# Characterizing the Uncertainty of Web Data: Models and Experiences

Lorenzo Blanco Valter Crescenzi Paolo Merialdo Paolo Papotti
Università degli Studi Roma Tre
Via della Vasca Navale, 79 – Rome, Italy
blanco,crescenz,merialdo,papotti@dia.uniroma3.it

## ABSTRACT

An increasing number of web sites offer structured information about recognizable concepts, relevant to many application domains, such as finance, sport, commercial products. However, web data is inherently imprecise and uncertain, and conflicting values can be provided by different web sources. Characterizing the uncertainty of web data represents an important issue and several models have been recently proposed in the literature. The paper illustrates state-of-the-art Bayesan models to evaluate the quality of data extracted from the Web and reports the results of an extensive application of the models on real life web data. Our experimental results show that for some applications even simple approaches can provide effective results, while sophisticated solutions are needed to obtain a more precise characterization of the uncertainty.

## Categories and Subject Descriptors

H.4.3 [**Information Systems Applications**]: Information browsers

## General Terms

Documentation, Experimentation

## Keywords

Web Data Extraction, Data Reconciliation, Probabilistic Data

## 1. INTRODUCTION

The Web is offering increasing amounts of data, which are becoming more and more important in several human activities. Consumers consult online catalogs to choose products they are willing to buy; individuals and institutions rely on the financial data available on the Web to take decisions about their trading activities; many people collect information from specialized web sites for leisure interests and hobbies. Several tools and techniques are now available to extract data from the Web, and many applications and services are built by integrating data provided by multiple sources. However, web data are inherently imprecise and uncertain, and different sites often provide conflicting data.

| number of distinct values | Open value | Volume | Price |
|---|---|---|---|
| 1 | 81 (10%) | 0 (0%) | 261 (32%) |
| 2 | 579 (71%) | 1 (0%) | 477 (58%) |
| 3 | 126 (15%) | 44 (5%) | 57 (7%) |
| 4 | 20 (2%) | 139 (17%) | 15 (2%) |
| 5 | 10 (1%) | 100 (12%) | 5 (1%) |
| 6 | 3 (0.5%) | 389 (48%) | 4 (0.5%) |
| $> 6$ | 0 (0%) | 146 (18%) | 0 (0%) |

**Figure 1: Number of distinct values reported by 34 web site for the *open value* on a random sample of 819 NASDAQ stocks quotes: the 34 sources agree on the same Open value only for 81 (about 10%) objects and there is no stock for which they all propose the same Volume value.**

To give a concrete example, let us consider the financial domain. On a fixed date, we looked at the Open trade of the `Apple Inc.` stock quote on 34 web sites that provide information about NASDAQ stocks: 24 sites reported `204.61`, 7 sites reported `204.51`, 3 sites reported `204.57`. Figure 1 illustrates the numbers of distinct values published by the 34 web sites for the open value on a random sample of 819 NASDAQ stock quotes: only for 81 stock quotes all the 34 sources agreed on a unique value; for 579 stock quotes 2 values were reported, and for a significant number (about 20%) more than 2 values were provided (with a peak of 6 distinct values for 3 stocks). Similar results can be observed for the Price and the Volume values.

We compared the open values published by each source with the official ones,[1] obtaining that the accuracy (i.e. the error rate) of the 34 source ranges from 0.18 to 0.98 (0.73 on average). Surprisingly, among the sources with the lowest accuracies, we found several popular web sites. Indeed, traditional and popular ranking methods (such as Google's *PageRank*, or Alexa's *Traffic Rank*) provide an indication about the overall popularity of the source, but they rely on properties that do not refer to quality of the published data. Even when sources are authoritative, the quality of data

[1]The open value of a stock quote is an official information provided by NASDAQ.


*WebQuality* '11, March 28, 2011 Hyderabad, India.

delivered in their pages can be compromised by editorial choices (e.g. numeric values might be deliberately approximated) and by the complexity of the publishing process, which can introduce errors and imprecisions. The problem is even exacerbated because many sources harness and publish data integrating information from other sources, introducing further complications in the process and propagating possible errors.

To evaluate the accuracy of a web source there is the need of an authority that provides the true values for the data of interest. In the above example, we were able to compute the accuracies of the sources with respect to the open values of the NASDAQ stock quotes because the NASDAQ web site publishes the official values. However, in general this is not the case: an authority could be missing, or data consumers could not be aware of its presence.

Another important issue that emerges from the example is that the uncertainty of data coming from conflicting web sources should be characterized by a probability distribution function that associates every value found in the sources with a truth probability. In our example, the probability distribution would express the probabilities that `204.61`, `204.51`, or `204.57` is the true open value for the `Apple Inc.` stock quote.

The database research community recently addressed these issues and developed several approaches to characterize the uncertainty of data coming from multiple sources. However, the proposed methods have been tested on real-life web data only to a limited extent, while systematic studies have been done mostly on synthetic data sets. The goal of this paper is twofold: first, we illustrate the principles of state-of-the-art approaches for evaluating the accuracy of web data and computing a probability distribution for the values they provide. Then, we present the results of the application of an implementation of these approaches on real-life web data from several domains. Our results show that in real life scenarios, even the simplest approach produces reliable results, especially in estimating the accuracy of the sources. More involved solutions outperform simple solution in general and are certainly needed when there is the need to compute precise probability distribution functions. Also, the experiments show that the accuracy of web data source is only marginally related to some popular web ranking indices.

The paper is organized as follows. Section 2 illustrates the principles and the intuitions of methods for computing the accuracy of data sources by observing the data they provide. Section 3 presents the results of a experimental activity that we have conducted on web data. Section 4 discusses related works. Section 5 concludes the paper.

## 2. PROBABILISTIC MODELS FOR UNCERTAIN WEB DATA

Web sources usually provide values for some properties of a large number of objects. For example, financial web sites publish the values for several stocks properties, such as volume, open, max and min values, etc.. Different sources can report inconsistent values of the properties for the same object making data published on the Web inherently uncertain.

The uncertainty of data can be characterized by probability distribution functions: data are associated with functions reporting the probabilities that a property assumes certain values for a given object. These possible values are collected from a set of sources publishing information about that object. Therefore, there is one probability distribution function for each pair property of each object.

| Object | A: *authority* | I: *independent* | IC: *ind. copied* | C1: *copier 1* | C2: *copier 2* | NAIVE | ACCU | DEP |
|---|---|---|---|---|---|---|---|---|
| | Sources | | | | | Model | | |
| obj1 | a | c | b | b | b | b | a\|b | a |
| obj2 | b | b | c | c | c | c | b | b |
| obj3 | c | b | c | c | c | c | c | c |
| *accuracy* | 1 | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | | | |

**Figure 2: Running Example:** *authority* **always reports the correct value;** *independent* **and** *independent copied* **provide, independently from the other sources, one correct value out of three;** *copier 1* **and** *copier 2* **copy their values from** *independent copied*.

The models to characterize the uncertainty of web data have a twofold goal. They aim at computing (*i*) the probability distributions for data provided by the sources, and (*ii*) the accuracy of the sources with respect to the data of interest, that is, the probability that a source provides the correct values for a set of objects.

State of the art models to characterize the uncertainty of Web data can be classified according to which of the following three main factors they take into account:

**consensus** given an object, the larger is the number of sources that agree for the same value, the higher is the probability that the value is correct;

**accuracy** the agreement of the sources' observations contributes in raising the probability that a value is correct in a measure that also depends on the accuracy of the involved sources;

**copiers** the presence of copiers, that is, sources that publish data copied by other sources, can generate misleading consensus on the values proposed by the most copied sources.

As an example, consider 5 sources publishing the values of a property for the same 3 objects, as shown in Figure 2. The first source A is an *authority* and provides the true value for each object (a, b, and c respectively), while all the other sources (I, IC , C1, C2) provide a correct value only for one object out of three. The sources I and IC (*independent* and *independent copied*) provide their values independently from the other sources, while the remaining sources C1 and C2 merely copy the values proposed by IC. Notice that in normal settings, the role of the sources is not available as input to the models (i.e., they are not aware that A is an authority, I is independent, and so on).

The simplest model, called NAIVE,[2] for estimating the accuracy of sources and computing data distribution probabil-

[2]The names of the models presented in this paper are inspired to those used in [11].

**Web Source 1**

| | *genre* | *publisher* | *release date* |
|---|---|---|---|
| The Sims 3 | Simulation | EA | June 5 2009 |
| Doom 3 | **FPS** | Activision | August 3 2004 |
| StarCraft 2 | RTS | Blizzard | July 27 2010 |

**Web Source 2**

| | *genre* | *publisher* | *release date* |
|---|---|---|---|
| The Sims 3 | Simulation | EA | June 5 2009 |
| Doom 3 | **FPS** | Activision | August 3 2004 |
| StarCraft 2 | RTS | Blizzard | July 27 2010 |

**Web Source 3**

| | *genre* | *publisher* | *release date* |
|---|---|---|---|
| The Sims 3 | Simulation | EA | June 5 2009 |
| Doom 3 | Shooter | **Steam** | **August 2004** |
| StarCraft 2 | Strategy | Blizzard | **July 2010** |

**Web Source 4**

| | *genre* | *publisher* | *release date* |
|---|---|---|---|
| The Sims 3 | Simulation | **Sims Studio** | June 5 2009 |
| Doom 3 | Shooter | **Steam** | **August 2004** |
| StarCraft 2 | Strategy | Blizzard | **July 2010** |

*True values*

| | genre | publisher | release date |
|---|---|---|---|
| The Sims 3 | *Simulation* | *EA* | *June 5 2009* |
| Doom 3 | *Shooter* | *Activision* | *August 3 2004* |
| StarCraft 2 | *Strategy* | *Blizzard* | *July 27 2010* |

**Figure 3: Four web Sources reporting data for three video games. The last table represents the true values for the scenario.**

ities considers only the consensus: the most probable value is the one published by the largest number of sources, and the probability of a value is estimated by its frequency over the given set of sources. According to NAIVE, in our running example there are 2 possible values for obj2 provided by the 5 sources considered: b is provided by sources A and I, while b is provided by 3 sources (IC and its copiers C1, C2). Similarly for obj1 the most likely value would erroneously be the value b. The example shows how in presence of many sources with different accuracy, a naive voting strategy could lead to incorrect conclusions. The probability distribution of the NAIVE model corresponds to the frequencies of the values. In the running example, the probability for obj1 is a$\rightarrow 1/5$, b$\rightarrow 3/5$, c$\rightarrow 1/5$.

A more involved model, ACCU, considers at the same time the first two factors, consensus and accuracies, and produces as output an estimation of the source accuracies together with the probabilities of the values [15, 16, 17]. Indeed, consensus among sources and sources' accuracy are mutually dependent: the greater is the accuracy of the sources, the more they agree for a large number of objects and the more they will affect the general consensus. Similarly, the more the sources agree on a large number of objects, the greater is their accuracy.

The role of the accuracies consist in weighting the consensus of the sources. A voting strategy similar to that used with the NAIVE model can be used to estimate the probabilities by means of the consensus: the only difference is that the votes are weighted according to the accuracies of the sources. The accuracy of a source can be estimated by comparing its observations with those of other sources for a set of objects. A source that frequently agrees with other sources is likely to be accurate, and similarly, the most accurate sources will be given the higher weights during the computation of the probabilities of the true values. In our running example, consider the accuracies given at the bottom of the table in Figure 2: 3 sources (IC, C1, and C2) provide the wrong value c for obj2, and they will be given a overall accuracy of 1, while 2 sources (A,I) provide the correct value b with an overall accuracy of $\frac{4}{3}$. However, even if the accuracies are known, the model still cannot decide which value, between a and b, is the most likely value for obj1.

A more complex model also considers the presence of copiers, that is, sources that publish values copied by one or more other sources. The presence of copiers makes harder the problem of computing the true values and the accuracies of the sources since they can create "artificial" consensus on values. A copier, even in good faith, can propagate a wrong value originated in one of the sources from which it copies. Provided that there is enough evidence about which are the correct values, it is possible to detect which sources are copying observing that copiers publish the same false values of the sources from which they copy. For instance, if b is considered the most likely value for obj2, the fact the IC, C1 and C2 publish the same false value attests that there are two copiers. The same argument cannot be used for obj3, for which the three sources publish the same value c: since this is a true value, it is not necessarily an evidence of coping.

DEP is a model that considers all the three factors above: consensus, accuracy and copiers [11]. It tries to detect possible copiers by analyzing the dependencies among the sources. Once the copiers has been detected, the consensus created by their presence will be ignored during the computation of the probabilities. The dependence analysis has to consider the mutual feedbacks amongst consensus, accuracy and dependencies: the accuracy of a source depends on the consensus over the values it provides; the dependencies between sources depends on sources accuracy and the consensus over the values they provide; finally, the consensus should take into account both the accuracy of sources and the dependencies between sources. For instance, once that it has been detected that IC, C1 and C2, copy one each other, the voting expressed by two sources will be ignored, and then it can be established that the most likely true value of obj1 is a.

In general, identifying the copiers is a challenging task for two main reasons. First, if in the considered sources there is a lack of evidence, copiers can be missed. Second, if the available evidence is misleading, false copiers can be detected. In Figure 3 we make use of an example to illustrate these issues: four distinct web Sources report data about the same three video games. For each video game three attributes are reported: *genre*, *publisher*, and *release date*. The fifth table shows the true values for the considered scenario. We remark that such information is not provided in general, in this example we consider it as given to facilitate

| | Soccer Players (20 sources) | | | | Video Games (30 sources) | | | Stock Quotes (30 sources) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Birthdate | Height | Weight | Avg | ESRB | Publisher | Avg | Price | Open Value | Volume | Avg |
| *# objects* | 976 | 980 | 972 | 976 | 288 | 166 | 227 | 819 | 819 | 819 | 819 |
| *# symbols* | 1435 | 47 | 50 | 510 | 5 | 75 | 40 | 1892 | 2011 | 4812 | 2902 |

**Figure 4: Statistics about the data extracted from the Web.**

the discussion.

Consider now the first attribute, the genre of the game. It is easy to notice that web Source 1 and web Source 2 are reporting the same false value for the genre of Doom 3 (errors are in bold). Following the intuition from [11], according to which copiers can be detected as the sources share false values, they should be considered as copiers. Conversely, observe that web Source 3 and web Source 4 report only true values for the genre and therefore there is not any significant evidence of dependence. The scenario radically changes if we look to the other attributes. Web Source 3 and web Source 4 are reporting the same incorrect values for the release date attribute, and they also make a common error for the publisher attribute. Web Source 4 also reports independently an incorrect value for the publisher of The Sims 3. In this scenario our approach concludes that web Source 3 and web Source 4 are certainly dependent, while the dependency between web Source 1 and web Source 2 would be very low. Using the DEP model, therefore by looking only to a single attribute at the time, web Source 1 and web Source 2 would been reported as copiers for the genre attribute because they share the same formatting rule for such data (i.e., false copiers detected), while web Source 3 and web Source 4 would been considered independent sources (i.e., real copiers missed).

Starting from the above observations, the dependence analysis has been further investigated and a more complex model M-DEP has been introduced to consider not only single attributes at a time, but whole tuples [5, 10].

# 3. EXPERIENCING THE MODELS ON WEB DATA

We have developed a Java prototype that implements the models described in the above section, to experience the models on the data provided by real life web data sources. We used collections of data extracted from web sites from 3 distinct domains: soccer players, video games, and stock quotes. Data were collected by means of Flint, a system to extract and integrate web data [4, 3]. The experiments were executed on a FreeBSD machine with Intel Core i7 2.66GHz CPU and 4GB memory. For all the considered models (except NAIVE) we set the probability of making an error on an independently provided value $\epsilon$=0.5; moreover for DEP and M-DEP we set the a-priori probability of dependence between two data sources $\alpha$=0.2 and the percentage of copied values over all values provided by a copier $c$=0.1.

## 3.1 Experimental Settings

For each domain, we downloaded pages from the Web and extracted the data by means of automatically generated wrappers manually refined to assure the correctness of the extraction rules. The attributes extracted were Height, Weight, and BirthDate for soccer players; Publisher and ESRB for video games; Price, Open Value, and Volume for stock quotes. Overall we collected 50,900 pages: statistics for the extracted data are reported in Figure 4. For each domain we produced the correct (true) values for a set of 861 stock quotes, 200 video games and 100 soccer players. Object were selected randomly, making sure that both popular and rare objects were part of the set. We believe that this is an important requirement, as famous objects are more likely to be curated and updated by the web sites maintainers. For the video game and stocks quote domains, the true values were collected by means of their authoritative sources, www.nasdaq.com and www.esrb.com respectively, which are the sites of the official organizations always providing correct information. The authoritative source for the stock quote domain is part of the set of sources considered for the experiment, but we keep it out of the set for the video games scenario. We will discuss later how the presence of the authority in the input set can impact the performance of the models. For soccer players, since an authoritative source does not exist, the true values of the considered attributes (Height, Weight, and BirthDate) were manually collected by inspecting the official web site of every soccer player whenever available, and the official web site of his current team club, otherwise. In any case, in the soccer domain the sources providing the true value of the players are not part of the set of sources considered for the experiments.

## 3.2 Evaluation Metrics

Given the truth vector $T = [t_1, \ldots, t_n]$ of correct values, for our experimental evaluation of the quality of web data, we define the *sampled accuracy* as the fraction of true values correctly reported by the site over the number of objects in the truth vector for which it publishes a value. For example, suppose we want to compute the sampled accuracy of a soccer web site $w$ reporting Height values for 1000 objects. We match this set of objects with the true values for Height in $T$ and identify 80 soccer players in the intersection of the two sets. We can now compute the sampled accuracy $\overline{a_i}$ for the source $i$: if, for example, the values reported by the source coincide with values in $T$ for 40 objects, than we estimate that the source reports true values for 50% of the cases and therefore $\overline{a_i}$=0.5. We compute in a similar way the sampled accuracy $a_i^m$ for every evaluated model $m$, the only difference is that the set of values matched with $T$ is the one made by the most probable values computed by $m$. In other words, even if a model returns a probability distribution for a value, we also consider the most probable one. We then obtain that a model can be treated as a single source and we can compute its sampled accuracy.

We rely also on two metrics, called *Probability Concentration* (PC) and *Accuracy Distance* (AD), to measure the performances of the models in computing the probability distributions and the accuracy of the sources, respectively.

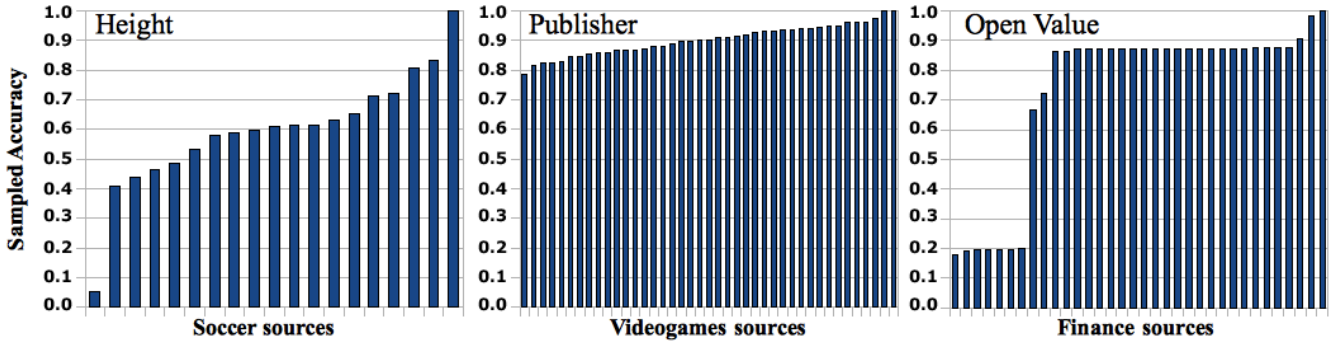


**Figure 5: The sampled accuracies of the sources of Open Value for stock quotes, Publisher for video games, and Height for soccer players.**

### *Probability Concentration (PC)*

The *Probability Concentration* measures the performance of the models in computing the probability distributions for $n$ observed objects. Given the truth vector $T$, the probability concentration for the model $m$ is the average probability associated to the correct value:

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

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

### *Accuracy Distance (AD)*

In order to measure the average quality of the accuracies computed for the sources, for each attribute we compare the sampled accuracy $a^m$ computed by every probabilistic model $m$ with the sampled accuracy $\overline{a}$ of the $k$ sources considered:

$$AD(m) = \frac{1}{k}\sum_{i=1}^{k} |a_i^m - \overline{a}_i|.$$

Note that if the estimated values for the accuracy of the sources are identical to the actual ones, then $AD$ equals to 0.

## 3.3 Accuracy of Web Data Sources

Figure 5 shows sampled accuracies for each source exposing values for it, fixed an attribute for every domain. The results show how attributes from distinct domains may assume quite different behaviours. Conversely, the sampled accuracies of the other attributes in the same domain behave quite similarly, and therefore we do not depict them here.

Overall, the average source accuracy is 70.21% for the soccer domain, 85.78% for stock quotes, and 89.22% for video games. It is important to observe that the sampled source accuracy seems to be better for domains where at least an authority exists. For example, the video game's Publisher exhibits high source accuracy (more than 78%) for every source, while in the case of the soccer players' Height and Weight the source accuracies are sensibly lower in all the sources. It is also worth noting that the better accuracy is expected for information that does not change over time (as Publisher for games).

In the finance domain, it can be observed that the source accuracies reflect the presence of clusters of sources that take their data from the same data providers. The source accuracy for the Open Value seems to be affected by two main factors: different sites publish the same value with a different number of digits, and the semantics of this attribute is sometimes confused with a very closely related attribute, that is, the price of the first trade.

In the soccer domain there is a peculiar case, a site whose sampled source accuracy is around zero: by manually inspecting that site, we observed that the site publishes randomly generated data about soccer players and that the published data change every time a page is reloaded.[3]

| | Alexa-links | Alexa-traffic | Pagerank |
|---|---|---|---|
| **Truth (Open Value)** | 0,15 | -0,22 | 0,5 |
| **Truth (Publisher)** | -0,04 | 0,05 | 0,11 |
| **Truth (Height)** | 0,13 | -0,31 | 0,21 |

**Figure 6: The correlation between of Alexa-Incoming-Links, Alexa-Traffic, and Google-PageRank with the sampled source accuracies is negligible. A value of 1 indicates a perfect positive relationship, while a value of -1 represents a perfect negative relationship.**

To evaluate how of popular ranking models for the Web, namely Google-PageRank[6], Alexa-Incoming-Links, Alexa-Traffic[4], relate to the accuracy of the sources, we computed the Pearson correlation coefficient[5] between these ranking indices and the data accuracy we computed for the web sites. As shown in Figure 6 the correlation coefficients of these ranking models with the data accuracy of the sources are negligible. These results suggest that the quality of the data

[3] e.g., http://soccer.azplayers.com/players/R/Ronaldo
[4] http://www.alexa.com
[5] A coefficient that represents the relationship between two variables that are measured on the same interval. It is defined as the covariance of the two variables divided by the product of their standard deviations.

| | Sampled Accuracy | | | | Probability Concentration | | | | Accuracy Distance | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | NAIVE | ACCU | DEP | M-DEP | NAIVE | ACCU | DEP | M-DEP | NAIVE | ACCU | DEP | M-DEP |
| **Birthdate** | 0.98 | 0.97 | 0.98 | 0.98 | 0.82 | 0.97 | 1.00 | 1.00 | 8.58 | 2.28 | 2.28 | 2.72 |
| **Height** | 0.66 | 0.67 | 0.67 | 0.67 | 0.51 | 0.66 | 0.67 | 0.67 | 4.6 | 13.86 | 13.93 | 16.49 |
| **Weight** | 0.59 | 0.66 | 0.66 | 0.66 | 0.49 | 0.67 | 0.67 | 0.67 | 6.37 | 10.53 | 10.76 | 12.11 |
| **ESRB** | 0.89 | 0.88 | 0.89 | 1.00 | 0.88 | 0.88 | 0.89 | 0.98 | 5.52 | 9.21 | 9.02 | 0.99 |
| **Publisher** | 0.97 | 0.97 | 0.98 | 1.00 | 0.92 | 0.97 | 0.98 | 1.00 | 2.62 | 1.50. | 1.33 | 0.07 |
| **Price** | 0.96 | 0.95 | 0.96 | 0.96 | 0.95 | 0.95 | 0.95 | 1 | 2.1 | 2.3 | 3.18 | 1.58 |
| **Open Value** | 0.97 | 0.95 | 0.97 | 0.97 | 0.73 | 0.95 | 0.96 | 0.96 | 11.01 | 9.56 | 9.74 | 8.39 |
| **Volume** | 1 | 1 | 1 | 1 | 0.83 | 1 | 1 | 1 | 13.12 | 0 | 0 | 0 |
| **AVG** | 0.88 | 0.88 | 0.89 | 0.91 | 0.77 | 0.88 | 0.89 | 0.91 | 6.74 | 6.16 | 6.28 | 5.29 |

**Figure 7: The Sampled Accuracy (higher is better), the Probability Concentration (higher is better), and the Accuracy Distance (lower is better).**

exposed by web sites is not reflected by these popular indices.

## 3.4 Experiments with Probabilistic Models

Figure 7 reports how the probabilistic models NAIVE, ACCU, DEP, and M-DEP perform by using the three metrics introduced before.

To compute the *Sampled Accuracy* for each model, a vector of candidate true values proposed by the model is needed. Such a vector is obtained by considering as candidate true values the ones that have the highest probabilities according to the computed probability distribution functions.[6] Its results show that on average all models are able to identify the correct values with a quite high precision in most cases. As expected, more complex models present better results for all domains, but surprisingly they all perform with similar results, with the best model, M-DEP, outperforming the simplest model, NAIVE, only by a 3.5% on average. It is also worth noting that score for all models are high for the domains where an authoritative source exists, but this is not directly related to the presence of such authority in the input set of sources. In fact, notice that in the reported results the authority was part of the input only for the stock quote scenario. Moreover, all models performed only slightly better in the video games scenario with an alternative input set containing also the authoritative source.

The *Probability Concentration* measures how much probability a model has concentrated on the known true value. In this case we assume that the NAIVE model propose probability distribution functions that merely reflect the frequencies of the observed values. This metrics shows that on average the most complex models ACCU, DEP, and M-DEP constantly outperform NAIVE by 14.2%, 15.5%, and 18.2%, respectively, as tools for estimating probability distribution functions. Again, we notice that the most sensible differences are in the domain where an authoritative source exists.

Finally, the *Accuracy Distance* metrics measures the distance between the sources quality estimation made by a model and their real sampled accuracy. The result shows that on average NAIVE is only marginally outperformed by the other models in estimating the correct accuracies of the input sources. Surprisingly, NAIVE shows comparable or even better results for half of the attributes (e.g., Height, Weight, Price). The most sensible improvements we observe with more complex models are in the stock quote attributes plus Birthdate. This behavior is due to the very different characteristic of the observed data as reported in Figure 4. In the stock quotes scenario, there is a very large number of distinct symbols in the alphabet and a larger number of instances. Also Birthdate has a much larger alphabet than Height or Weight. We discuss the role of the alphabet in more detail in the following.

To conclude our analysis, we examine the execution times for the models considered in the evaluation. Unsurprisingly, Figure 8 shows that NAIVE is always the faster, requiring only a few seconds for all the scenarios. Complex models perform equally well for video games attributes, but are slower in the soccer players scenario, and significantly slower in the stock quotes domain. It is easy to notice that this behavior reflects again the complexity of the data in terms of symbols in the alphabet and number of instances.

## 3.5 Discussion of the results

In general, achieving a good accuracy in the estimation of the most probable values is the easy part in characterizing the uncertainty of web data and is generally performed well by all approaches. We then conclude that for applications requiring only a good estimate of the true values, even a simple approach such as the NAIVE model can be used. However, whenever a more precise characterization of the uncertainty is required, for example to populate a probabilistic database, complex models exhibit significant advantages. In particular, over the three domains complex models show an improvements up to 36.7% for the *Probability Concentration* w.r.t. NAIVE; similar and better results can be observed for the *Accuracy Distance* estimation.

When deciding on what technique to use, we can distinguish three dimensions that can guide the choice: the characteristics of the domain that is used, the requirements on the result, the execution times.

1. If an authoritative source exists for the domain of interest, all models obtain very good results, even if the authority is not part of the input data. However, if the data of interest have large alphabets, complex models are likely to obtain better results. The intuition behind this behavior is quite simple: if the domain of possible values is very large and two sources present the same value it is very unlikely that they agree by chance. Either the value is correct or the fact that they agree on a wrong value is a very strong evidence

[6] If multiple values have the same maximum probability, a random value among them is chosen.

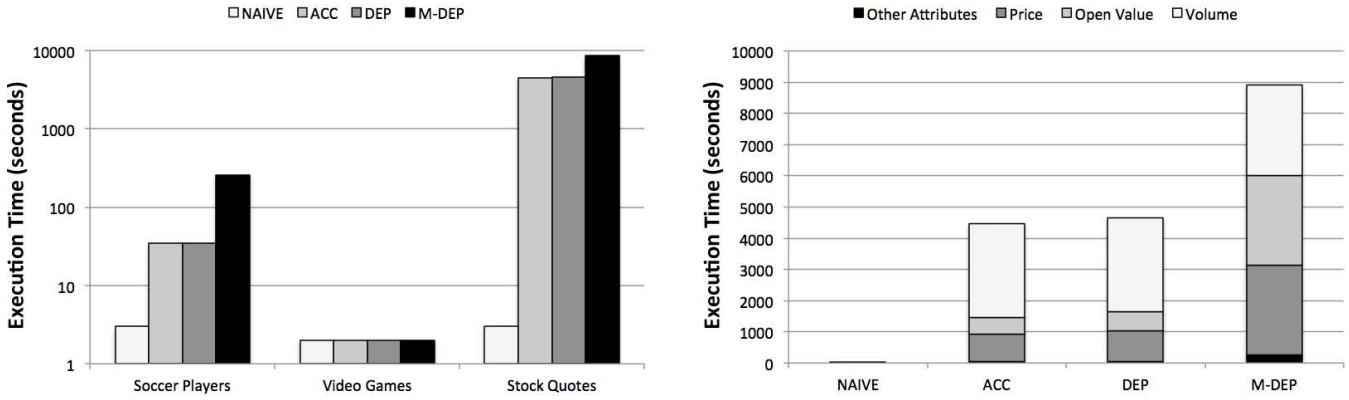


**Figure 8: On the left: execution times for the models over the three domains (logarithmic scale). On the right: detailed analysis of the execution times for different attributes.**

that they are copiers. As in general identifying copies improves the quality of the results, we can conclude that large alphabets lead to better results.

2. On average, more complex models guaranteed results with better quality. This is particularly evident for the estimation of the Probability Concentration and accuracy of the quality of the sources. But, if the desired output is only the most probable value, then the simple model in many cases returns satisfying results.

3. Execution times depends on the number of instances and the size of the alphabet for the attribute of interest. Experimental activity seems to suggest that only with very large alphabets the execution times of complex models become very expensive (up to hours). In settings with strict efficiency requirements it is crucial to examine carefully the characteristics of the data.

# 4. RELATED WORK

Our work explores the application of probabilistic techniques to assign truth probabilities to values gathered from conflicting sources of information. Such information is then used to evaluate the quality of the web sources in terms of their data accuracy. The problem is related to the broader context of data quality [2] and to the issue of combining probability distributions expressed by a group of experts, which has been studied in the statistics community (e.g., [8]).

Many projects have recently been active in the study of imprecise databases and have achieved a solid understanding of how to represent and process uncertain data (see [9] 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 [13] 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 [14].

On the contrary, only recently there has been some 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 [7], but they do not consider the problems of combining different probability distributions and evaluating the reliability of the sources.

TruthFinder [17] was the first project to address the problem of truth discovery in the presence of multiple web sources providing conflicting information. TruthFinder considers both the consensus on values and the accuracy of sources, and it can be considered as the first work that realizes and exploits their mutual dependency. Based on some heuristics, an iterative algorithm computes the trustiness of values and the accuracy of the sources. A similar direction has been also explored by [16] and [15] which present fix-point algorithms to estimate the true value of data reported by a set of sources, together with the accuracy of the sources.

Some of the intuitions behind TruthFinder were formalized in a probabilistic Bayesan framework by Dong *et al.* [11], who also considered how the presence of copiers (i.e. sources that copy from other sources) affects the evaluation of the source accuracy. While in TruthFinder the effects of possible copying dependencies between sources are handled by means of a simple heuristic, the authors of [11] develop a more principled approach to detect source dependencies. To achieve these results, their model (which corresponds to the DEP model illustrated in Section 2) computes the probability that a pair of sources are dependent by analyzing the co-occurrences of errors. A further variant by the same authors also consider the variations of truth values over time [12]. This latter investigation can lead to identify outdated sources and its first application in a real world scenario shows promising results.

The model behind DEP has been extended to improve the quality of the source dependencies detection. In fact, in [11] 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 video games. However, it is assumed that objects are described by just one attribute, e.g. the publisher of a video game. On the contrary, data sources usually provide complex data, i.e. collections of tuples with many attributes, and it has been shown that by considering this information the quality of the results can be improved [5, 10].

Notice that in our evaluation we tested for the first time

all the models on three common datasets.[7] In fact, evaluations of the above proposals were mainly conducted on synthetic data sets because, as noticed in [15], real-world data sets are hard to find, since they should be annotated with the real truth value in order to carry out the evaluation. Both TruthFinder and the models described in [11] were experimented just on one real data set composed by a collection of data about computer science books taken from `www.abebooks.com` (by the authors of TruthFinder), with the goal of discovering the books' authors. TruthFinder was also evaluated on a data set composed by a collection of data (the runtime) about movies. The algorithms described in [15] were experimented on a data constructed from the data published by a web-based prediction market: the data set was composed by users' answers on a given topic. Also in [5] the authors tested their solution with only one finance scenario, while in [10] authors used both the `www.abebooks.com` dataset and another dataset containing weather data.

An experimental comparison of authority and quality results for web sites has been done in [1]. Our work differs from this study in two important points. First, in our comparison against common popularity metrics we exploit the accuracy of the data offered by the web sources, while they compare quality in term of human judgement provided by experts. Second, we study the effectiveness of statistical models for the automatic evaluation of the sources, without requiring any user interaction.

# 5. CONCLUSIONS

In this paper we have presented an experimental evaluation of state-of-the-art techniques for assessing the quality and accuracy of web data sources. We then used evaluation metrics to compare the considered models on three real-life datasets taken from the web and manually cured to guarantee precision in the results.

Our evaluations suggest that sophisticated models always compute better results than simple voting strategies, but the decision on which model to use in an application should be done only after an analysis of the desired requirements and of the characteristics of the domain of interest.

# 6. REFERENCES


[1] B. Amento, L. G. Terveen, and W. C. Hill. Does "authority" mean quality? predicting expert quality ratings of web documents. In *SIGIR*, pages 296–303, 2000.

[2] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies, and Techniques*. Springer-Verlag, 2008.

[3] L. Blanco, M. Bronzi, V. Crescenzi, P. Merialdo, and P. Papotti. Exploiting information redundancy to wring out structured data from the web. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 1063–1064, New York, NY, USA, 2010. ACM.

[4] L. Blanco, M. Bronzi, V. Crescenzi, P. Merialdo, and P. Papotti. Redundancy-driven web data extraction and integration. In *WebDB*, 2010.


[7] Our real-world data sets, as well as our implementation of the models are available on request.


[5] L. Blanco, V. Crescenzi, P. Merialdo, and P. Papotti. Probabilistic models to reconcile complex data from inaccurate data sources. In *CAiSE*, pages 83–97, 2010.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks*, 30(1-7):107–117, 1998.

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

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

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

[10] X. Dong, L. Berti-Equille, Y. Hu, and D. Srivastava. Global detection of complex copying relationships between sources. *PVLDB*, 3(1):1358–1369, 2010.

[11] X. L. Dong, L. Berti-Equille, and D. Srivastava. Integrating conflicting data: The role of source dependence. *PVLDB*, 2(1):550–561, 2009.

[12] X. L. Dong, L. Berti-Equille, and D. Srivastava. Truth discovery and copying detection in a dynamic world. *PVLDB*, 2(1):562–573, 2009.

[13] D. Downey, O. Etzioni, and S. Soderland. A probabilistic model of redundancy in information extraction. In *IJCAI*, pages 1034–1041, 2005.

[14] D. Florescu, D. Koller, and A. Y. Levy. Using probabilistic information in data integration. In *VLDB*, pages 216–225, 1997.

[15] A. Galland, S. Abiteboul, A. Marian, and P. Senellart. Corroborating information from disagreeing views. In *Proc. WSDM*, New York, USA, 2010.

[16] M. Wu and A. Marian. Corroborating answers from multiple web sources. In *WebDB*, 2007.

[17] X. Yin, J. Han, and P. S. Yu. Truth discovery with multiple conflicting information providers on the web. *IEEE Trans. Knowl. Data Eng.*, 20(6):796–808, 2008.