From Statistical Knowledge Bases to Degrees of Belief: An Overview*

Joseph Y. Halpern†
Computer Science Dept.
Cornell University
http://www.cs.cornell.edu/home/halpern
halpern@cs.cornell.edu

ABSTRACT

An intelligent agent will often be uncertain about various properties of its environment, and when acting in that environment it will frequently need to quantify its uncertainty. For example, if the agent wishes to employ the expected-utility paradigm of decision theory to guide its actions, she will need to assign degrees of belief (subjective probabilities) to various assertions. Of course, these degrees of belief should not be arbitrary, but rather should be based on the information available to the agent. This paper provides a brief overview of one approach for inducing degrees of belief from very rich knowledge bases that can include information about particular individuals, statistical correlations, physical laws, and default rules. The approach is called the *random-worlds* method. The method is based on the *principle of indifference*: it treats all of the worlds the agent considers possible as being equally likely. It is able to integrate qualitative default reasoning with quantitative probabilistic reasoning by providing a language in which both types of information can be easily expressed. A number of desiderata that arise in direct inference (reasoning from statistical information to conclusions about individuals) and default reasoning follow directly from the semantics of random worlds. For example, random worlds captures important patterns of reasoning such as specificity, inheritance, indifference to irrelevant information, and default assumptions of independence. Furthermore, the expressive power of the language used and the intuitive semantics of random worlds allow the method to deal with problems that are beyond the scope of many other non-deductive reasoning systems. The relevance of the random-worlds method to database systems is also discussed.

Categories and Subject Descriptors

H.2.4 [**Database Management**]: Systems—*relational databases*; H.2.8 [**Database Management**]: Database Applications—*Statistical databases*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods—*Representation languages.*

General Terms

Theory.

Keywords

Statistical information, degrees of belief, maximum entropy, principle of indifference, probabilistic databases, statistical databases.

1. INTRODUCTION

Consider an agent with a knowledge base, KB, who has to make decisions about its actions in the world. For example, a doctor may need to decide on a treatment for a particular patient, say Eric. The doctor's knowledge base might contain information of different types, including statistical information, e.g., "80% of patients with jaundice have hepatitis"; first-order information, e.g., "all patients with hepatitis have jaundice"; default information, e.g., "patients with hepatitis typically have a fever"; and information about the particular patient at hand, e.g., "Eric has jaundice". In most cases, the knowledge base will not contain complete information about a particular individual. For example, the doctor may be uncertain about the exact disease that Eric has. Since the efficacy of a treatment will almost certainly depend on the disease, it is important for the doctor to be able to quantify the relative likelihood of various possibilities. More generally, to apply standard tools for decision making such as *decision theory* (see, e.g., [21, 26]), an agent must assign probabilities, or *degrees of belief*, to various events. For example, the doctor may wish to assign a degree of belief to an event such as "Eric has hepatitis".

Fahiem Bacchus, Adam Grove, Daphne Koller and I [4] gave one particular method that allows such an agent to use its knowledge base to assign degrees of belief in a principled

*This is a brief overview of an invited talk, which is largely based on joint work with Fahiem Bacchus, Adam Grove, and Daphne Koller [4]

†Supported in part by NSF under grants CTC-0208535 and IIS-0534064, by ONR under grant N00014-01-10-511, by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the ONR under grants N00014-01-1-0795 and N00014-04-1-0725, and by AFOSR under grant F49620-02-1-0101.

PODS'06, June 26–28, 2006, Chicago, Illinois, USA.

manner; we call this method the *random-worlds method.* My talk will focus on this work because, although it is not so recent, it does seem quite relevant to recent work in the database community. Here I briefly review the approach (almost all the discussion is taken from [4], with very little change) and discuss the connection to database work.

There has been a great deal of work addressing aspects of this general problem. Two large bodies of work that are particularly relevant are the work on *direct inference*, going back to Reichenbach [25], and the various approaches to *nonmonotonic reasoning.* Direct inference deals with the problem of deriving degrees of belief from statistical information, typically by attempting to find a suitable *reference class* whose statistics can be used to determine the degree of belief. For instance, a suitable reference class for the patient Eric might be the class of all patients with jaundice. While direct inference is concerned with statistical knowledge, nonmonotonic reasoning deals mostly with knowledge bases that contain default rules. None of the systems proposed for either reference-class reasoning or nonmonotonic reasoning can deal adequately with the large and complex knowledge bases that we are interested in. In particular, none can handle rich knowledge bases that may contain a combination of first-order, default, and statistical information.

We assume that the information in the knowledge base is expressed in a variant of the language introduced by Bacchus [1]. Bacchus's language augments first-order logic by allowing statements of the form $\|Hep(x)|Jaun(x)\|_x = 0.8$, which says that 80% of patients with jaundice have hepatitis. Notice, however, that in finite models this statement has the (probably unintended) consequence that the number of patients with jaundice is a multiple of 5. To avoid this problem, we use approximate equality rather than equality, writing $\|Hep(x)|Jaun(x)\|_x \approx 0.8$, read "approximately 80% of patients with jaundice have hepatitis". Intuitively, this says that the proportion of jaundiced patients with hepatitis is close to 80%: i.e., within some tolerance τ of 0.8. This interpretation seems quite appropriate for database applications, where it is unlikely that we will be completely confident in the statistics that we have. The use of approximate equality has the further advantage of letting us express default information. We interpret a statement as "Birds typically fly" as expressing the statistical assertion "Almost all birds fly". Using approximate equality, we can represent this as $\|Fly(x)|Bird(x)\|_x \approx 1$. (This interpretation is closely related to various approaches applying probabilistic semantics to nonmonotonic logic; see Pearl [24] for an overview.)

Having described the language in which our knowledge base is expressed, we now need to decide how to assign degrees of belief given a knowledge base. Perhaps the most widely used framework for assigning degrees of belief (which are essentially subjective probabilities) is the Bayesian paradigm. There, one assumes a space of possibilities and a probability distribution over this space (the *prior* distribution), and calculates *posterior* probabilities by conditioning on what is known (in our case, the knowledge base). To use this approach, we must specify the space of possibilities and the distribution over it. In Bayesian reasoning, there is relatively little consensus as to how this should be done in general. Indeed, the usual philosophy is that these decisions are subjective.

Our approach is different. We assume that the KB contains all the knowledge the agent has, and we allow a very expressive language so as to make this assumption reasonable. This assumption means that any knowledge the agent has that could influence the prior distribution is already included in the KB. As a consequence, we give a single uniform construction of a space of possibilities and a distribution over it. Once we have this probability space, we can use the Bayesian approach: To compute the probability of an assertion ϕ given KB, we condition on KB, and then compute the probability of ϕ using the resulting posterior distribution.

So how do we choose the probability space? One general strategy, discussed by Halpern [18], is to give semantics to degrees of belief in terms of a probability distribution over a set of *possible worlds*, or first-order models. This semantics clarifies the distinction between statistical assertions and degrees of belief. As suggested above, a statistical assertion such as $\|Hep(x)|Jaun(x)\|_x \approx 0.8$ is true or false in a particular world, depending on how many jaundiced patients have hepatitis in that world. On the other hand, a degree of belief is neither true nor false in a particular world—it has semantics only with respect to the entire set of possible worlds and a probability distribution over them. There is no necessary connection between the information in the agent's KB and the distribution over worlds that determines her degrees of belief. However, we clearly want there to be some connection. In particular, we want the agent to base her degrees of beliefs on her information about the world, including her statistical information. The random-worlds method turns out to be a powerful technique for accomplishing this.

To define our probability space, we have to choose an appropriate set of possible worlds. Given some domain of individuals, we stipulate that the set of worlds is simply the set of all first-order models over this domain. That is, a possible world corresponds to a particular way of interpreting the symbols in the agent's vocabulary over the domain. In our context, we can assume that the "true world" has a finite domain, say of size N. In fact, without loss of generality, we assume that the domain is $\{1, \ldots, N\}$.

Having defined the probability space (the set of possible worlds), we must construct a probability distribution over this set. For this, we give perhaps the simplest possible definition: we assume that all the possible worlds are equally likely (that is, each world has the same probability). This can be viewed as an application of the *principle of indifference.* Since we are assuming that all the agent knows is incorporated in her knowledge base, the agent has no *a priori* reason to prefer one world over the other. It is therefore reasonable to view all worlds as equally likely. Interestingly, the principle of indifference (sometimes also called the *principle of insufficient reason*) was originally promoted as part of the very definition of probability when the field was originally formalized by Jacob Bernoulli and others; the principle was later popularized further and applied with considerable success by Laplace. (See [17] for a historical discussion.) It later fell into disrepute as a general definition of probability, largely because of the existence of paradoxes that arise when the principle is applied to infinite or continuous probability spaces. However, the principle of indifference can be a natural and effective way of assigning degrees of belief in certain contexts, and in particular, in the context where we restrict attention to a finite collection of worlds.

In any case, combining the choice of possible worlds with the principle of indifference, we obtain a prior distribution. We can now induce a degree of belief in ϕ given KB by conditioning on KB to obtain a posterior distribution and then computing the probability of ϕ according to this new distribution. It is easy to see that, since each world is equally likely, the degree of belief in ϕ given KB is the fraction of possible worlds satisfying KB that also satisfy ϕ.

One problem with the approach as stated so far is that, in general, we do not know the domain size N. Typically, however, N is known to be large. We therefore approximate the degree of belief for the true but unknown N by computing the limiting value of this degree of belief as N grows large. The result is the random-worlds method. (Of course, if a database includes information about the domain size, then we can just use it.)

The key ideas in the approach are not new. Many of them can be found in the work of Johnson [19] and Carnap [6, 7], although these authors focus on knowledge bases that contain only first-order information, and for the most part restrict their attention to unary predicates. Related approaches have been used in the more recent works of Shastri [28] and of Paris and Vencovska [23], in the context of a unary statistical language. Chuaqui [10] shares the idea of basing a theory of probabilistic reasoning upon notions of indifference and symmetry, although the details of his approach are quite different from ours.

Having defined the method, how do we judge its reasonableness? Fortunately, as we mentioned, there are two large bodies of work on related problems from which we can draw guidance: reference-class reasoning and default reasoning. While none of the solutions suggested for these problems seems entirely adequate, the years of research have resulted in some strong intuitions regarding what answers are intuitively reasonable for certain types of queries. Interestingly, these intuitions often lead to identical desiderata. In particular, most systems (of both types) espouse some form of preference for more specific information and the ability to ignore irrelevant information. We show that the random-worlds approach satisfies these desiderata. Moreover, in the absence of information, the random-world method makes attributes independent. Rather than having to assume independence, as is done in many applications, it is a provable consequence of the approach. In fact, all these properties follow from two general theorems. We prove that, in those cases where there is a specific piece of statistical information that should "obviously" be used to determine a degree of belief, random worlds does in fact use this information. The different desiderata, such as a preference for more specific information and an indifference to irrelevant information follow as easy corollaries. We also show that random worlds provides reasonable answers in many other contexts, not covered by the standard specificity and irrelevance heuristics. Thus, the random-worlds method is indeed a powerful one, that can deal with rich knowledge bases and still produce the answers that people have identified as being the most appropriate ones.

Of course, to the extent that we are going to use the method, we have to calculate degrees of belief. In general, the problem of deciding whether the degree of belief even exists (that is, whether there is a limiting probability) is undecidable [16, 20]. We can get decidability if we restrict to *unary* knowledge bases, where there are no function symbols and all predicates are unary [15, 20]. More importantly, as shown in [14], there is a close connection between random worlds and *maximum entropy* in the case of unary knowledge bases. Based on this connection, we show that in many cases of interest a maximum-entropy computation can be used to calculate an agent's degree of belief.

The connection between random worlds on maximum entropy relies critically on the assumption that we are conditioning on a unary formula. In fact, in almost all applications where maximum entropy has been used (and where its application can be best justified in terms of the random-worlds method) the knowledge base is described in terms of unary predicates (or, equivalently, unary functions with a finite range). For example, in physics applications we are interested in such predicates as quantum state (see [13]). AI applications and expert systems also often use only unary predicates [9] such as symptoms and diseases. In general, many properties of interest can be expressed using unary predicates, since they express properties of individuals. Indeed, a good case can be made that statisticians tend to reformulate all problems in terms of unary predicates, since an event in a sample space can be identified with a unary predicate [27]. In fact, in most cases where statistics are used, we have a basic unit in mind (an individual, a family, a household, etc.), and the properties (predicates) we consider are typically relative to a single unit (i.e., unary predicates). Thus, results concerning computing the asymptotic conditional probability if we condition on unary formulas are significant in practice.

On the other hand, for database applications, it is quite standard to see binary relation like "Manager-of". Note that the random-worlds method makes sense for predicates of arbitrary arity; it is just the connection to maximum entropy that is lost.

More generally, how does the random-worlds method relate to work in databases? I briefly discuss a few points of contact:

- One application is obvious: There are many statistical databases available, such as those derived from census data and economic data. The random-worlds method gives a principled method of deriving conclusions about particular individuals based on the statistical information.

- There has also been a great deal of work, especially recently, on probabilistic and imprecise databases; see, for example, [8] and the more recent [5] and [12] and the references therein.[1] Probabilistic databases allow probabilities to be associated with tuples. Roughly speaking, the probability represents the likelihood that the tuple is in the database. In some settings, the probability is best interpreted as statistical information. For example, Burdick et al. [5] give an example of an automobile database that lists makes of cars and the problem from which they are suffering ("brake problem", "transmission problem", etc.). The probabilistic information could be interpreted as summarizing statistical information, in which case the techniques used here apply immediately to draw conclusions. Alternatively, it could be interpreted as representing a degree

[1] Interestingly, Cavallo and Pittarelli [8] suggest using maximum entropy approaches to handle probabilistic databases, although they do not give independent motivation for doing so.

of belief (an agent's subjective belief that, say, the car is suffering from a brake problem). The method as described cannot deal with knowledge bases that include degrees of belief; however, in [3], we discuss three ways that the random-worlds method can be extended to handle degrees of belief.

- Random-worlds style methods have been used to define notions of privacy, where all instances of a database consistent with what is known are considered equally likely; cf. [22].

It seems that there is now a great deal of interest in the database community in finding ways to draw inferences from databases that include incomplete and imprecise information. The random-worlds method and other related approaches (see [2]) may prove to be a useful tool for doing this. Indeed, as I have mentioned, random-worlds style methods have already been used in the context of privacy; they have also been used in analyzing probabilistic databases [11]. I suspect that more applications will be found as well.

2. REFERENCES

[1] F. Bacchus. *Representing and Reasoning with Probabilistic Knowledge*. MIT Press, Cambridge, Mass., 1990.

[2] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistics to belief. In *Proc. Tenth National Conference on Artificial Intelligence (AAAI '92)*, pages 602–608. 1992.

[3] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. Generating new beliefs from old. In *Proc. Tenth Conference on Uncertainty in Artificial Intelligence (UAI '94)*, pages 37–45, 1994.

[4] F. Bacchus, A. J. Grove, J. Y. Halpern, and D. Koller. From statistical knowledge bases to degrees of belief. *Artificial Intelligence*, 87(1–2):75–143, 1996.

[5] D. Burdick, P. M. Deshpande, T. S. Jayram, R. Ramakrishnan, and S. Vaithyanathan. OLAP over uncertain and imprecise data. In *VLDB '05, Proceedings of the 31st Int. Conference on Very Large Databases*, pages 970–981, 2005.

[6] R. Carnap. *Logical Foundations of Probability*. University of Chicago Press, Chicago, 1950.

[7] R. Carnap. *The Continuum of Inductive Methods*. University of Chicago Press, Chicago, 1952.

[8] R. Cavallo and M. Pittarelli. The theory of probabilistic databases. In *VLDB '87, Proceedings of the 13th Int. Conference on Very Large Databases*, pages 71–87, 1987.

[9] P. C. Cheeseman. A method of computing generalized Bayesian probability values for expert systems. In *Proc. Eighth International Joint Conference on Artificial Intelligence (IJCAI '83)*, pages 198–202. 1983.

[10] R. Chuaqui. *Truth, Possibility, and Probability: New Logical Foundations of Probability and Statistical Inference*. North-Holland, Amsterdam, 1991.

[11] N. Dalvi and D. Suciu. Efficient query evaluation on probabilistic databases. In *VLDB '04, Proceedings of the 30th Int. Conference on Very Large Databases*, pages 864–875, 2004.

[12] N. Dalvi, G. Miklau, and D. Suciu. Asymptotic conditional probabilities for conjunctive queries. In *Proc. International Conference on Database Theory*, pages 287–303, 2005.

[13] K. G. Denbigh and J. S. Denbigh. *Entropy in Relation to Incomplete Knowledge*. Cambridge University Press, Cambridge, U.K., 1985.

[14] A. J. Grove, J. Y. Halpern, and D. Koller. Random worlds and maximum entropy. *Journal of A.I. Research*, 2:33–88, 1994.

[15] A. J. Grove, J. Y. Halpern, and D. Koller. Asymptotic conditional probabilities: the non-unary case. *Journal of Symbolic Logic*, 61(1):250–275, 1996.

[16] A. J. Grove, J. Y. Halpern, and D. Koller. Asymptotic conditional probabilities: the unary case. *SIAM Journal on Computing*, 25(1):1–51, 1996.

[17] I. Hacking. *The Emergence of Probability*. Cambridge University Press, Cambridge, U.K., 1975.

[18] J. Y. Halpern. An analysis of first-order logics of probability. *Artificial Intelligence*, 46:311–350, 1990.

[19] W. E. Johnson. Probability: The deductive and inductive problems. *Mind*, 41(164):409–423, 1932.

[20] M. I. Liogon'kiĭ. On the conditional satisfiability ratio of logical formulas. *Mathematical Notes of the Academy of the USSR*, 6:856–861, 1969.

[21] R. D. Luce and H. Raiffa. *Games and Decisions*. Wiley, New York, 1957.

[22] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramanian, ℓ-diversity: privacy beyond k-anonymity. In *Proc. Int. Conf. Data Engineering (ICDE 2006)*, 2006.

[23] J. B. Paris and A. Vencovska. On the applicability of maximum entropy to inexact reasoning. *International Journal of Approximate Reasoning*, 3:1–34, 1989.

[24] J. Pearl. Probabilistic semantics for nonmonotonic reasoning: a survey. In *Proc. First International Conference on Principles of Knowledge Representation and Reasoning (KR '89)*, pages 505–516, 1989. Reprinted in G. Shafer and J. Pearl (Eds.), *Readings in Uncertain Reasoning*, pp. 699–710. San Francisco: Morgan Kaufmann, 1990.

[25] H. Reichenbach. *The Theory of Probability*. University of California Press, Berkeley, 1949. Translation and revision of German edition, published as *Wahrscheinlichkeitslehre*, 1935.

[26] L. J. Savage. *Foundations of Statistics*. Wiley, New York, 1954.

[27] G. Shafer. Personal communication, 1993.

[28] L. Shastri. Default reasoning in semantic networks: a formalization of recognition and inheritance. *Artificial Intelligence*, 39(3):285–355, 1989.