Università degli Studi di Roma Tre
Dipartimento di Informatica e Automazione
Via della Vasca Navale, 84 – 00146 Roma, Italy.

A Probabilistic Model to Characterize the Uncertainty of Web Data Integration: What Sources Have The Good Data?

Lorenzo Blanco, Valter Crescenzi, Paolo Merialdo, Paolo Papotti

RT-DIA-146-2009 **June 2009**

Dipartimento di Informatica e Automazione
Università di Roma Tre
00146 Roma – Italy

http://www.dia.uniroma3.it

ABSTRACT

There is a large amount of data that is published on the web. Several techniques have been developed to extract and integrate data from Web sources. However, Web data is inherently imprecise and uncertain. Novel approaches to deal with the uncertain data have been recently proposed. However, they assume an uncertain degree is already associated with the data. This paper addresses the issue of characterizing the uncertainty of data extracted from a number of Web sources. We developed a probabilistic model to compute a probability distribution for the extracted values, and the reliability of the sources. We also report the results of several experiments on both synthetic and data extracted from real-life Web sites.

Contents

List of Figures

1 Introduction

As the Web is offering increasing amounts of data, several research projects have concentrated on the development of scalable techniques for extracting and integrating data from Web sources.

An interesting feature of the Web is the redundancy of information, which occurs both at the intensional and at the extensional levels, as many sources provide similar information about the same objects. For example, in the financial domain there are several Web sites that publish detailed data about attributes such as last trade, volume, and market capitalization for the NYSE stock quotes.

The redundancy of information among a large number of sources represents an interesting and specific feature that can be leveraged to address the data integration issue. However, redundancy among autonomous and heterogeneous sources also implies inconsistences in the integrated data: sources can indeed provide different values for the same property of a given object. To give a concrete example, on April 21st 2009, the open trade for the Sun Microsystem Inc. stock quote published by the CNN Money, Google Finance, and Yahoo! Finance Web sites, was `9.17`, `9.15` and `9.15`, respectively.

Inconsistency makes Web data inherently uncertain. The database community has recently proposed several solutions for modeling and processing uncertain data. However, as we discuss in Section 2, they assume that a degree of uncertainty is already associated with the data. When we extract data from the Web, we have just the raw values, without any characterization about the inherent data uncertainty nor the source reliability. Ranking schemes (such as *pageRank*) provide an indication of the reliability of the source, but they refer to properties that mainly deal with information retrieval goals and concepts.

This paper addresses the issue of characterizing the uncertainty of data extracted from Web sources. We have developed a probabilistic model, whose goal is twofold: it associates the possibly inconsistent values proposed by the various sources with a probability distribution, and it computes a reliability score for the sources. Our model analyzes the consensus among the sources and their reliability: the overall intuition is that the agreement of many sources over a certain value is reason for thinking there is a high probability that value is truth, depending on the reliability of the sources.

2 Related Work

Our work explores the application of probabilistic techniques to the results of Web data integration processes.

Many projects have recently been active in the study of imprecise databases and have achieved a solid understanding of how to represent [15] and process [7] uncertain data (see [8] for a survey on the topic). On the contrary, 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 field of data extraction and integration.

The development of effective data integration solutions based on probabilistic approaches has been addressed by several projects in the last years. Cafarella et al. have described a system to populate a probabilistic database with data extracted from the Web [4]. However they do not consider the problems of combining different probability distributions and evaluating the reliability of the sources. Also in [10] 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 [11], or other forms of dependencies between sources [16].

Our work is related to the issue of combining probability distributions expressed by a group of experts [6, 5, 12], which has been studied in the statistics community. These works follow either *behavioral* or *mathematical* approaches. Behavioral approaches attempt to generate agreement among the experts by having them interact (e.g. [9]). Behavioral approaches consider the quality of individual expert judgments, but they cannot be applied in our context as they rely on the interaction among sources. On the other hand, mathematical methods, which build on the Morris's seminal work [14], propose processes and models, mostly Bayesian, to combine individual probability distributions and produce a single distribution. These methods differ from our approach, since they consider the reliability of the experts as given or, when such information is not available, they propose solutions that do not take into account the experts reliability at all. Our proposal is mainly inspired by [13, 3] and strongly relies on the idea that the reliability of the sources plays a crucial role in the computation of the probability distribution for the value of interest.

Berti-Equille et al. have recently tacked the issue of detecting dependencies between Web data sources [1]. Their techniques could be applied to extend our setting to include dependencies between sources. Anyway, we observe that the role of plagiarism among web sources is worth debating because sources tend to copy from good sources.

3 A Probabilistic Model for Uncertain Web data

In our settings, a Web source that provides the value of a certain property for a real world object is modeled as a *witness* that reports an *observation*. For example, there are several Web sources that report the value of the last trade for the NYSE stock quotes. We say that these sources are witnesses of the last trade for the NYSE stock quotes.

Sources can provide values for many properties of a large number of objects. For example, financial Web sites usually publish the values for several stock quote properties, such as volume, max and min values, etc.. However, without loss of generalization, we develop the discussion considering only one property. For the sake of readability, in the following we may omit to specify the property an observation refer to, and by the value of an object is meant the value of the property for that object.

As we have discussed in the previous section, different witnesses can report inconsistent observations, that is, they can provide inconsistent values for the same objects. We characterize the uncertainty of data by computing the probability that the observed property of an object assumes certain values, given a set of observations that refer to that object from a collection of witnesses. With respect to the running example, we compute the probabilities distribution for the open trade of the Sun Microsystem Inc. stock quote, given the three observations illustrated in the previous section. Also, we characterize the reliability of a witness, by computing the probability that it provides the correct values for a set of objects, given the observations of other witnesses for the same set of objects.

To compute the probability that the value of an object is correct, our model considers two factors: the consensus among the witnesses' observations and the reliability of the sources. Given an object, the larger is the number of witnesses that agree for the same value, the higher is the probability that is the correct value. However, the agreement of the witnesses' observations contributes in raising the probability that a value is correct in a measure that depends also on the reliability of the involved witnesses.

To evaluate the reliability of a witness we compute how its observations compare with the observations of other witnesses for a set of objects. A witness that frequently agrees with other witnesses is likely to be reliable.

Therefore, consensus among sources and sources' reliability are mutually dependent: the greater is the reliability of the sources, the more they agree for a large number of objects. Similarly, the

more the sources agree on a large number of objects, the greater is their reliability. Based on these ideas, we have developed an algorithm, called PRE, that computes the distribution probability for the value of every observed object and the reliability of the witnesses. Our algorithm takes as input the observations of some witnesses on a set of objects (e.g. the open of the NYSE stock quotes) and the prior distribution of the observed property values. The algorithm is composed of the following three main steps:

1. **Consensus**: based on the agreement of sources among their observations on individual objects and on the current reliability of sources, compute the probability distribution for the value of every object;

2. **Reliability**: based on the current probability distributions of the observed objects, evaluate the reliability of the sources;

3. if the reliability of sources did not change in step 2, terminate. Otherwise, go to step 1.

In the following, we first present our probabilistic model for computing the probability distribution for the values of the observed objects. Then, we illustrate how we evaluate the witnesses' reliability.

3.1 Probability Distribution of the Values

The following development refers to the computation of the probability distribution for the values of one object, given the observations of several witnesses, and the witnesses' reliability. The same process can be applied for every object observed by the witnesses.

We use a discrete random variable X to model the possible values of the observed object. $\mathcal{P}(X = x)$ denotes the prior probability distribution of X on the $x_1, \ldots, x_n$ possible values. For the sake of readability, the symbol $\dot{x}$ denotes the event $X = x$, i.e. the event "x is the correct value for X".

The individual observation of a witness is denoted o; also, $v(o)$ is used to indicate the reported value.

The *reliability* of a witness w, denoted r, corresponds to the conditional probability that the witness reports x, given $\dot{x}$; that is: $r = P(o|\dot{x})$, with $v(o) = x$.

Given a set of witnesses $w_1, \ldots, w_k$, with reliability $r_1, \ldots, r_k$ that report a set of observations $o_1, \ldots, o_k$ our goal is to calculate:

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

That is, 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:

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

The events $\dot{x}_i$ forms a partition of the event space. Thus, according to the Law of Total Probability:

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

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

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

Therefore:

$$P\Big(\dot{x} \Big| \bigcap_{i=1}^{k} o_i\Big) = \frac{P\Big(\dot{x}\Big) \prod_{i=1}^{k} P\Big(o_i \Big| \dot{x}\Big)}{\sum_{j=1}^{n} P\Big(\dot{x}_j\Big) \prod_{i=1}^{k} P\Big(o_i \Big| \dot{x}_j\Big)} \qquad (1)$$

The term $P(\dot{x})$ is the prior probability that X assumes the value x, obviously equals to $\mathcal{P}(\dot{x})$.

The term $P(o|\dot{x})$ represents the probability distribution that a witness reports a value $v(o)$. Observe that if $v(o) = x$ (i.e. the witness reports the correct value) the term coincides with the reliability r of the witness. Otherwise, i.e. if $v(o) \neq x$, $P(o|\dot{x})$, it corresponds to the probability that the witness reports the incorrect value $v(o)$. In this case, we assume that $v(o)$ has been selected randomly from the $n-1$ incorrect values of X.

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

$$\sum_{v(o)\in\{x_1,\ldots,x_n\}} P(o|\dot{x}) = 1;$$

therefore,

$$\sum_{v(o)\neq x} P(o|\dot{x}) = 1 - r.$$

We can assume that every incorrect value is selected proportionally to the prior probability distribution $\mathcal{P}$, that is:

$$P\Big(o\Big|\dot{x}\Big) = \Delta\mathcal{P}\Big(X = v(o)\Big).$$

Then:

$$\sum_{v(o)\neq x} P\Big(o\Big|\dot{x}\Big) = \sum_{v(o)\neq x} \Delta\mathcal{P}\Big(X = v(o)\Big) = 1 - r.$$

From which:

$$\Delta = \frac{1-r}{1-\mathcal{P}(\dot{x})}.$$

We can then conclude:

$$P(o_i|\dot{x}) = \begin{cases} r_i & , v(o_i) = x \\ \frac{1-r_i}{1-\mathcal{P}(\dot{x})}\mathcal{P}(X = v(o)) & , v(o_i) \neq x \end{cases} \qquad (2)$$

Combining (1) and (2), we obtain the final expression to compute $P\Big(\dot{x}\Big| \bigcap_{i=1}^{k} o_i\Big)$.

Example The above formula can detect and exploit the consensus to compute how probable are the value observed by the witnesses. In the following we apply the formula using a very small dataset for shortness reasons, but we remind that the formula is intended to work over a bigger dataset.

In this example the prior probability distribution of the values is uniform and four witnesses report a (possibly null) value for each of the five observed objects $\mathcal{O}_1, ..., \mathcal{O}_5$. As we have no evidence of the witnesses' reliability, we set each of them to the same constant value 1.0.

[1]This assumption is a simplification of the domain, because one source might copy data from other sources. However, if an external subject provides the correlation of sources, it is possible to extend our model to settings where sources may copy.

	w_1	w_2	w_3	w_4	a	b	c	d	e
$\mathcal{O}_1$	c	a		c	0.0	0.0	1.0	0.0	0.0
$\mathcal{O}_2$	d	b	b	c	0.0	1.0	0.0	0.0	0.0
$\mathcal{O}_3$	a	c	e	c	0.0	0.0	1.0	0.0	0.0
$\mathcal{O}_4$		d	d		0.0	0.0	0.0	1.0	0.0
$\mathcal{O}_5$	c	e	b	d	0.0	0.25	0.25	0.25	0.25

On the left side of the table the witnesses and their observations are shown. On the right side we report, for each object, the probability distribution that associates a probability with every possible value of the object.

For $\mathcal{O}_4$ the witnesses reported always the same value "d", this is the simplest combination. The consensus is total and the formula returns 1.0 for "d" and 0.0 for every different value. An incomplete consensus is reached for objects $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3$, as two witnesses report the same value. Therefore, as the reliability of the witnesses is constant, this consensus excludes any other value. For the object o_5 there is no consensus and the reported values are equiprobable.

In this scenario the reliability of the witnesses have no influence because are constant. In the following we shall illustrate how we estimate them and we show in an example their key role in the generated probability distributions.

3.2 Witnesses Reliability

We now illustrate the evaluation of the reliability of the witnesses, given their observations for a set of objects, and the probability distributions associated with the values of each object.

Our approach is based on the intuition that the reliability 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 NYSE 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 witness' reliability.

We previously defined the reliability r_i of a witness w_i as the probability that w_i reports the correct value. Now, given the set of n objects for which the source w_i reports its observations $o_1, ..., o_n$, and the corresponding probability distributions $P_1(\dot{x}), ..., P_n(\dot{x})$, computed from the observations of many witnesses with the formula described above, we estimate the reliability of w_i as the average of the probabilities associated with the values reported by w_i:

$$r_i = \frac{1}{n} \sum_{j=1}^{n} P_j\Big(X = v_j(o_i)\Big) \qquad (3)$$

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

We have implemented the PRE algorithm in a Java working prototype. The algorithm initializes the values of the sources' reliability to a constant value, then it starts the iteration that computes the probability distribution for the value of every object (using (2)) and the reliability of sources, using (3).

Example Let us consider again the example introduced in Section 3.1. The following table illustrates this is the evolution of the reliability of the witnesses:

step	w_1	w_2	w_3	w_4
0	1.0	1.0	1.0	1.0
1	0.31	0.65	0.56	0.56
2	0.16	0.66	0.52	0.41
3	0.07	0.71	0.54	0.29
...	...	...	...	...
27	0.00	1.00	0.50	0.25
28	0.00	1.00	0.50	0.25

Notice how the consensus gradually drives the estimation until it converges to a certain set of reliability that does not change any more in two subsequent iterations. Eventually, the witness w_2 appears as the most reliable, it always reports the estimated correct values. At first sight its consensus was not clearly visible, but it is possible to verify that for the objects $\mathcal{O}_1, \mathcal{O}_2, \mathcal{O}_3, \mathcal{O}_4$ the observations of w_2 correspond to at least another one of $\mathcal{O}_3, \mathcal{O}_4$ or $\mathcal{O}_5$.

In order to appreciate the role of the witness reliability we now show the probability distributions computed after the third iteration:

	w_1	w_2	w_3	w_4	a	b	c	d	e
$\mathcal{O}_1$	c	a		c	0.73	0.08	0.03	0.08	0.08
$\mathcal{O}_2$	d	b	b	c	0.01	0.92	0.03	0.05	0.01
$\mathcal{O}_3$	a	c	e	c	0.01	0.04	0.70	0.04	0.21
$\mathcal{O}_4$		d	d		0.02	0.02	0.02	0.92	0.02
$\mathcal{O}_5$	c	e	b	d	0.09	0.27	0.01	0.05	0.58

Every value that in subsequent iterations will be supposed correct is evaluated with a probability higher than 0.5.

The probability distribution for $\mathcal{O}_1$ and $\mathcal{O}_5$ are remarkable. The former is a rare case of consensus over uncorrect values, but generally the consensus rewards w_2. Therefore, the consensus of w_1 and w_4 is overcome by the reliability of w_2. The latter probability distribution is a common case of total lack of consensus, in this situation the witness with higher reliability prevails.

4 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 were done with synthetic data while the second set were performed with real world data: we used collections of data extracted from web sites from three distinct domains, i.e. soccer players, stock quotes, and videogames.

The goal of the experiments with synthetic data was to analyze the algorithms behavior under varying conditions: number of sources, percentages of null values, reliability of the sources. In particular, we choose configurations of synthetic data that were comparable to those found in real world sources.

4.1 Synthetic scenarios

We have developed a tool for the generation of synthetic scenarios. The tool simulates the data provided by a collection of k sources, by generating the values for a set of n objects. The generation process is configured with the following parameters: prior distribution for the values and cardinality of the values alphabet; also, for each source of the k sources, we can configure the percentage of null values, and the source reliability $\overline{r}$.

For every experiment, in an initialization step, the tool generates a sequence T of values $t_1, \ldots, t_n$, according to the configured prior distribution.[2]These values are taken as the truth for the current experiment: they represent the correct (true) values for the observed objects. Then, the tool simulates the generation of the values produced by the k sources, according to the input configuration. For every object the tool produces either a correct value, or an uncorrect value, or a null value, according to the source configuration.

We conducted three sets of experiments to study the performance of the algorithm with different configurations. In all the experiments we generated values for $n = 500$ objects. Figure 1 summarizes the configurations of the three sets of experiments. The three sets of experiments EXP1, EXP2, EXP3 aim at studying the influence of the percentage of null values, the number of sources, the number of alphabet values, respectively.

	%null	#sources	#symbols
EXP1	**0%–90%**	30	15
EXP2	25%	**4–36**	15
EXP3	25%	30	**2–40**

Figure 1: Configurations for the synthetic scenarios.

In order to evaluate the influence of reliable and unreliable sources, for each of these configurations, we also varied the ratio between *bad sources* and *good sources*. The reliability of bad sources is 0.4 (that is, 60% of the reported values are uncorrect), whereas the reliability of good sources is 0.9. We used three combinations of good and bad sources with the following percentages of good ones: 30%, 60%, and 90%.

We run our algorithm to compute the probability distribution for each object, and the reliability of each source. In the following we call this procedure the *iterative approach*. To highlight the effectiveness of our algorithm, we also compared our solution with a *naive approach*, in which the probability distribution is computed ignoring the reliability for the sources. In practice, as done by approaches that do not provide any solution to evaluate the reliability, we impose a constant reliability (0.6) for all the witnesses and compute the distributions in just one step, without any iteration.

4.1.1 Evaluation Metrics

We rely on two metrics, called *probability concentration* (PD) and *reliability precision* (RP), to measure the performances of the algorithms in computing the probability distributions and the reliability of the sources, respectively.

Probability Concentration (PC) The Probability Concentration measures the performance of the algorithms in computing the probability distributions for the observed objects. Given the truth vector T, the probability concentration is the average probability associated to the correct value:

$$PD = \frac{1}{n} \sum_{j=1}^{n} P_j\Big(X = t_j\Big).$$

Note that if all the probability distributions associate a probability value of 1 to the correct value, PD equals 1. Conversely the lower is PD the more the probability distributions are scattered over uncorrect values, i.e they associate probability to uncorrect values.

[2]For the sake of simplicity we always adopt a uniform distribution (which is therefore derived from the cardinality of the values alphabet).

Reliability Precision (RP) We compare the results of the algorithm with the actual reliability of the sources, which is specified by the input configuration, in order to measure the quality of the computed reliability:

$$RP = \frac{1}{k} \sum_{i=1}^{k} |r_i - \overline{r}_i|.$$

Note that if the estimated values for the reliability of the sources are identical to the real ones, then RP equals to 0.

4.1.2 Results

The results of the our experiments on the synthetic scenarios are illustrated in Figure 2. For each set of experiments we report three PC graphics to plot the Probability Concentration, where each graph refers to different percentage of good sources, and one RP graphic to plot the Reliability Precision. Each PC graphic reports the performances of both the naive and the iterative approaches. The RP graphics plot the curves for the reliability evaluation with the three different percentages of good sources.

It is apparent that the iterative approach always outperforms the naive one, and that in all the considered settings, the Probability Concentration descends below 0.9 only when either the percentage of null values is greater than 75%, or the number of sources is below 8. In all other cases, by looking at the RP graphics, it can be noted that the average errors in the reliability estimation is well below 0.02. Overall, the algorithm reacts slowly to the worsening of the sources, and even with only 30% of good sources, the Probability Concentration is around 0.75 for only 4 sources. Compared to the naive approach, the results of the iterative approach are only marginally affected by the number of symbols in the alphabets, and it is able to reach optimal results already with 3 symbols.

4.2 Real-World Web data

To gather data from dozens of Web sites we used *Flint*, a system for the automatic extraction and integration of Web data [2].[3]

The settings for the real-world experiments are reported in Figure 3, which shows the list of attributes that we studied for each domain.[4]. `#objects` reports the number of objects that have at least two sources reporting a value for the actual attribute. The lower bound for this dimension is a few hundreds.

It is worth observing that the percentage of null values (`%null`) is very low for all the stock quotes attributes, while it is rather high for the other two domains. Also the size of the values alphabet (`#symbols`) differs among the attributes, ranging between five and 2079.

We ran our algorithm over the three scenarios with interesting results. For space limitations we cannot report the individual results of the experiments; thus we discuss some important results of this experimental activity.

Stock quotes Web sources reporting financial data shown very high average reliability for all attributes. For example, the average reliability over all the sources for the attributes *daily high price* and *last trade price* were `0.98` and `0.99`, respectively. For the *last trade price* the lowest

[3]Pages were downloaded on April 20th 2009. Since financial data change during the trading sessions, we have downloaded the pages while the markets were closed.

[4]The number of sites considered for each attribute varies (even in the same domain) because not all the sources present the same set of attributes (this is particularly evident for the Soccer domain) or because of limitations of the data extraction process

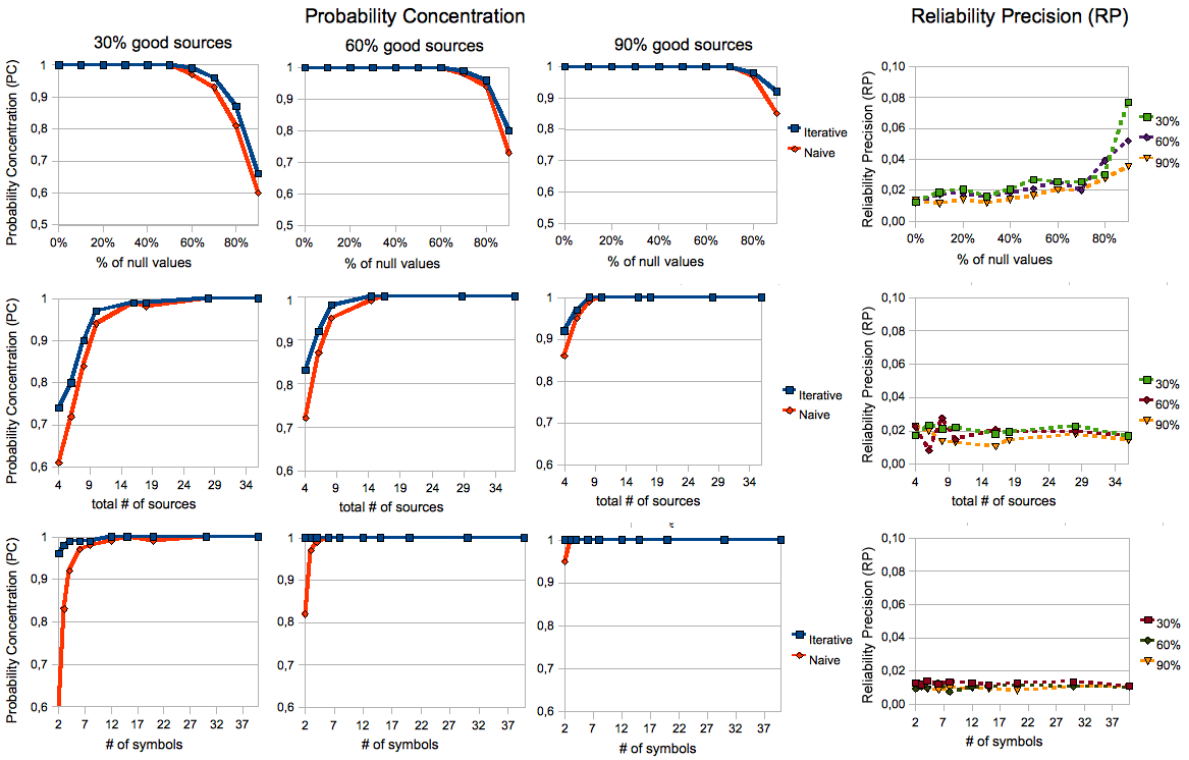

Figure 2: Synthetic experiments.

	#sites	%null	#symbols	#objects
Finance				
last trade	20	12.08	545	525
day high	15	12.32	545	519
day low	14	10.94	541	512
$ variation	22	16.53	311	567
volume	17	10.90	575	515
Soccer				
height	19	85.06	49	1221
weight	11	78.06	61	717
club	19	85.94	345	831
position	15	84.98	5	213
birthdate	21	85.05	2696	1706
nationality	18	87.02	112	566
VGames				
ESRB	16	83.22	6	1003
soft. house	20	83.34	296	2089

Figure 3: Settings for the three real-world experiments.

reliability value was 0.98, and it was 1 for eight sources. For the *volume* the reliability values range between 0.81 and 0.99, with 15 sources above 0.90.

Soccer players In the sport domain, sources do not manifest a strong consensus, and the estimated reliability values were more heterogeneous than in the stock quote domain. In this domain, we also had a notable example: the reliability of one source was extremely low for all the attributes. For example, we obtained a reliability value of 0.08 for *height* and of 0.01 for *birthdate*. We investigated the source and it turned out that the Web site exposes random data: values for players' attributes change every time a page is loaded. [5] Excluding the reliability of this fake source, the reliability for *height* ranges between 0.33 and 0.90, for *birthdate* between 0.90 and 0.99, for *weight* between 0.51 and 0.76.

Videogames The results for this domain place between the two previous ones. The sources expose the data with significant accuracy and for both attributes the majority of the sources have reliability greater than 0.90.

We remark that an evaluation of the performance for the real-world configurations can be derived from a comparison with the results of the synthetic experiments. For the stock quotes scenario, the low amount of duplicates, the large alphabet sizes, and the number of sources (which is always greater than 14) for each attribute guarantee that the algorithm estimated the reliability for such sources with an average error below 0.02. For the soccer and videogames scenarios, the settings present a much higher amount of null values, but in terms of average error the results have an estimated error below 0.05 wrt the real reliability of the sources.

References

[1] Laure Berti-Equille, Anish Das Sarma, Xin Dong, Amélie Marian, and Divesh Srivastava. Sailing the information ocean with awareness of currents: Discovery and application of source dependence. In *CIDR*, 2009.

[2] L. Blanco, V. Crescenzi, P. Merialdo, and P. Papotti. Flint: Google-basing the web. In *EDBT*, 2008.

[3] Laurence BonJour. *The structure of empirical knowledge*. Harvard University Press, 1985.

[4] Michael J. Cafarella, Oren Etzioni, and Dan Suciu. Structured queries over web text. *IEEE Data Eng. Bull.*, 29(4):45–51, 2006.

[5] Robert T. Clemen and Robert L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19(2):187 – 203, 1999.

[6] Roger Cooke. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press US, 1991.

[7] Nilesh N. Dalvi and Dan Suciu. Efficient query evaluation on probabilistic databases. *VLDB J.*, 16(4):523–544, 2007.

[8] Nilesh N. Dalvi and Dan Suciu. Management of probabilistic data: foundations and challenges. In *PODS*, pages 1–12, 2007.

[5] E.g., http://soccer.azplayers.com/players/F/Flaco

[9] Morris H. DeGroot. Reaching a consensus. *Journal of the American Statistical Association*, 69(345):118–121, 1974.

[10] Doug Downey, Oren Etzioni, and Stephen Soderland. A probabilistic model of redundancy in information extraction. In *IJCAI*, pages 1034–1041, 2005.

[11] Daniela Florescu, Daphne Koller, and Alon Y. Levy. Using probabilistic information in data integration. In *VLDB*, pages 216–225, 1997.

[12] Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986.

[13] Michael Huemer. Probability and coherence justification. *Southern Journal of Philosophy*, 35(4):463–72, 1997.

[14] Peter A. Morris. Combining expert judgments: A bayesian approach. *Management Science*, 23(7):679–693, 1977.

[15] Anish Das Sarma, Omar Benjelloun, Alon Y. Halevy, and Jennifer Widom. Working models for uncertain data. In *ICDE*, page 7, 2006.

[16] Anish Das Sarma, Xin Dong, and Alon Halevy. Data integration with dependent sources. Technical report, Stanford InfoLab, December 2008.