Maximum Entropy in Support of Semantically Annotated Datasets

Paulo Pinheiro da Silva, Vladik Kreinovich, and
Christian Servin

Department of Computer Science
University of Texas, El Paso, TX 79968, USA
paulo@utep.edu, vladik@utep.edu

Abstract. One of the important problems of semantic web is checking whether two datasets describe the same quantity. The existing solution to this problem is to use these datasets' ontologies to deduce that these datasets indeed represent the same quantity. However, even when ontologies seem to confirm the identify of the two corresponding quantities, it is still possible that in reality, we deal with somewhat different quantities. A natural way to check the identity is to compare the numerical values of the measurement results: if they are close (within measurement errors), then most probably we deal with the same quantity, else we most probably deal with different ones. In this paper, we show how to perform this checking.

Key words: semantic web, ontology, uncertainty, probabilistic approach, Maximum Entropy approach

Checking whether two datasets represent the same data: formulation of the problem. In the semantic web, data are often encoded in Resource Description Framework (RDF) [2]. In RDF, every piece of information is represented as a triple consisting of a *subject*, a *predicate*, and an *object*. For example, when we describe the result of measuring the gravitation field, the coordinates at which we perform the measurements for a subject, a predicate is a term indicating that the measured quantity is a gravitational field (e.g., a term *hasGravityReading*), an the actual measurement result is an object.

In general, an RDF-based scientific dataset can be viewed as a (large) graph of RDF triples. One of the hard-to-solve problems is that triples in two different datasets using the same predicate *hasGravityReading* may not mean the same thing just because the predicates have the same name. One way to check this is to use semantics, i.e., to specify the meanings of the terms used in both datasets by an appropriate ontology, and then use reasoning to verify that the meaning of the terms is indeed the same. In the gravity example, we conclude that the predicate *hasGravityReading* has the same meaning in both datasets if in both datasets, this meaning coincides with *sweet:hasGravityReading*, the meaning of this term in one of the the Semantic Web for Earth and Environmental Terminology (SWEET) ontologies [3] that deals with gravity.

Need to take uncertainty into account. Even when ontologies seem to infer that we are dealing with the same concept, there is still a chance that the two datasets talk about slightly different concepts. To clarify the situation, we can use the fact that often, the two datasets contain the values measured at the same (or almost the same) locations. In such cases, to confirm that we are indeed dealing with the same concept, we can compare the corresponding measurement results $x'_1, \ldots, x'_n$ and $x''_1, \ldots, x''_n$. Due to measurement uncertainty, the measured values x'_i and x''_i are, in general, slightly different.

The question is: *Based on the semantically annotated measurement results and the known information about the measurement uncertainty, how can we use the uncertainty information to either reinforce or question whether two datasets namely representing the same data may not be the same data.*

Probabilistic approach to measurement uncertainty. To answer the above question, we must start by analyzing how the measurement uncertainty is represented. In this paper, we consider the traditional probabilistic way of describing measurement uncertainty.

In the engineering and scientific practice, we usually assume that for each measuring instrument, we know the probability distribution of different values of measurement error $\Delta x'_i \stackrel{\text{def}}{=} x'_i - x_i$. This assumption is often reasonable, since we can *calibrate* each measuring instrument by comparing the results of this measuring instrument with the results of a "standard" (much more accurate) one. The differences between the corresponding measurement results form the sample from which we can extract the desired distribution.

Often, after the calibration, it turns out that the tested measuring instrument is somewhat *biased* in the sense that the mean value of the measurement error is different from 0. In such cases, the instrument is usually re-calibrated – by subtracting this bias (mean) from all the measurement results – to make sure that the mean is 0. Thus, without losing generality, we can also assume that the mean value of the measurement error is 0: $E[\Delta x'_i] = 0$.

The degree to which the measured value x'_i differs from the actual value x_i is usually measured by the *standard deviation* $\sigma'_i \stackrel{\text{def}}{=} \sqrt{E[(\Delta x'_i)^2]}$.

Gaussian distribution: justification. The measurement error is usually caused by a large number of different independent factors. It is known that under certain reasonable conditions, the joint effect of a large number of small independent factors has a probability distribution which is close to Gaussian; the corresponding results (*Central Limit Theorems*) are the main reason why Gaussian (normal) distribution is indeed widely spread in practice [4]. So, it is reasonable to assume that the distribution for $\Delta x'_i$ is Gaussian.

Towards a solution. We do not know the actual values x_i, we only know the measurement results x'_i and x''_i from the two datasets. For each i, the difference between these measurement results can be described in terms of the measurement errors: $\Delta x_i \stackrel{\text{def}}{=} x'_i - x''_i = (x'_i - x_i) - (x''_i - x_i) = \Delta x'_i - \Delta x''_i$. It is reasonable to

assume that this difference is also normally distributed. Since the mean values of $\Delta x_i'$ and $\Delta x_i''$ are zeros, the mean value of their difference Δx_i is also 0, so it is sufficient to find the standard deviation $\sigma_i = \sqrt{V_i}$ of Δx_i. In general, for the sum of two Gaussian variables, we have $\sigma_i^2 = (\sigma_i')^2 + (\sigma_i'')^2 + 2r_i \cdot \sigma_i' \cdot \sigma_i''$, where $r_i = \dfrac{E[\Delta x_i' \cdot \Delta x_i'']}{\sigma_i' \cdot \sigma_i''}$ is the correlation between the i-th measurement errors. It is known that the correlation r_i can take all possible values from the interval $[-1, 1]$: the value $r_i = 1$ corresponds to the maximal possible (perfect) positive correlation, when $\Delta x_i'' = a \cdot \Delta x_i' + b$ for some $a > 0$; the value $r_i = 0$ corresponds to the case when measurement errors are independent; the value $r_i = -1$ corresponds to the maximal possible (perfect) negative correlation, when $\Delta x_i'' = a \cdot \Delta x_i' + b$ for some $a < 0$. Other values correspond to imperfect correlation. The problem is that usually, we have no information about the correlation between measurement errors from different datasets.

First idea: assume independence. A usual practical approach to situations in which we have no information about possible correlations is to assume that the measurement errors are independent.

A possible (somewhat informal) justification of this assumption is as follows. Each correlation r_i can take any value from the interval $[-1, 1]$. We would like to choose a single value r_{ij} from this interval.

We have no information why some values are more reasonable than others, whether non-negative correlation is more probable or non-positive correlation is more probable. Thus, our information is invariant with respect to the change $r_i \to -r_i$, and hence, the selected correlation value r_i must be invariant w.r.t. the same transformation. Thus, we must have $r_i = -r_i$, thence $r_i = 0$. A somewhat more formal justification of this selection can be obtained from the Maximum Entropy approach; see, e.g., [1]. Under the independence assumption, we have $(\sigma_i)^2 = (\sigma_i')^2 + (\sigma_i'')^2$.

Once we know the values, we can use the χ^2 criterion (see, e.g., [4]) to check whether with given degree of confidence α, the observed differences are consistent with the assumption that these differences are normally distributed with standard deviations σ_i: $\sum\limits_{i=1}^{n} \dfrac{(\Delta x_i)^2}{(\sigma_i)^2} \le \chi^2_{n,\alpha}$. If this inequality is satisfied, i.e., if $\sum\limits_{i=1}^{n} \dfrac{(\Delta x_i)^2}{(\sigma_i')^2 + (\sigma_i'')^2} \le \chi^2_{n,\alpha}$, then we conclude that the two datasets indeed describe the same quantity. If this inequality is not satisfied, then most probably, the datasets describe somewhat different quantities.

On the other hand, there is another possibility: that the two datasets do describe the same quantity, but the measurement errors are indeed correlated.

An alternative idea: worst-case estimations. If the above inequality holds for some values σ_i, then it holds for larger values σ_i as well. To take into account the possibility of correlations, we should only reject the similarity hypothesis when the above inequality does not hold even for the largest possible values σ_i.

Since $|r_i| \le 1$, we have $(\sigma_i)^2 \le V_i \stackrel{\text{def}}{=} (\sigma_i')^2 + (\sigma_i'')^2 + 2\sigma_i' \cdot \sigma_i''$. The value V_i is attained for $\Delta x_i'' = -\dfrac{\sigma_i''}{\sigma_i'} \cdot \Delta x_i'$. So, the largest possible value of σ_i^2 is equal to V_i. One can easily check that $V_i = (\sigma_i' + \sigma_i'')^2$. Thus, in this case, if $\sum\limits_{i=1}^{n} \left(\dfrac{\Delta x_i}{\sigma_i' + \sigma_i''}\right)^2 \le \chi^2_{n,\alpha}$, then we conclude that the two datasets indeed describe the same quantity. If this inequality is not satisfied, then most probably, the datasets describe somewhat different quantities.

Conclusion. Based on the semantically annotated measurement results and the known information about the measurement uncertainty, how can we use the uncertainty information to either reinforce or question whether two datasets namely representing the same data may not be the same data?

We assume the some values from the two datasets contain the results of measuring the same quantity at the same locations and/or moments of time. Let n denote the total number of such measurements, let $x_1', \ldots, x_n'$ denote the corresponding results from the first dataset, and let $x_1'', \ldots, x_n''$ denote the measurement results from the second dataset. We assume that we know the standard deviations σ_i' and σ_i'' of these measurements, and that we have no information about possible correlation between the corresponding measurement errors. In this case, we apply the Maximum Entropy approach, and conclude that if $\sum\limits_{i=1}^{n} \dfrac{(\Delta x_i)^2}{(\sigma_i')^2 + (\sigma_i'')^2} \le \chi^2_{n,\alpha}$, where $\chi^2_{n,\alpha} \approx n$ is the value of the χ^2-criterion for the desired certainty α, then this reinforces the original conclusion that the two datasets represent the same data. If the above inequality is not satisfied, then we conclude that either the two datasets represent different data (or, alternatively, that the measurement uncertainty values σ_i' and σ_i'' are underestimated).

If we have reasons to suspect that the measurement errors corresponding to two databases may be correlated, then can be more cautious and reinforce the original conclusion even when a weaker inequality is satisfied: $\sum\limits_{i=1}^{n} \left(\dfrac{\Delta x_i}{\sigma_i' + \sigma_i''}\right)^2 \le \chi^2_{n,\alpha}$.

Acknowledgments. This work was partly supported by NSF grant HRD-0734825 and by NIH Grant 1 T36 GM078000-01. The authors are thankful to the anonymous referees for valuable suggestions.

References

1. Jaynes, E. T.: Probability Theory: The Logic of Science, Cambridge University Press (2003)
2. Resource Description Framework (RDF) http://www.w3.org/RDF/
3. Semantic Web for Earth and Environmental Terminology SWEET ontologies http://sweet.jpl.nasa.gov/ontology/
4. Sheskin, D.: Handbook of Parametric and Nonparametric Statistical Procedures, Chapman & Hall/CRC, Boca Raton, Florida (2004)