UNCERTAINTY MANAGEMENT IN INFORMATION SYSTEMS

FROM NEEDS TO SOLUTIONS

EDITED BY

Amihai Motro
George Mason University
Fairfax, VA, USA

Philippe Smets
Université Libre de Bruxelles
Brussels, Belgium

KLUWER ACADEMIC PUBLISHERS
Boston/London/Dordrecht

1

INTRODUCTION

Amihai Motro

Department of Information and Software Systems Engineering
George Mason University
Fairfax, VA 22030, USA

1 SCOPE

As its title suggests, "Uncertainty Management in Information Systems" is a book about how information systems can be made to manage information permeated with uncertainty. This subject is at the intersection of two areas of knowledge: *information systems* is an area that concentrates on the design of practical systems that can store and retrieve information; *uncertainty modeling* is an area in artificial intelligence concerned with accurate representation of uncertain information and with inference and decision-making under conditions infused with uncertainty.

New applications of information systems require stronger capabilities in the area of uncertainty management. Our hope is that lasting interaction between these two areas would facilitate a new generation of information systems that will be capable of servicing these applications. Although there are researchers in information systems who have addressed themselves to issues of uncertainty, as well as researchers in uncertainty modeling who have considered the pragmatic demands and constraints of information systems, to a large extent there has been only limited interaction between these two areas.

As the subtitle, "From Needs to Solutions," indicates, this book presents viewpoints of information systems experts on the needs that challenge the uncertainty capabilities of present information systems, and it provides a forum to researchers in uncertainty modeling to describe models and systems that can address these needs.

The following subsections describe somewhat more formally the scope of this book. It should be noted that the term *uncertainty* has been in use as both a *broad* term covering any number of situations in which the information available is short of perfection (and we use it in this interpretation in the title of this book), and to describe a *specific* kind of imperfection (where the correctness of the available information is in doubt). In general, we try to use *imperfect information* for the broad term; otherwise, it should be clear from the context whether uncertainty is used in its broad or specific meaning.

1.1 Imperfect Information

Consider an information system D that attempts to represent some part of the real world W, and assume that both D and W are denoted in the same modeling formalism; for instance, both D and W could be relational databases.[1] We may consider W as the perfect information, and D as its *approximation*. This information D is *complete*, if D "contains" W; that is, the stored information includes *all* the information that the system presumes to model. D is *sound*, if D is "contained" in W; that is, the stored information includes *only* true information. Roughly speaking, an information system is perfect if it includes the *whole* truth and *nothing but* the truth.

In most practical situations, information systems rarely meet this standard of perfection. When an information system is not sound and complete, the common practice is to *ignore* the imperfections; that is, to continue operation as if the system were sound and complete. This implies that queries might be answered inaccurately: answers might exclude correct information, and some of the information they include might be incorrect.

Often, however, the negative consequences of inaccurate answers can be reduced by improving the modeling techniques used in the approximation of W.

Assume an initial model without any features for representing anything but sound and complete information. An example of such a model is this simple binary data model, in which, a database of employees is a set of employee numbers, each identifying a unique employee, and several two-column tables that store information about employees; for instance, *(Employee_number, Name)*, *(Employee_number, Department)*, and *(Employee_number, Age)*. The semantics of this database are that the set of employee numbers is sound and complete (i.e., it agrees perfectly with the employee numbers of the actual set of employ-

[1] Of course, only D is available; W is only hypothetical.

ees), and that each database pair provides accurate and reliable information about a specific employee. Thus, the pair (9128,30) in the table *Age* indicates that the age of the employee whose number is 9128 is indeed 30.

Consider now these examples.

- Assume that the age of a particular employee is not known precisely, only that it is in the range 25–40. In the model just described, the corresponding age pair would be excluded from the database. Yet, if the model is extended to allow such imprecise information, the available information could be stored. A query that lists all the employees older than 40 could be answered perfectly, and a query to list the age of this employee could provide this range. In both cases, the answers are more informative.
- Assume now that the age of an employee is given as 35, but there are some doubts about this information. For example, it might be estimated that with probability 0.8 the age is 35, but with probability 0.2 it is some other value. If the model is extended to include such uncertain information, the available information could be stored, and a query regarding the age of this employee could be answered more informatively.
- Finally, assume that the age of the male employees is known precisely, but the age of the female employees is correct only in an estimated 75% of the cases. Moreover, it is known that an estimated 5% percent of the temporary employees are not even recorded in the database. Again, if the model is extended to include such quality information, then many queries could be answered more informatively.

These examples illustrate three situations in which the information that is available is not sound or complete. Yet, rather than completely exclude all information that is short of perfection, it is advantageous to include in the system whatever relevant knowledge is available.

In the first example, the model extensions would enable the system to represent *imprecise* information; in the second, the system would be extended to represent *uncertain* information; and in the third, the system would take into consideration the *quality* of the information (statements that *qualify* the soundness and completeness of D).

These three kinds of extensions, respectively address three major kinds of imperfect information. Yet, there are various other flavors of imperfect information, and Chapter 8 provides a thorough analysis of the subject. In fact, many other chapters begin with a classification, simple or intricate, of the various kinds of imperfections that are addressed.

1.2 Manipulating Imperfect Information

Of course, once an information model provides structures for *representing* imperfect information, it must also show how to process *retrieval requests* against such information.

Returning to our abstract representation with D and W, we may define a retrieval request as a function m that derives from the description D, another description D', which is the answer to the request. Then, the definition of m must be extended to the case where D describes imperfect information.

In our simple binary model, assume that retrieval requests are issued with this simple statement: **retrieve** *attribute* **where** *attribute condition value*, where both attributes are from the same two-column table. For example, **retrieve** *Age* **where** *Employee_number* = 9128, or **retrieve** *Employee_number* **where** *Age* $\geq$ 45. If the model is extended to allow imprecise or uncertain information, then the processing of such queries would have to be extended as well. For example, in the latter query, would employee number 6310 be returned, if the age is given as in the range 40–50? Would it be returned, if the age is given as 35 with probability 0.8 and some other value with probability 0.2? And if the model is extended with quality information, then the processing of queries must be extended to determine the quality of the answers it issues.

1.3 Imperfect Retrieval Requests

Most information systems provide retrieval languages for formulating precise, well-understood queries. Once information systems are capable of storing imperfect information, it is only natural to extend their query languages to allow "imperfect" queries as well. Consider these examples.

- Retrieve the employees who are *certain* to have a given skill.
- Retrieve the five employees who match *most closely* a given qualification.
- Retrieve all the employees who are *young and well paid.*

Note the difference between extensions that allow conventional query languages to work with more general kinds of information (discussed earlier) and extensions that allow more general kinds of queries (discussed here).

Altogether, the scope of this book includes the representation and manipulation of imperfect information, and the definition and execution of imperfect retrieval requests.

2 STRUCTURE

This book comprises 15 chapters. Except for this introductory chapter and the final chapter, which provides a bibliography, the core of the book is divided roughly into two parts. The first part (Chapters 2–7) was authored by researchers who work in information systems, and the second part (Chapters 8–14) was authored by researchers in uncertainty modeling in artificial intelligence.

The purpose of the first part is twofold:

1. To describe the issues and challenges in the area of imperfect information that confront information systems.
2. To describe the state of the art with respect to this area; that is, the solutions that have already been applied and the experience that has been gained from these solutions.

The purpose of the second part is twofold as well:

1. To describe the principal theories for modeling imperfect information.
2. To show how these theories may be adapted to information systems, responding to their particular challenges, and, in some cases, to discuss actual experimentation with prototype systems that demonstrate these theories.

The discussion in the first part is divided mostly by specific kinds of information systems. Chapter 2 is a general introduction to the problems and issues of imperfect information in information systems. Using an abstract model of an information system it classifies the various kinds of imperfection in information systems, and it provides a quick survey of the most notable solutions that have been attempted or suggested. Chapters 3–7 focus, respectively, on relational databases, intelligent databases, scientific and statistical databases, expert systems, and information-retrieval systems.

The most common type of information system are databases, and the most popular kind of databases are relational databases. Consequently, considerable attention has already been paid to various types of imperfect information in relational databases; mostly, missing information. Chapter 3 provides a thorough survey of this subject. The term *intelligent databases* is used in Chapter 4 to describe extensions to the relational model within the framework of mathematical logic. Essentially, these models introduce rules into the relational framework, thereby providing abilities to capture "knowledge." Whereas the discussion in Chapters 3 and 4 is, for the most part, independent of specific applications, Chapter 5 focuses on databases that are used in scientific and statistical applications. These two areas are a source for a variety of challenging problems of imperfect information.

Chapter 6 considers a second kind of information systems: expert systems. More accurately, it considers problems of imperfect information that are encountered in acquired knowledge: whether this knowledge is *declared* by human experts, as in most expert systems, or *discovered* in databases, by means of mostly automatic knowledge mining techniques.

Finally, Chapter 7 focuses on a third kind of information system: information-retrieval systems. Such systems, for example, systems for storing and retrieving documents, were early to develop models for representing objects whose contents are often open to subjective interpretations, and then develop retrieval techniques that "neutralize" this subjectiveness. This enables users to request documents on a specified subject, and retrieve relevant documents even when they have been classified under "close" subjects.

We note that both knowledge discovery and information retrieval are receiving much attention these days, with the explosive growth of databases and document repositories on the Internet and the World Wide Web. In their different ways, knowledge discovery and information retrieval both attempt to help us tame the enormous amount of information now available.

The discussion in the second part of this book is centered around the different theories that have been developed to manage imperfect information. Chapter 8 is an introduction to these theories. It provides a thorough classification of the various kinds of imperfect information, and it presents the various approaches that have been proposed to model imperfect information. Chapters 9–12 are devoted to models that are based, respectively, on probability theory, fuzzy set and possibility theory, mathematical logic, and belief functions. All but the logic-based models may be classified as *quantitative* models; logic-based models are *qualitative*.

Chapter 9 discusses uncertainty models that use probabilistic methods. The emphasis is on pragmatic issues. It is observed that there are a variety of probabilistic and Bayesian methods, each requiring different effort and providing different benefits, and the approach is to choose a model so that its costs will be commensurate with its benefits. Chapter 10 is devoted to uncertainty models that are based on fuzzy set theory and its associated theory of possibility. These theories offer a unified framework for managing a wide range of imperfect information, as well as for formulating retrieval requests of the kind discussed in Section 1.3. Various nonclassical logics have been proposed for modeling imperfect information, and Chapter 11 focuses on two such logics. Paraconsistent logic can represent and manipulate inconsistent information, and default logic can manage default information, such as rules that are usually true but are allowed to have exceptions. Chapter 12 concludes the discussion of uncertainty models with its transferable belief model. In the transferable belief model agents express their subjective opinions that a subset of the universe of possible values indeed contains the correct value. As in the Bayesian approach, this degree of belief is usually quantified by a probability measure, except that it does not rely on probabilistic quantification but on belief functions.

The uncertainty modeling part concludes with chapters 13 and 14, which survey the area of uncertainty modeling from personal perspectives. Chapter 13 discusses progress made in the field of approximate reasoning systems and their relevance and applicability to information systems. It considers mostly quantitative models, such as those based on probability theory, fuzzy set theory, or belief functions. Chapter 14 considers the question of whether it is possible to classify the different uncertainty techniques by the application needing uncertainty management; i.e., whether it is feasible to construct a matrix, whose columns are the different uncertainty models, and whose rows are the different kinds of applications, and whose values will indicate the suitability of the model to the application.

Chapter 15 concludes this book with an extensive, classified bibliography on the subject of imperfect information.

This book is the culmination of a project that brought together researchers in the scientific communities of information systems and uncertainty modeling to study the needs of the information systems community and to tap the expertise of the uncertainty modeling community for solutions that respond to these needs [1, 2].

REFERENCES

[1] A. Motro and P. Smets, editors. *Proceedings of Workshop on Uncertainty Management in Information Systems: From Needs to Solutions* (Puerto de Andraitx, Mallorca, Spain, September 23–26), 1992.

[2] P. Smets and A. Motro, editors. *Proceedings of Workshop on Uncertainty Management in Information Systems: From Needs to Solutions* (Avalon, California, April 2–5), 1993.