) start significantly mixing dependencies at 0.8 and 0.5 error rates, respectively.

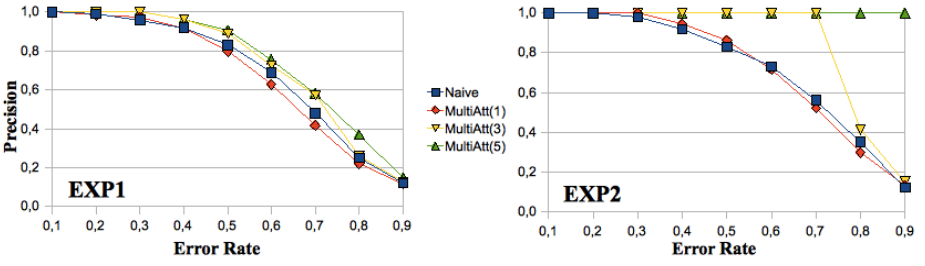

Fig. 3. Synthetic experiments: MultiAtt(5) outperforms alterative configurations in all scenarios

It is worth remarking that our algorithm does not introduce regressions with respect to previous solutions. In fact, we have been able to run all the synthetic examples in [6] obtaining the same results with all the configurations of MultiAtt. This can be explained by observing that in those examples the number of copiers is minor than the number of independent sources and MultiAtt(1) suffices for computing correctly all the dependencies. In the following, we will show that real data are significantly affected by the presence of copiers, but there are cases where considering only one attribute does not suffice to find the correct dependencies between sources.

5.2 Real-World Web Data

We used collections of data extracted from web sites about NASDAQ stock quotes.[4] All the extraction rules were checked manually, and the pages were downloaded on November 19th 2009.[5]

The settings for the real-world experiments are reported in Figure 4, which shows the list of attributes we studied. Among hundreds of available stock quotes we have chosen the subset that maximizes the inconsistency between sources.

It is worth observing that in this domain an authority exists: it is the official NASDAQ website (http://www.nasdaq.com). We ran our algorithm over the available data and we evaluated the results considering the data published by that source as the truth. The experiments were executed on a FreeBSD machine with Intel Core Duo 2.16GHz CPU and 2GB memory.

To test the effectiveness of our approach we executed the algorithm considering one attribute at a time, considering all the 10 possible configurations of three attributes, and, finally, considering five attributes at the same time. In Figure 5.a are reported the average of the precisions obtained over the five attributes by these configurations. The worst average precision (0.39) is obtained considering only one attribute at a time: this is due to the lack of clear majorities in the

[4] We relied on *Flint*, a system for the automatic extraction of web data [2].

[5] Since financial data change during the trading sessions, we downloaded the pages while the markets were closed.

Attribute	#sites	%null	#symbols	#objects
last price	39	0.3	544	250
open price	34	16.09	568	250
52 week high	34	16.59	531	250
52 week low	34	16.59	487	250
volume	39	1.98	1259	250

Fig. 4. Settings for the real-world experiments

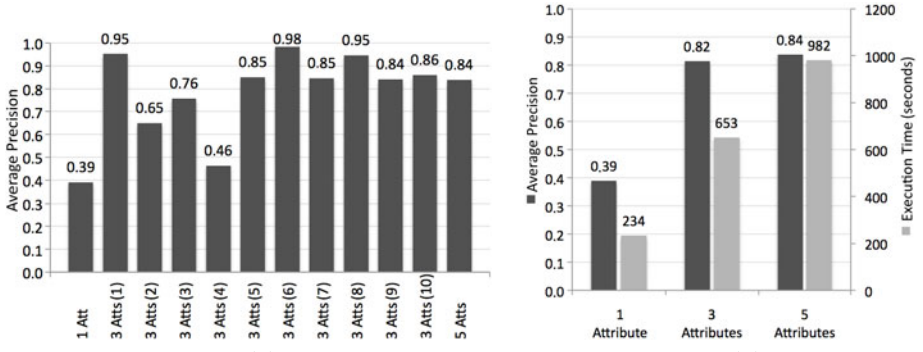

Fig. 5. Real-world experiments results

setting and the consequent difficulty in the discovery of the dependencies. We obtained interesting results considering the configurations of three attributes. In fact, it turned out that some configurations perform significantly better than others. This is not surprising, since the quality of the data exposed by an attribute can be more or less useful in the computation of the dependencies: for example, an attribute does not provide information to identify copiers if either all the sources provide the correct values or all the sources provide different values. However, it is encouraging to notice that considering all the five attributes we obtained a good precision (0.84). This shows that even if there exist attributes that do not contribute positively (or provide misleading information), their impact can be absorbed if they are considered together with the good ones.

Figure 5.b reports the average precision scores for the three configurations compared with their execution times (the average in the cases with one and three attributes). It can be observed that the execution times increase linearly with the number of attributes involved in the computation, with a maximum of 16 minutes for the configuration with five attributes.

6 Conclusions and Future Work

We developed an extension of existing solutions for reconciling conflicting data from inaccurate sources. Our extension takes into account complex data, i.e.

tuples, instead of atomic values. Our work shows that the dependence analysis is the point in which the state-of-the-art models can be extended to analyze several properties at a time. Experiments showed that our extension can greatly affect the overall results.

We are currently studying further developments of the model. First, we are investigating an extension of the model beyond the uniform distribution assumption. Second, we are studying more complex forms of dependencies such as those implied by integration processes that include a record linkage step.

References

1. Berti-Equille, L., Sarma, A.D., Dong, X., Marian, A., Srivastava, D.: Sailing the information ocean with awareness of currents: Discovery and application of source dependence. In: CIDR (2009)
2. Blanco, L., Crescenzi, V., Merialdo, P., Papotti, P.: Flint: Google-basing the web. In: EDBT (2008)
3. Blanco, L., Crescenzi, V., Merialdo, P., Papotti, P.: A probabilistic model to characterize the uncertainty of web data integration: What sources have the good data? Technical report, DIA - Roma Tre - TR146 (June 2009)
4. Cafarella, M.J., Etzioni, O., Suciu, D.: Structured queries over web text. IEEE Data Eng. Bull. 29(4), 45–51 (2006)
5. Dalvi, N.N., Suciu, D.: Management of probabilistic data: foundations and challenges. In: PODS, pp. 1–12 (2007)
6. Dong, X.L., Berti-Equille, L., Srivastava, D.: Integrating conflicting data: The role of source dependence. PVLDB 2(1), 550–561 (2009)
7. Dong, X.L., Berti-Equille, L., Srivastava, D.: Truth discovery and copying detection in a dynamic world. PVLDB 2(1), 562–573 (2009)
8. Downey, D., Etzioni, O., Soderland, S.: A probabilistic model of redundancy in information extraction. In: IJCAI, pp. 1034–1041 (2005)
9. Florescu, D., Koller, D., Levy, A.Y.: Using probabilistic information in data integration. In: VLDB, pp. 216–225 (1997)
10. Galland, A., Abiteboul, S., Marian, A., Senellart, P.: Corroborating information from disagreeing views. In: Proc. WSDM, New York, USA (2010)
11. Wu, M., Marian, A.: Corroborating answers from multiple web sources. In: WebDB (2007)
12. Yin, X., Han, J., Yu, P.S.: Truth discovery with multiple conflicting information providers on the web. IEEE Trans. Knowl. Data Eng. 20(6), 796–808 (2008)