

# Probabilistic Model-Based Assessment of Information Quality in Uncertain Domains<sup>\*</sup>

Steffen Michels, Marina Velikova, and Peter J.F. Lucas

Institute for Computing and Information Sciences  
Radboud University Nijmegen, The Netherlands  
`s.michels@science.ru.nl`

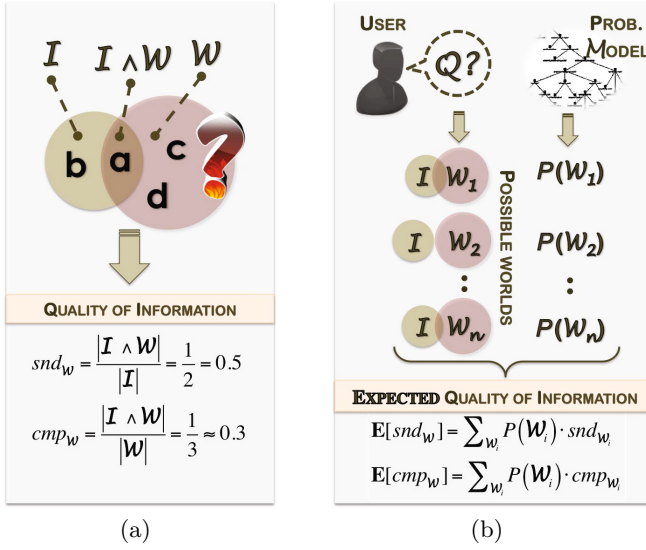
**Abstract.** In various domains, such as security and surveillance, a large amount of information from heterogeneous sources is continuously gathered to identify and prevent potential threats, but it is unknown in advance what the observed entity of interest should look like. The quality of the decisions made depends, of course, on the quality of the information they are based on. In this paper, we propose a novel method for assessing the quality of information taking into account uncertainty. Two properties – *soundness* and *completeness* – of the information are used to define the notion of information quality and their expected values are defined using a probabilistic model output. Simulation experiments with data from a maritime scenario demonstrates the usage of the proposed method and its potential for decision support in complex tasks such as surveillance.

## 1 Introduction

Information is collected virtually in any domain and task to increase our knowledge about reality and to provide bases for decisions. The quality of these decisions depends, of course, on the quality of the information they are based on and the estimation of the latter has been the focus of research for many years. Very early work can be found in the science of experimental measurement, where information is largely collected in a controlled setting with priorly defined expected outcomes, and information quality is typically quantified by statistical measures such as systematic and statistical errors. Later the field of information retrieval emerged based on survey studies and database management tasks, where measures such as precision and recall are used to determine the goodness of answers extracted with respect to a pre-defined question [8]. These measures are clearly defined with respect to ground-truth information, assumed to be provided by an expert or another information source with certainty [3]. Studies in information management have further defined information quality by a large variety of dimensions, e.g., completeness, accuracy, relevancy, timeliness, as summarised

---

<sup>\*</sup> This publication was supported by the Dutch national program COMMIT. The research work was carried out as part of the Metis project under the responsibility of the Embedded Systems Institute with Thales Nederland B.V. as the carrying industrial partner.



**Fig. 1.** Two approaches for estimating the quality of information  $\mathcal{I}$  in terms of soundness ( $snd$ ) and completeness ( $cmp$ ): a) standard method used in information retrieval, assuming that the hypothetical true world  $\mathcal{W}$  is known, and b) our model-based method yielding expected values based on probabilistic estimates of the possible true worlds  $\mathcal{W}_i$  generated from a set of user queries  $Q$

in [7]. The definitions of these dimensions include uncertainty but no methods for the practical assessment of the dimensions are provided.

In recent years, we observe wide availability and heterogeneity of information sources in new domains, such as security, surveillance and object recognition, where one has to handle a continuous stream of information and it is unknown in advance what the observed entity of interest should look like. This leads to *uncertainty* in the information obtained and in the estimation of its quality. Hence, new measures are required for defining and assessing information quality such that uncertainty is explicitly taken into account. The notions and probabilistic method we propose in this paper are a contribution in this direction.

The representation language of our method is based on first-order logic with binary predicates, which allow a straightforward expression of object-attribute-value relations, typically referred as facts in the information obtained from sources. We define the notion of “information quality” in terms of two properties – *soundness* and *completeness* – that are based on the estimation of set cardinalities. The conceptual idea is similar to the ones proposed earlier in [2,6], but we propose a novel formalisation and estimation of these properties taking into account uncertainty, as illustrated in Figure 1, namely:

- We define the intersection set of the information gathered and the ground-truth with respect to a set of user queries and not to a fixed database as

done before. Hence, we consider “quality” as a dynamic property of the information, which depends on the questions at hand.

- Unlike the related studies, we do not assume that the true world is provided, for example by a human or a database, but we generate the set of possible true worlds based on the set of user queries. Then for each world we compute the probability for being true using a predefined model using domain knowledge and other information. The probabilities obtained are finally used to compute the expected value for soundness and completeness of the information gathered. Here, we do not restrict the structure or the nature of the model, as long as it provides a proper probability distribution of the possible world states.

The proposed method is general and can be applied to any domain where uncertainty of information plays a role, but its development was motivated by our research on estimating information quality in maritime surveillance. Therefore, we next present a scenario from the maritime domain to motivate this research. Section 3 presents the language and main notions used to define information quality. The main theoretical contribution—probabilistic method for assessment of information quality—is described in Section 4. Experiments described in Section 5 illustrate the application of the method. Section 6 concludes the paper.

## 2 Motivating Scenario

The detection of drug smuggling is one of the biggest challenges that maritime surveillance and the coast guard operators nowadays face. Typical hypotheses that increase the suspiciousness of a ship are falsified/missing ship identity information or the type of cargo that allows to hide drugs easily (e.g., bulk dry). The operators need to check the quality of information reported by the ship via the automatic identification system (AIS) or radio contact in their daily tasks to examine these hypotheses.

The Dutch bulk-carrier “*Seaman*” approaches the port of Rotterdam carrying cement, sand and stone, and sends the following AIS information to the coast guard:

**i:** ID = 23378, name = *Seaman*, type = *bulk-carrier*, cargo = {*sand*, *gravel*}.

An intelligence report of the same day warns an operator that a bulk carrier, most likely carrying cement and registered in Bahamas, is expected to arrive within a day at the Rotterdam port and it is suspected for drugs smuggling. To decide whether the *Seaman* may be the suspected vessel, the operator wants to examine the quality of information reported by the ship with respect to various questions:

- Does *Seaman* carry cement?
- Is *Seaman* a bulk-carrier registered in Bahamas?

To get answers to these questions, the operator uses additional information from other sources such as a ship registry database and visual observations and finds out that the *Seaman* is a Dutch bulk-carrier carrying also cement and stone but not gravel on board.

Thus, the operator concludes that the *Seaman* has reported incorrect and incomplete cargo, prompting some attention for further monitoring.

Performing such manual check-ups in practice, however, would make the work of operators practically impossible. Furthermore, some sources may not be easily accessible or their information not fully trusted, thus making the conclusions uncertain. Therefore, an automated approach is needed to evaluate the quality of information inherent with uncertainty.

### 3 Preliminaries

#### 3.1 Representation Language

We use *first-order logic* as a flexible way to represent information and formulate queries. This abstracts from how data is stored in an actual implementation which could for instance be in some kind of database system. Our logic consists of a countable set of *predicate symbols*. We do not use functions, only constants like *seaman*. This restriction is chosen to ensure that domains are finite, as is common in probabilistic logics. We use upper case labels for variables, like  $X$ . We further use the common logical connectives  $\forall$ ,  $\exists$ ,  $\wedge$ ,  $\vee$ ,  $\implies$  and  $\iff$  to build up formulas.

We assume that information consists of logical facts described as object-attribute-value relations, which are represented as binary predicates in the form  $a(O, V)$ , where  $a$  refers to an attribute of object  $O$  and  $V$  is the attribute value. We do this without loss of generality as binary predicate logics is as expressive as full predicate logic. In the paper we use the name of objects, like *seaman*, as object labels for better readability. Adding new information means taking the conjunction of the previous and the new one. Instantiation of  $O$  and  $V$  means assigning a value to them. We explicitly do not use a *closed world assumption*, but distinguish wrong information from not reported information which is true in the real world.

#### 3.2 Information Quality

In Table 1 we define the notions for representing information used in the rest of the paper. Although (sets of) queries and the true world also represent information, for the remaining of the paper we will use the term *information* only for what has been reported, thus defining and estimating its quality. For the meta properties we use to define information quality we use upper case labels for variables, like  $X$  and lower case labels for instantiations ( $x$ ). We use bold notation for variables and instantiations representing sets ( $\mathbf{X}$  and  $\mathbf{x}$ ). We finally use logic entailment, denoted by  $\models$ , to indicate that a query is positively answered by given information.

Ideally, all queries positively answered by the information one has are true in the real world, too. The other way around, everything true in the real world should be derivable from the information. So we consider information quality as

**Table 1.** Definition of notions for representing information

| Notion                      | Definition                                                                                                                        | Scenario examples                                                                                                                                                                                                                                                                      |
|-----------------------------|-----------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Information $I$             | A self-contained statement about the world that can be expressed by a logical ground sentence (a formula without free variables). | $i_1 = \text{type}(\text{seaman}, \text{tanker})$<br>$\wedge \text{name}(\text{seaman}, \text{seaman})$<br>$\wedge \text{id}(\text{seaman}, 23378)$<br>$i_2 = \text{cargo}(\text{seaman}, \text{sand})$<br>$\wedge \text{cargo}(\text{seaman}, \text{gravel})$<br>$i = i_1 \wedge i_2$ |
| Query $Q$                   | A logical ground sentence, representing a question, which is positively answered if it is entailed by the information             | $q_1 = \text{cargo}(\text{seaman}, \text{sand})$<br>$q_2 = \text{type}(\text{seaman}, \text{bulk-carrier})$<br>$\wedge \text{flag}(\text{seaman}, \text{Bahamas})$                                                                                                                     |
| Set of queries $\mathbf{Q}$ | It can be abbreviated by a formula with a free variable, if all possible groundings of the variable are known.                    | $\mathbf{q}_1 = \{q_1, q_2\}$<br>$\mathbf{q}_2 = \text{cargo}(\text{seaman}, \text{TYPE})$<br>$\mathbf{q}_3 = \{\text{cargo}(\text{seaman}, \text{cement}),$<br>$\text{cargo}(\text{seaman}, \text{sand}),$<br>$\text{cargo}(\text{seaman}, \text{stone})\}$                           |
| World $\mathcal{W}$         | The information representing the state of the real world.                                                                         | $\mathcal{W} = \{\text{flag}(\text{seaman}, \text{NL}),$<br>$\text{type}(\text{seaman}, \text{bulk-carrier}),$<br>$\text{cargo}(\text{seaman}, \text{cement}),$<br>$\text{cargo}(\text{seaman}, \text{sand}),$<br>$\text{cargo}(\text{seaman}, \text{stone}), \dots\}$                 |

a two dimensional property. We introduce measures of how much those desired meta properties are true, referred to as *soundness* and *completeness* of information  $I$  with respect to a set of queries  $\mathbf{Q}$  and they are formalised as follows:

$$\text{snd}_{\mathcal{W}}(\mathbf{Q}, I) \equiv \frac{\left| \left\{ Q \in \mathbf{Q} \mid I \models Q \wedge \mathcal{W} \models Q \right\} \right|}{\left| \left\{ Q \in \mathbf{Q} \mid I \models Q \right\} \right|} \quad (1)$$

$$\text{cmp}_{\mathcal{W}}(\mathbf{Q}, I) \equiv \frac{\left| \left\{ Q \in \mathbf{Q} \mid I \models Q \wedge \mathcal{W} \models Q \right\} \right|}{\left| \left\{ Q \in \mathbf{Q} \mid \mathcal{W} \models Q \right\} \right|} \quad (2)$$

In the case the denominator is 0 the soundness or completeness becomes 1. The values for soundness and completeness range from 0 to 1, where 0 is the worst and 1 the best possible quality. The soundness intuitively measures how much of what the information tells is actually true, the completeness measures how much of what is true is included in the information. In the special case that the set contains one query only, the result is always 0 or 1. The completeness never decreases when additional information  $\mathbf{I}'$  is added to the information  $\mathbf{I}$ , whereas for soundness such a general property cannot be stated.

*Example 1.* The soundness and completeness of  $i_2$  given  $\mathbf{q}_3$  defined in Table 1 are:

$$\text{snd}_{\mathcal{W}}(\mathbf{q}_3, i_2) = \frac{1}{1} = 1 \quad \text{cmp}_{\mathcal{W}}(\mathbf{q}_3, i_2) = \frac{1}{3} \approx 0.33$$

The definitions in (1) and (2) are closely related to the notions of precision and *recall* in information retrieval, and soundness and completeness as defined in [2,6], which are also based on the comparison of set cardinalities. The main difference in our definitions is that we construct the sets with respect to the set of queries. In other words, the elements of the sets depend on whether a query is entailed by the information and/or the true world, instead of directly comparing the information and the true world as done in the other approaches.

## 4 Probabilistic Method for Assessing Information Quality

In practice the entire state of  $\mathcal{W}$  is often not known, especially in the domain we deal with. Previous work [2] for instance solves this by using data for which the world is assumed to be true, for example based on expert evaluation. There are two problems with this approach. First, the quality of received information cannot be assessed instantaneously in an automated way. Second, it cannot deal with uncertainty about how the world looks like.

Therefore, instead of requiring hard decisions of what is true or not in the real world, we construct a set of all possible worlds, which is restricted only to the relevant part given by all groundings of predicates relevant for the set of queries  $\mathbf{Q}$ . We denote this set of possible worlds by  $\mathbf{W}_{\mathbf{Q}}$ .

*Example 2.* The possible parts of all relevant worlds given  $\mathbf{q}_3$  are:

$$\begin{aligned} \mathbf{w}_{\mathbf{q}_3} = \{ & \text{cargo}(\text{seaman}, \text{cement}) \wedge \text{cargo}(\text{seaman}, \text{sand}) \wedge \text{cargo}(\text{seaman}, \text{stone}), \\ & \text{cargo}(\text{seaman}, \text{cement}) \wedge \text{cargo}(\text{seaman}, \text{sand}) \wedge \neg \text{cargo}(\text{seaman}, \text{stone}), \\ & \dots \} \end{aligned}$$

Now for every possible world  $W \in \mathbf{W}_{\mathbf{Q}}$  we want to compute the probability for being true. To achieve this, we use a probabilistic model  $\mathcal{M} = \langle \mathcal{S}, P \rangle$ , where  $\mathcal{S}$  denotes the structure of the model and  $P$  is a probability distribution such that  $P : \mathbf{W}_{\mathbf{Q}} \rightarrow [0, 1]$ . The probability of a possible world is denoted by  $P(W)$  and we have  $\sum_{W \in \mathbf{W}_{\mathbf{Q}}} P(W) = 1$ . In general,  $P$  is the joint probability distribution of the specific grounding terms in  $W$ .

Building such a probabilistic model can itself be a complex task, but there is a large body of work about how to build such models by incorporating expert knowledge or learning from data. Here we do not restrict the structure of the model, as long as the model produces a valid probability distribution, but in Section 4.2 we discuss some possible alternatives. A strong point of our approach is that once such a model is build, it can be reused for different vessels and sets of queries.

### 4.1 Expected Information Quality

We can then use  $\mathcal{M}$  to derive the *expected soundness* and *completeness* of a set of queries, as defined in (1) and (2). For this we sum over all possible worlds and

parametrise  $snd$  and  $cmp$  with a possible world. We use the notation  $snd_W$  and  $cmp_W$  replacing  $\mathcal{W}$  by the possible world  $W$  in (1) and (2), respectively:

$$\mathbf{E}[snd_{\mathcal{W}}(\mathbf{Q}, I)] \equiv \sum_{W \in \mathbf{W}_{\mathbf{Q}}} P(W) \cdot snd_W(\mathbf{Q}, I) \quad (3)$$

$$\mathbf{E}[cmp_{\mathcal{W}}(\mathbf{Q}, I)] \equiv \sum_{W \in \mathbf{W}_{\mathbf{Q}}} P(W) \cdot cmp_W(\mathbf{Q}, I) \quad (4)$$

Note that if the probability distribution  $P$  gives a probability of 1 for a possible world then the formulas for the expected values of soundness and completeness reduce to those in (1) and (2), respectively.

*Example 3.* In this example we use  $i_2$ ,  $\mathbf{q}_3$  and  $\mathcal{W}$  from Table 1. Given  $\mathbf{q}_3$ , we generate the set of all relevant possible worlds  $\mathbf{W}_{\mathbf{q}_3}$  and we assume that a probabilistic model gives the probability  $P(W)$  of each possible world  $W$  being true; see the table below. This probabilities could for instance be frequencies obtained from past observations. We abbreviate the possible world by only giving the cargo the Seaman is carrying instead of giving the full predicates. To compute the expected values of soundness and completeness we count the number of queries  $q \in \mathbf{q}_3$  satisfying  $i_2 \models q \wedge W \models q$ ,  $i_2 \models q$  and  $W \models q$ , shortly abbreviated as  $|i_2 \wedge W|$ ,  $|i_2|$  and  $|W|$ .

| $W$                                                                | $P(W)$ | $ i_2 \wedge W $ | $ i_2 $ | $ W $ | $snd_W(\mathbf{q}_3, i_2)$ | $cmp_W(\mathbf{q}_3, i_2)$ |
|--------------------------------------------------------------------|--------|------------------|---------|-------|----------------------------|----------------------------|
| $\neg\text{cement} \wedge \neg\text{sand} \wedge \neg\text{stone}$ | 0.56   | 0                | 1       | 0     | 0                          | 1                          |
| $\neg\text{cement} \wedge \neg\text{sand} \wedge \text{stone}$     | 0.06   | 0                | 1       | 1     | 0                          | 0                          |
| $\neg\text{cement} \wedge \text{sand} \wedge \neg\text{stone}$     | 0.07   | 1                | 1       | 1     | 1                          | 1                          |
| $\neg\text{cement} \wedge \text{sand} \wedge \text{stone}$         | 0.1    | 1                | 1       | 2     | 1                          | 1/2                        |
| $\text{cement} \wedge \neg\text{sand} \wedge \neg\text{stone}$     | 0.04   | 0                | 1       | 1     | 0                          | 0                          |
| $\text{cement} \wedge \text{sand} \wedge \neg\text{stone}$         | 0.09   | 1                | 1       | 2     | 1                          | 1/2                        |
| $\text{cement} \wedge \neg\text{sand} \wedge \text{stone}$         | 0.05   | 0                | 1       | 2     | 0                          | 0                          |
| $\text{cement} \wedge \text{sand} \wedge \text{stone}$             | 0.03   | 1                | 1       | 3     | 1                          | 1/3                        |

Using (3) and (4) the expected soundness and completeness are then:

$$\mathbf{E}[snd_{\mathcal{W}}(\mathbf{q}_3, i_2)] = 0.29 \quad \mathbf{E}[cmp_{\mathcal{W}}(\mathbf{q}_3, i_2)] = 0.735$$

We notice that the estimations are not close to the actual qualities computed in Example 1, as  $P(W)$  is based on prior distributions and does not account for additional information, for instance, that the vessel is a bulk-carrier.

An advantage of our approach is that it allows for a lot of flexibility in what one is interested in. One could for instance ask whether two types of cargo are on board of the Seaman or at least one of them.

*Example 4.* Some examples for more complex queries are given, together with the true and estimated qualities of  $i_2$ , given the estimates of Example 3:

| $\mathbf{Q}$                                                                                                    | $snd_{\mathcal{W}}(\mathbf{Q}, i_2),$<br>$cmp_{\mathcal{W}}(\mathbf{Q}, i_2)$ | $\mathbf{E}[snd_{\mathcal{W}}(\mathbf{Q}, i_2)],$<br>$\mathbf{E}[cmp_{\mathcal{W}}(\mathbf{Q}, i_2)]$ |
|-----------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|
| $\mathbf{q}_4 = \{\text{cargo}(\text{seaman}, \text{cement}) \wedge \text{cargo}(\text{seaman}, \text{sand})\}$ | 1.0, 0.0                                                                      | 1.00, 0.88                                                                                            |
| $\mathbf{q}_5 = \{\text{cargo}(\text{seaman}, \text{cement}) \vee \text{cargo}(\text{seaman}, \text{sand})\}$   | 1.0, 1.0                                                                      | 0.19, 1.00                                                                                            |

## 4.2 Probabilistic Model

Although in this paper we do not restrict ourselves to a particular type of a probabilistic model, below we discuss some alternatives for estimating the probability distribution of the set of possible worlds depending on available information.

**No Information—Uniform  $P(\mathbf{W}_Q)$ .** When no knowledge and information about the domain is given, the simplest assumption is that all possible worlds are equally likely:

$$\forall W \in \mathbf{W}_Q : P(W) = \frac{1}{|\mathbf{W}_Q|} \quad (5)$$

*Example 5.* Consider again Example 3. Given the assumption in (5), the expected soundness and completeness are  $\mathbf{E}[snd_W(\mathbf{q}_3, i_2)] = 0.5$  and  $\mathbf{E}[cmp_W(\mathbf{q}_3, i_2)] \approx 0.42$ .

In practice, this assumption is most likely violated, making the estimation of information quality less reliable.

**Estimate  $P(\mathbf{W}_Q)$  Using Prior Distributions.** Information is often available as statistics over a population, which can be used as prior distribution for estimating  $P(\mathbf{W}_Q)$ , when no other specific information is available. For example, in Example 3, one can derive the distribution by taking the fraction of vessels carrying the particular set of cargoes out of the bulk dry vessels documented at a particular port.

Although prior distributions could be a good first step for estimating  $P(\mathbf{W}_Q)$ , we note that this may be suboptimal since the quality of information about unlikely cases will be considered low, even if the quality is good; this is illustrated in the examples in Section 5.2.

**Update  $P(\mathbf{W}_Q)$  Using New Information.** If new information becomes available, in terms of probability theory, we can condition on this information to update the estimate of  $P(\mathbf{W}_Q)$ .

*Example 6.* Continuing Example 3, the prior distribution estimate can be improved by using the type of the vessel. Next to the type adding also information obtained from a visual observation that the vessel definitely carries sand, further updates the probability distribution of the possible worlds, as shown in the table below:

| $W$                                                                | $P(W)$ | $P(W \text{bulk\_carrier})$ | $P(W \text{bulk\_carrier, sand})$ |
|--------------------------------------------------------------------|--------|-----------------------------|-----------------------------------|
| $\neg\text{cement} \wedge \neg\text{sand} \wedge \neg\text{stone}$ | 0.56   | 0.06                        | 0.00                              |
| $\neg\text{cement} \wedge \neg\text{sand} \wedge \text{stone}$     | 0.06   | 0.02                        | 0.00                              |
| $\neg\text{cement} \wedge \text{sand} \wedge \neg\text{stone}$     | 0.02   | 0.08                        | 0.10                              |
| $\neg\text{cement} \wedge \text{sand} \wedge \text{stone}$         | 0.15   | 0.04                        | 0.05                              |
| $\text{cement} \wedge \neg\text{sand} \wedge \neg\text{stone}$     | 0.04   | 0.05                        | 0.00                              |
| $\text{cement} \wedge \text{sand} \wedge \neg\text{stone}$         | 0.09   | 0.03                        | 0.0375                            |
| $\text{cement} \wedge \neg\text{sand} \wedge \text{stone}$         | 0.05   | 0.07                        | 0.00                              |
| $\text{cement} \wedge \text{sand} \wedge \text{stone}$             | 0.03   | 0.65                        | 0.8125                            |

How the probabilities are updated depends on the model, which is not the focus in this paper but we note that the probabilities of  $P(W|\text{bulk\_carrier, sand})$  can be derived from the probabilities for  $P(W|\text{bulk\_carrier})$ . This example demonstrates that more

relevant information contributes to getting a distribution closer to the true world. Below are shown the actual qualities together with the expected ones using the different distributions:

| Model distribution                       | $snd_{\mathcal{W}}(\mathbf{q_3}, i_2) = 1$<br>$\mathbf{E}[snd_{\mathcal{W}}(\mathbf{q_3}, i_2)]$ | $cmp_{\mathcal{W}}(\mathbf{q_3}, i_2) = 0.33$<br>$\mathbf{E}[cmp_{\mathcal{W}}(\mathbf{q_3}, i_2)]$ |
|------------------------------------------|--------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|
| $P(W)$                                   | 0.29                                                                                             | 0.735                                                                                               |
| $P(W \text{bulk\_carrier})$              | 0.8                                                                                              | 0.392                                                                                               |
| $P(W \text{bulk\_carrier}, \text{sand})$ | 1                                                                                                | 0.415                                                                                               |

This shows that having more specific information leads to better estimates of the expected values of information quality.

**Derivation of  $P(W_{\mathbf{Q}})$  Using Independence Assumptions.** Learning the probabilities for all possible worlds may quickly become impractical because their number grows exponentially with the number of queries. One way to tackle this is to assume that the probability  $P(g)$  for a grounding  $g$  of a predicate, representing a query in  $\mathbf{Q}$ , is true is independent of all the others. In this case, only one estimated probability is needed for each query to define the model. The probability that a possible world is true can then be defined as:

$$P(W) \equiv \prod_{W \models g} P(g) \cdot \prod_{W \models \neg g} (1 - P(g)) \quad (6)$$

*Example 7.* Consider again Example 3. Suppose that the model cannot give a probabilistic estimate of each possible world as before but only to the grounding of each predicate in the queries, i.e.,

$$\begin{array}{ccc} P(\text{cargo}(\text{seaman}, \text{cement})) & \Big| & P(\text{cargo}(\text{seaman}, \text{sand})) & \Big| & P(\text{cargo}(\text{seaman}, \text{stone})) \\ 0.21 & & 0.29 & & 0.24 \end{array}$$

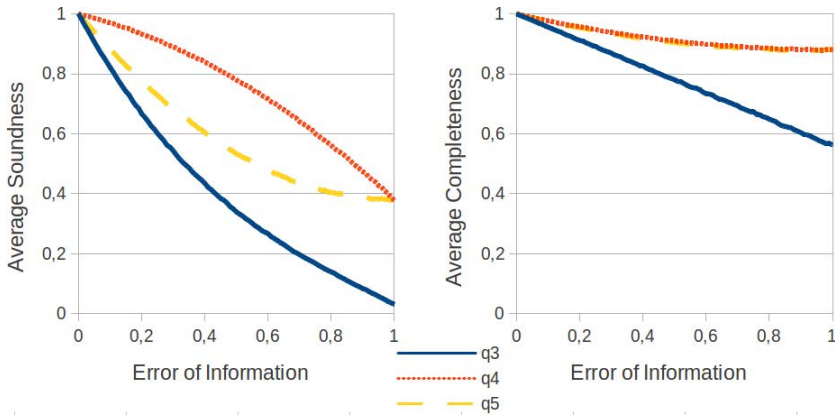
Here, the probabilities are computed as the marginals for the sake of comparison. Now using (6) and the three probabilities above, we compute the probabilities for each possible world. Then  $\mathbf{E}[snd_{\mathcal{W}}(\mathbf{q_3}, i_2)] = 0.29$  and  $\mathbf{E}[cmp_{\mathcal{W}}(\mathbf{q_3}, i_2)] \approx 0.66$ . The expected soundness is still the same, while the expected completeness changes.

Although in practice, the independence assumption may be too strict for all variables, it can still help reduce complexity while providing useful estimations.

Typical examples of probabilistic models that are able to incorporate domain knowledge, manage dependencies, and provide prior and posterior (joint) probability distributions include probabilistic graphical models such as Bayesian networks [5] and probabilistic logic [4,1].

## 5 Experiments

We perform two experiments to show that (i) the queries matter for estimating information quality (Exp-1), showing that our flexible approach to define queries is necessary, and (ii) a distribution conditioned on observations of the actual world is needed to give proper estimates (Exp-2).



**Fig. 2.** The effect of the query on the soundness and completeness of information

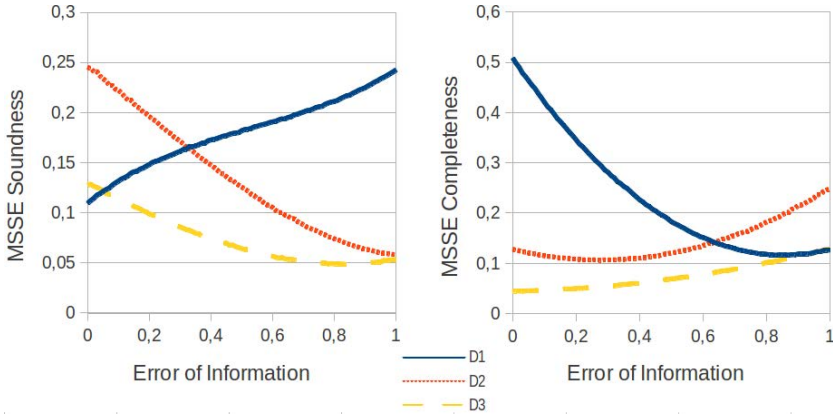
For both experiments we generate  $N$  samples using the distribution given in the table of Example 3. Each sample  $S_1, \dots, S_N$  therefore consists of three Boolean values indicating whether the ship carries cement, sand or stone. We fix  $N = 100000$  for all experiments. We simulate pieces of information  $I_1, \dots, I_N$  by observing whether a type of cargo is present in the actual world or not. The information can be wrong, which we simulate by introducing an error  $\epsilon$ , i.e. there is a probability  $\epsilon$  that the observation is wrong, which means that the cargo is reported while not present or the other way around. In the dynamic domain we deal with the actual error rates are unknown. We therefore vary  $\epsilon$  from 0 to 1 in steps of 0.01 in the experiments to get results for various error rates.

### 5.1 Exp-1: Effect of Query on Information Quality

The information quality decreases with more discrepancies between the information and the actual world, i.e., with increasing  $\epsilon$ . The rate of change of soundness and completeness given the error, however, depends on the query, as demonstrated in this experiment.

We compare the different sets of queries: **q3** from Table 1 and **q4**, **q5** from Example 4. For the different queries we determine the actual soundness and completeness  $snd_{S_i}(\mathbf{Q}, I_i)$  and  $cmp_{S_i}(\mathbf{Q}, I_i)$  for each sample and compute the averages. We repeat this for different  $\epsilon$ . Figure 2 shows the average soundness and completeness for the different queries against  $\epsilon$ .

The quality of information obviously decreases as the error increases. If there is no error the quality is perfect, but it is interesting that it never goes to 0 even if the reported information is completely wrong. For instance, the mean completeness for **q3** and  $\epsilon = 1.0$  is above 0.5. This is because according to the distribution in 56% of the cases none of the three cargo types is on board of a ship. The reported completely wrong information reports all three kinds of cargo, which is per definition complete. This is true for soundness in an analogous way.



**Fig. 3.** The effect of the distribution used to estimate information quality on the MSSE

The set of query containing the conjunction  $\mathbf{q}_4$  has a better soundness, as in less situations the query can be positively answered. In this case the soundness is 1.0. The difference between the completeness of  $\mathbf{q}_4$  and  $\mathbf{q}_5$  is very small. This is not a general property of the queries, but changes with a different distribution used to generate the samples.

## 5.2 Exp-2: Comparison of Different Distributions for Prediction

In this section we compare how different distributions perform when used to estimate the information quality. This shows that as already shown in Example 6 conditioning on some knowledge about the actual world is needed to get proper estimates. We only use  $\mathbf{q}_3$  and use the following distributions:

- $\mathcal{D}_1$ : the uniform distribution (Equation 6)
- $\mathcal{D}_2$ : the distribution used to generate the data (table of Example 3)
- $\mathcal{D}_3$ : the distribution used to generate the data conditioned on the observation whether the ship is carrying sand ( $P(W|\text{sand})$  or  $P(W|\neg\text{sand})$ )

For the different distributions we determine for each sample the actual and estimated soundness and completeness and compute the *mean sum of squared errors* (MSSE). We repeat this for different  $\epsilon$ ; see Figure 3.

The result clearly shows that using a distribution which is dynamically conditioned on information from the actual world, outperforms fixed distributions in nearly all cases. It might be surprising that the uniform distribution for some  $\epsilon$  works better than the one used to generate the data. We focus on the soundness for the case  $\epsilon = 0$  to give a better insight of why this is the case. In the following table we give for each pair of estimated and actual soundness the frequency of how often this pair occurs for all samples  $S_1, \dots, S_N$ :

|                                      | $\mathcal{D}_1$                                  |           | $\mathcal{D}_2$                                  |           |
|--------------------------------------|--------------------------------------------------|-----------|--------------------------------------------------|-----------|
| $snd_{\mathcal{W}}(\mathbf{q}_3, I)$ | $\mathbf{E}[snd_{\mathcal{W}}(\mathbf{q}_3, I)]$ | frequency | $\mathbf{E}[snd_{\mathcal{W}}(\mathbf{q}_3, I)]$ | frequency |
| 1.0                                  | 1.0                                              | 0.562     | 1.0                                              | 0.562     |
| 1.0                                  | 0.5                                              | 0.438     | $0.247 \pm 0.026$                                | 0.438     |

For the cases where no information is reported the information is always sound, so the difference is in the other cases. Using  $\mathcal{D}_1$  for those cases all predictions are 0.5 and for  $\mathcal{D}_2$  it is always smaller. Since 0.5 is close to the actual soundness 1.0,  $\mathcal{D}_1$  performs better for this case. This shows that a fixed distribution is not sufficient even if the actual distribution is perfectly known.

## 6 Conclusions

We propose a novel methodology for systematically representing and assessing the quality of information obtained from various sources in uncertain domains such as maritime surveillance. We give a concise, formal definition of information quality based on logical entailment and two sub-properties—soundness and completeness—indicating the goodness of answers derived from information compared to the ground truth. Unlike previous work the definition allows to consider information quality as a dynamic property of the information, which depends on the questions at hand. Also we do not require certain knowledge about reality and allow to deal with uncertainty, by proposing method to estimate information quality using a probabilistic model. An application on a maritime scenario demonstrated basis properties of the proposed method.

In future, we plan to study the effect of the estimates from the probabilistic model on the quality of information by developing a realistic model based on the maritime domain.

## References

1. Laskey, K.B.: MEBN: A language for first-order bayesian knowledge bases. *Artif. Intell.* 172(2-3), 140–178 (2008)
2. Motro, A., Rakov, I.: Estimating the Quality of Databases. In: Andreasen, T., Christiansen, H., Larsen, H.L. (eds.) *FQAS 1998. LNCS (LNAI)*, vol. 1495, pp. 298–307. Springer, Heidelberg (1998)
3. Naumann, F., Leser, U., Freytag, J.C.: Quality-driven integration of heterogenous information systems. In: *Proc. of the 25th Int. Conf. on VLDB*, pp. 447–458 (1999)
4. Nilsson, N.J.: Probabilistic logic. *Artif. Intell.* 28(1), 71–87 (1986)
5. Pearl, J.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann (1988)
6. Peim, M., Franconi, E., Paton, N.W.: Estimating the quality of answers when querying over description logic ontologies. *Data & Knowl. Eng.* 47(1), 105–129 (2003)
7. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. *Communications of the ACM* 45(4), 211–218 (2002)
8. Salton, G., McGill, M.J.: *Introduction to Modern Information Retrieval*. McGraw-Hill (1983)