BELIEVABILITY AS AN INFORMATION QUALITY DIMENSION

(Research-in-progress)
IQ Concepts

Shekhar Pradhan
Department of Computer Science, Marist College
shekhar.pradhan@marist.edu

Abstract: This paper argues that believability is a more important dimension of information quality than accuracy because of what is called the representation gap between information and data, where "data" is defined as what is entered into an information system and "information" is defined as what is returned by the system in response to decision level queries. It provides an analysis of believability that ties it to the philosophical field of epistemology. It identifies a type of believability, which is called procedural believability and discusses what implications this type of believability has for the design, maintenance, and use of information systems.

INTRODUCTION

It is widely accepted that information is a product manufactured ultimately out of the raw material of data. [9, 10] As with any product, information is produced to be "consumed" by users. There are certain attributes that users regard as desirable in the information product. One such attribute is believability, which is defined by [10] as "the extent to which data are accepted or regarded as true, real, and credible."

In the literature on information quality, the IQ attribute of believability has not received much study or analysis. In this paper, we seek to rectify this situation by proposing an analysis of believability.

We speculate that there may be two reasons why believability as an IQ dimension may not have received much attention. First, if believability is taken as the property of being believed by users, then the study of this property seems to fall within the domain of psychology rather than information science. Second, believability may be seen by some people as secondary to the IQ attribute of accuracy. For instance, [5] describes believability as expected accuracy. We address both of these points in turn.

PRELIMINARIES

To ground this discussion, we propose as a preliminary definition of believability: *the property of being **worthy** of being believed by rational and informed users.* We regard this as a better definition than the one quoted from [10] because it distinguishes between the property of *being believed* and the property of being *believable.* Just as *being trustworthy* does not mean being trusted, but being worthy of trust, so also with believability. Another advantage of this way of defining believability is that it allows us draw on philosophical work on believability. Of course, it requires further investigation to determine which properties of information make it worthy of being believed by rational and informed users. This investigation can be done empirically, as described below. But, in addition, there is a rich tradition of work in the branch of philosophy called epistemology which also provides many insights into this question.

In this paper we will distinguish between the terms “data” and “information.” By “data” we mean what is entered in an information system. We will use the term “information” to mean decision making level representations of reality. Typically, such decision level representation of reality is produced by an information system in response to decision level queries. Hence, we will sometimes use “information” to mean answers to decision level queries. For example, the data may be that certain sensors have displayed certain readings, whereas the information that can be obtained from this data is that there is a high level of carbon monoxide in a certain environment. The distinction is not absolute—what is data in one context can be information in another context. But information is a product made from data, where the product is relevant to decision making in some context.

ACCURACY AND BELIEVABILITY

Next we examine the relationship between accuracy and believability. It is commonly thought that of the two, accuracy is the more important attribute, both conceptually and in practice. Although different definitions of accuracy have been given in the literature, accuracy has been most commonly described as correctness of data. And correctness is commonly understood as correspondence with reality [10]. It is commonly assumed that a piece of information is a representation of some part of reality and that whether a piece of information is accurate is verified by direct observation of this aspect of reality [10]. An example of this way of looking at things would be the information that there are 37 sprockets in stock presently can be verified to be accurate by going to the stockroom and doing a count of the sprockets there.

If we look at accuracy in this way, then it is natural to ask why we need believability as a quality attribute. It would appear that if information has the attribute of accuracy, then it should be believed, at least, by those who accept that it has the attribute of accuracy. On the other hand, if information lacks the attribute of accuracy then it should not be believed, at least, by those who think it lacks the attribute of accuracy. So it would appear that believability of information comes into play only when it is not known whether that information is accurate or not. In this way the attribute of believability seems secondary to the attribute of accuracy.

Furthermore, the concept of believability seems to be dependent on the concept of accuracy. To believe some claim is to believe it to be true, i.e., accurate. Thus, it would seem that believability is a secondary quality attribute, both conceptually and operationally, which comes into play only when a determination of the primary quality attribute of accuracy cannot be made.

This account of the relationship between accuracy and believability needs to be re-examined for several reasons. First, we need to be clearer about the account of accuracy given above. Accuracy is described as correspondence with reality. This can be understood in terms of mapping. If there is a mapping from the components of the information system (pieces of data or information) to components of reality (facts), then the information (system) is accurate [9]. But for any set of sentences, there can be many mappings from those sentences to some set of facts or other. What matters is what logicians call the *intended* mapping or interpretation. So until we know the intended mapping we are not in a position to verify some information as accurate or not. And the intended mapping may be to a set of facts that cannot be easily verified.

Consider a typical context of use of information. Suppose management of a company wants to know how many of its employees are members of ethnic minorities. To keep this simple, let us suppose that in the company database there is a table EMPLOYEES which has a column that has YES if the employee is regarded as a minority member and NO otherwise. It is a simple matter to compute an answer to management’s query from this table. But management didn’t ask for how many of the entries in this table had YES in the appropriate field. That information can be directly verified to be accurate by examining the table. Management, or any consumer of information, has certain intended meaning of the terms they use in the query. In this case, “employee” and “ethnic minority”. These terms have certain complex legal and social signification. What the poser of the query wants to know is how many of our employees are ethnic minorities in this sense. It is an assumption that by adding the tuples with YES in the appropriate field, we arrive at the answer in the intended sense. The intended interpretation of the information stored in the EMPLOYEE table is the interpretation that captures this complex legal and social signification of these terms. The point is that verifying the accuracy of the claim that so and so is an employee of the company and is a member of an ethnic minority is not a simple matter of scanning the facts. It often involves complex consideration of conflicting and interacting arguments and evidence. The end result of such considerations is a judgment of how believable this claim is.

Lest it be thought that this example is atypical, let us take the more familiar example briefly described above. Suppose the query is: How many sprockets are in stock currently? Let us say we extract a certain number from the information stored in the information system as the answer to this query. We go to the stockroom and count the sprockets there. This should verify the accuracy of the information that there are a certain number of sprockets in stock. And normally it does. But that doesn’t alter the fact that taking count of the sprockets in the stock room is an argument for the believability of the claim made by the information system. To see this, imagine that there are employees who occasionally store their own sprockets in the stock room. These sprockets are not owned by the company. So they can’t be considered part of the company stock. Thus, the intended meaning of the query and the answer is the sprockets that are owned by the company and currently available for use. Now it is easy to see that counting the sprockets in the stockroom only provides a strong argument for the believability of the claim made by the information system.

These two examples show that there is a gap between what the user of information wants to know and what can be directly verified. We call this the *representation gap.* This gap is in practice filled by making a number of assumptions. Thus, in the sprocket case an assumption being made is that the sprockets in the stockroom are all owned by the company. Another assumption being made is that all the sprockets owned by the company which are currently available for use are in the stockroom. The correctness of these assumptions cannot be verified as accurate by a simple scan since the facts of ownership are complex and diffuse. But they can have a certain degree of believability associated with them.

That there should be a representational gap between decision-level queries and the data entered in information systems is hardly surprising. Decision-level queries are posed by agents embedded in a complex social, economic, legal, cultural, etc., environment and such agents employ concepts that reflect the complexity of this environment. On the other hand, the data entered in an information system gets its semantic and conceptual content by the data model used in designing the information system and the constraints that relate parts of this data model. This data model with its associated constraints is a very impoverished version of the conceptual structure used by the poser of decision-level queries. The gap between the conceptual structure underlying queries and the data model is bridged in practice by data collection methods, such as questionnaires, interviews, etc., which use a richer conceptual structure than the data model used by information systems. But generally there is no systematic, disciplined effort to ensure that the conceptual structure used by the data collection methods does effectively bridge the representational gap between information and data.

The point being made here, that there is a representational gap between information and data, is very familiar to philosophers of science [1]. Scientists pose high-level queries such as, “How many particles of such and such type are present in this chamber?” To answer these queries they set up experiments. What can be directly observed and what can be verified are readings on dials, marks on specially prepared material, changes in the color of certain solutions, etc. But such directly verifiable information counts as an answer to the query only under lots of assumptions. These assumptions are called auxiliary theories. These theories cannot be directly verified, but they have a certain degree of credibility or believability associated with them. Thus, the gap between what scientists, or users of information systems, want to know and what can be directly verified as accurate makes believability of information a more important attribute in practice than accuracy of information.

It is not our intent to argue that there is no such thing as accuracy. Rather, we want to establish that in practice believability of *information* is the primary attribute and accuracy is the secondary attribute. Information is *taken* to be accurate if it is determined to be fully believable. This is not an analysis of the concept of accuracy. Accuracy can be still be analyzed as correspondence with reality, or a mapping with reality. Our point is that except in very simple cases, when we are considering certain types of data as merely *data*, we don't have a direct access to reality—we discern reality through evidence and argumentation, which establishes believability. That is, we discern reality, or accuracy, through believability.

This point is even more evident in considering the products of decision support systems. The output of such systems are judgments such as, so and so is creditworthy, or so and so is a security risk. But what does it mean to say that a person is creditworthy? This means, that person is such that if money was loaned to him, he would repay the loan in a timely fashion. Creditworthy is what philosophers call a dispositional term, like brittle [4]. To say something is brittle is to say that if it is struck with certain degree of force it will break. This claim may be true even if the thing in question is never hit with the requisite force. This is what is called a subjunctive conditional. Philosophers and logicians have not been able to agree on the truth conditions of subjunctive conditionals, or of their close cousin, the contrary-to-fact conditional, such as "if this vase had been hit with force, it would have broken." Thus, it is not clear what facts correspond to claims involving dispositional terms. Where do we go to determine the accuracy of the claim that someone is creditworthy? We have some idea of what sorts of considerations would support the claim that someone is creditworthy and what sorts of considerations would undermine it. But these considerations can at most establish the believability of the judgment that a certain person is creditworthy.

Hence, in studying information quality we cannot neglect believability. It cannot be regarded as secondary to accuracy, or any other IQ attribute.

BELIEVABILITY

Let us now try to get clearer about believability.

Some researchers regard believability as an intrinsic property [10], whereas other regard it as contextual, meaning that the degree of believability of some information depends on the context of its use. It is certainly true that whether some information has an acceptable degree of believability is a matter of context, because what is regarded as acceptable is a matter of context. But that it has that degree of believability, whether acceptable or not, may not be dependent on the context of use.

In another sense, believability is a contextual property. The believability of information is relative not to the context of use, but to the context of evidence. How believable some information is depends on what other information we take into account in assessing its believability. This is clear in a court of law where what evidence is admissible heavily determines the outcome of a trial. But it equally holds in other contexts. Whether a rational and informed user regards some information as worthy of belief depends on what other information such a user has at his disposal in considering this question.

It is possible to argue that the believability of the information means believability to this or that user. In this sense the term believable is implicitly referential—it always contains a reference to some user. So the proper locution should be “believable to user X.” To say that information is believable to user X means the information has those set of properties, P_X, which makes that information worthy of belief in X’s judgment. These properties would include those properties which are peculiar to X (such as X’s need to believe it, or need not to believe it) and those properties which any user will regard as making that information worthy of belief. That is, the set P_X will contain a subset, P, which is invariant from user to user.

More formally, let U be a set of users, for each user x in U, let P_X denote those properties that x regards as relevant to determining whether some information is worthy of belief, then the set P is the intersection of the members of $\{ P_X \mid x$ is a member of $U\}$.

This user invariant set will contain exactly those properties which are essential to determining whether some information is believable to a rational user because, presumably, rational users don’t use their personal needs or exigencies to determine whether something is worthy of belief. And these user invariant properties may properly be regarded as intrinsic features of believability.

In analyzing the concept of believability we have used the concept of a rational and informed user. We have provided an explanation above of what we mean by a rational user in terms of the user invariant set of properties regarded as relevant by any user to determining the believability of information. By an informed user we mean a user who has the requisite information to determine whether some information has those properties.

What might be such intrinsic aspects of believability? And how do we determine them? An empirical study can be devised to make this determination. A sample group of users can be surveyed to determine what properties must some body of information have for them to regard that information as worthy of belief? The properties common to all the answers are the likely candidates for inclusion as the intrinsic, user-invariant set of properties that a rational user would regard as making some information worthy of belief. We plan to conduct such a study in future research.

Another way of determining the intrinsic, user-invariant properties relevant to believability is through philosophical analysis. Philosophy has a rich tradition of analysis of the justification of belief [8]. The question they have asked is what makes some beliefs justified. In other words, what makes some beliefs worthy of believing? This is similar to our question: what makes some information worthy of belief?

There is a consensus in the philosophical literature that some of the factors relevant to the justification of a belief is the balance of evidence for or against that belief and the causal mechanism by which a belief was produced. However, there is much disagreement about the type of causal mechanism, its precise role in justification, and how it interacts with the evidential aspect of justification [8].

Using this idea, we can distinguish two aspects to the believability of information. *Content believability* is believability attaching to data or information as a result of the balance of evidence for and against it. *Procedural believability* is believability attaching to information as a result of the process by which that information was produced.

Pradhan [6, 7] has investigated how answers to a query to a distributed information system can be annotated with a confidence value which represents the degree of believability of that information taking into account all the evidence for and against that information, and taking into account all the evidence for or against the information that was used as evidence, and so on. That is, his method takes into account all the interacting evidential relationships that exist in a body of information to determine the overall evidence for the answers to a query. He has done this by casting the relations of evidence for and against as a system of arguments of varying degrees of strength. Since we regard information as answers to a decision-level query, his work can be used to determine the degree of content believability attaching to any information relative to the data or information in some information system.

In this paper we next investigate the role of the causal process by which information is produced as a contributing factor to the believability of information. That is, we investigate procedural believability.
We do not investigate in this work how content believability and procedural believability can be combined to give a measure of the believability of information. This will be addresses in future work.

INFORMATION PRODUCTION AND BELIEVABILITY

Recall that we defined information as the answers produced in response to decision-level queries. We defined data as what is entered into the information system. Now we examine the causal process by which information is produced as a contributor to the believability of that information. This causal process can be broken down into three steps:

a. The process by which data is entered into the system.
b. The process of translating a decision-level query into a query in a query language, such as SQL, and translating back the answers to the query language query as answers to the decision-level query.
c. All the processes involved in between data entry and query formulation, such as schema normalization, integrity constraint formulation, transaction processing (which would include query answering, query optimization, constraint checking, concurrency control, etc.).

Naumann and Roth [5] contains an excellent discussion of part c and its relation to IQ dimensions.

As we pointed out earlier there is a gap between the answers returned to a decision-level query, with its intended interpretation, and the data contained in the information system. This gap is bridged, we said, by making a number of assumptions about the process by which information is produced out of data. Any defect in any part of this process can affect the believability of the answers returned to the decision level query.

Naumann and Roth [5] discuss the principle *garbage in, garbage out* and the principle *quality in, quality out*. But there is a third possibility, which is *quality in, garbage out.* This is possible if there is any defect in any of the steps in information production process described above. They claim that the processes involved in the production of answers to queries posed to commercial DBMS are so reliable at this point that if the database contains quality data then it produces quality answers (*quality in, quality out).* We believe that this is correct with regard to the part c of the process described above, but not necessarily true with regard to parts a and b.

The process of entering data is not just a question of entering correct data, but entering it with the intended interpretation. This is achieved to some extent by entering the data into a table with a schema which has the intended interpretation. This intended interpretation can be partly captured by giving the schema a certain semantics in terms of integrity constraints such as foreign key constraints and assertions. But in many contexts the resources of integrity constraints are just not adequate to capture the intended meaning of the data entered. Thus, suppose a certain tuple is entered into a table called EMPLOYEE with a certain schema. But the intended meaning of “employee” might make a distinction between employees and consultants, who may also be on the payroll of the company. Thus, the process of entering data into the employee table has to ensure that data about that person is entered into that table only if he has a certain type of legal contract with the company. It may not be possible to ensure this by using a set of integrity constraints. Our point is that the data may be correct as far as it goes (it has quality), but may still result in garbage when used to answer queries about employees.

The process of translating a decision level query into a query language query and translating the answers to this latter query back into answers for the former query is even more liable to be defective. Consider the following example.[1] Management of a company wants to know how many of a certain part are available for use (stock on hand). This query can be understood in three ways:

i. How many of the part are on the shelf?

ii. How many of the parts are in the plant even if it left the shelf, but didn't get already used in a product?

iii. How many of the part are on the shelf, but haven't already been committed to a product.

Clearly, the same query (How many of such and such part are available for use?) can receive at least three different formulations in the query language, and the resulting answer may be garbage even though the data is quality if the query was not translated with the intended meaning.

Thus, the extent to which the process of producing information, understood as answers to decision-level queries, can be certified as non-defective to that extent that information can be regarded as worthy of belief.

BELIEVABILITY AND SYSTEM DESIGN

Information systems can be designed so as to enhance procedural believability. Our emphasis on the representation gap between information and data suggests practices that should be used in the development, maintenance, and use of an information system which will improve the procedural believability of the information produced by that information system. We propose below in a highly sketchy and preliminary manner the practices that should be used in the development, maintenance, and use of an information system.

Requirements for an information system should be gathered through use cases [3] which consist in posing typical queries to the information system by potential users. These queries should be posed in English as opposed to any sort of formal query language. The key concepts in these queries should be identified. These concepts must be analyzed to clarify the meaning of these concepts and to identify any ambiguities in these concepts. For example, ambiguity in the term "available for use" in the example described above. These concepts and their analysis should be stored in a semantic dictionary, which will have much richer information than data dictionaries.

Next it should be determined what sort of data should be stored in the system and how it should be organized. In considering any potential data, we must identify the representation gap between the information we seek from the system (determined by the user queries) and the data to be stored. Then we must identify the procedures that will bridge the representational gap and implement these procedures. Some of these procedures will consist in formulating integrity constraints which will give a semantics to the schema which is as close as possible to the semantics of the concepts which are used in the user queries. If there is any residual representation gap then the additional procedures used to identify these gaps must be specified. Typically, these procedures will pertain to what we have earlier called data entry and data collection. For instance, if some data is to entered in the EMPLOYEE table and the tuples stored in the EMPLOYEE table is to be used to answer user queries about employees in a certain intended sense specified in the dictionary then the residual representational gap between this sense of "employee" and the semantics of EMPLOYEE (as determined by the integrity constraints) must be filled by data collection constraints and practices. Thus, for instance, if users distinguish between employees of a company and consultants that are hired and there is no adequate way of capturing the distinction between them in the integrity constraints over schema, then before entering some information in the EMPLOYEE table there must be some appropriate process for determining that this information is about an employee as opposed to a consultant.

[1] We owe this example to Craig Fisher, who assures us that this situation actually occurred when he worked for a certain well known company.

Next, we describe practices to be followed in the posing of decision level queries to the system. If a decision level query contains a term defined in the semantic dictionary, then the definition of the term in that dictionary should be used to formulate the query in the query language. If the term is has several different definitions, then the user level query should be disambiguated before translation into query language.

To some extent, these practices are obvious and already followed in the design and use of information systems. But clearly identifying the specific features of the representational gap between information and data of a system aids in highlighting just which of these practices are critical to enhancing procedural believability of the information generated by the system.

CONCLUSIONS

In this paper we have argued that although data can many times easily be verified to be accurate, in general decision-level information cannot be. This is because there is a representational gap between data and information. This gap is bridged by the assumption that certain connections hold between the data and the information. Except in unusual circumstances, there is no direct way of verifying that these connections hold. So these connections can only be regarded as more or less believable. Thus, decision-level information can only be regarded as more or less believable.

We have introduced a distinction between content believability and procedural believability. We have described how procedural believability of information produced by an information system depends on the specific procedures implemented to bridge the representational gap between data contained in the system and the information produced by the system. Finally, we discussed the implications of this work for the design, use, and maintenance of information systems.

In future extensions of this work we plan to do the following work:

i. Design and carry out an empirical study to determine what properties are relevant to making information worthy of belief by rational and informed users.

ii. Devise functions for combining the content believability and procedural believability of some information so as to have a measure of the overall believability of this information.

iii. Since believability is a matter of degrees, we need a way of representing the degree of belief of information. We plan to do that.

iv. Since several pieces of information, each with its own degree of belief can be combined, a calculus for combining degrees of belief is required. We plan to do that.

v. Devise belief assurance practices, similar to software quality assurance practices[2], based on the procedural believability enhancing practices we have identified above for design, use, and maintenance of information systems.

REFERENCES

[1] Curd, M., Cover, J.A. *Philosophy of Science: The Central Issues*, W.W. Norton and Company, 2005.

[2] Galin, D. *Software Quality Assurance*, Pearson Addison-Wesley, 2004

[3] Larman, C. *Applying UML and Patterns, 2nd Edition*, Prentice-Hall, 2003

[4] Mumford, S. *Dispositions*, Oxford University Press, 1998.

[5] Naumann, F., Roth, M. "Information Quality: How Good are Off-the-shelf DBMS?" *proceedings of the 9th International Conference on Information Quality*, 1994, pp. 260-274.

[6] Pradhan, S. "Connecting Databases with Argumentation." *Web Knowledge Management and Decision Support*, Springer-Verlag, LNAI 2543, Berlin, 2003. pp.170-185

[7] Pradhan, S. "Argumentation Databases." *Logic Programming: Proceedings of ICLP 2003*, Springer-Verlag, LNCS 2916, Berlin, 2003. pp. 178-193

[8] Swinburne, R. *Epistemic Justification*, Oxford University Press, 2001.

[9] Wand, W., Wang, R.Y. "Anchoring Data Quality Dimensions in Ontological Foundations." *Communications of the ACM*, 39(11).1996. pp.86-95

[10] Wang, R.Y., Strong, D.M. "Beyond Accuracy: What Data Quality Means to Data Consumers." *Journal of Management Information Systems*, 12(4). 1996. pp.5-34