Aggregating Across Multiple Levels of Granularity to Meet Customer and Organizational Query Requirements

Suzanne M. Mahoney,
Innovative Decisions, Inc..
1945 Old Gallows Rd.
Vienna,VA 22182
smahoney@innovativedecisions.com

Abstract

A research organization responds to a variety of customer requests. Each high level request is broken down into a set of low level requests. For each low level request, the research organization follows a multi-step process: access, acquire, analyze and report. The customer is given an estimate of whether its request will be met, and if not, why not. The explanation may be provided at multiple levels of granularity including request level (e.g. low level request X cannot be fulfilled) and step level (e.g. the problem lies with acquisition). Available databases contain an incomplete history of past requests as well as information about current staffing and resources. The challenge is to provide an estimate of current capabilities for satisfying requests in general and to estimate the probability for meeting a specific incoming request.

This paper proposes an abstract process model and an aggregation strategy for combining evidence across multiple levels of granularity into estimates that meet these challenges. It specifies an approach for constructing query-specific Bayesian networks in response to queries from the customers and research organization management.

1 INTRODUCTION

A research organization invited Innovative Decisions to develop a method that predicts its ability to produce a research report requested by a customer in a timely manner. Currently, within a few weeks of receiving a request, the organization provides an estimate of its likelihood of producing a report. These estimates are generated by analysts drawing on both collected data and the experiential knowledge from their networks of experts.

The challenge is to produce an estimate immediately, relying only on the available on-line data. Several different approaches were considered, including drawing statistics from existing databases. However, pertinent statistics are generally not available at the level of a requested report. This led to the proposal to provide relative estimates by modeling the reporting process with a Bayesian network that uses indirect, but relevant, data as evidence. However, because the relevant data is at various levels of granularity, it was decided to construct process models for the lowest levels of request as well as for those levels consistent with relevant data. The proposal recommends using an aggregation strategy to connect relevant data to a process model representing the customer's request and automated network construction to limit the constructed network to relevant information.

This proposed solution meets the research organization's requirements. It has not yet been implemented. For reasons of confidentiality, the example given is a surrogate for the actual application.

The following section introduces a motivating example. The next section reviews related research. There are three aspects to the application: Process models; aggregation; query-specific network construction. In Section 4, the process and aggregation models are introduced while Section 5 discusses the construction of query-specific networks.

2 EXAMPLE

A research organization specializes in generating reports about artwork. A customer requests a tailored report. It may be restricted to a set of artists, a period in time, and/or a specified collection. At its most specific level, a report is for a specific artist, decade of completion, and the collection in which the artwork is located. A customer's request can be viewed as a set of very specific reports.

Furthermore, each customer requests specialized information that may or may not have been reported in the past. For example, one customer may be interested in the theme/subject of the artwork while another may be interested in the materials and methods used in producing the artwork. Therefore, past reports, while indicative of the organization's capacity to respond to a customer's request, are not predictive.

The process of responding to a request includes the following steps:

1. Gain <u>access</u> to the collections in which relevant works of art may be kept;

2. Identify relevant works of art within the collection and <u>acquire</u> relevant data such as measurements, photos, curatorial history, X-rays, and samples for spectrography

3. <u>Analyze</u> the relevant art works using the collected data;

4. Generate a <u>report</u> responding to the user's specific questions.

When a customer submits a request, she wishes to know whether or not the request is likely to be fulfilled in a timely fashion. The organization's response is necessarily couched in uncertain terms. By way of explanation, the organization needs to produce an estimate of the capability to perform each step of the process. To support drill-down the organization also wants to produce capability estimates for each process, for each artwork included in the request.

The organization's capability to accomplish each step has some uncertainty associated with it. First, a particular collection, especially if it is in private hands, may not be accessible. And even if access is granted to a collection believed to hold relevant art works, those art works may not be found because of destruction, loss or simply incorrect records. Even when some art work is found, a researcher with the appropriate expertise to analyze the art work may not be available.

In general, past performance is indicative of future capability. However, the performance data available for each process step is uncertain and incomplete. This is partly due to past poor record-keeping. But also, reports are usually prepared at the request level and may or may not include information on every artwork notionally included in the request.

At the same time, management of the organization wants reports documenting how well general capabilities of the organization are meeting current requests.

3 RESEARCH

3.1 Relevant applications

Researchers have used Bayesian networks to model various processes. Some examples include: Deventer et al (1999) modeled injection molding as a set of sub-processes; Weidl et al (2003) described an industrial process operation and asset management tool based on an object-oriented Bayesian network;. Wolbrecht, et al (2000) modeled a multistage manufacturing process.

3.2 Aggregation Approach

The underlying assumption that the capabilities of low level processes can be combined independently is modeled by independence of causal influence (ICI) functions. The first ICI function, a binary Noisy-OR (Pearl, 1988; Henrion, 1987), was generalized by Srinivas (1993) and Diez (1993) to non-binary variables. In general, an ICI function simply requires that the influence of parents on a dependent variable be independent of one another. (Heckerman and Breese, 1996; Zhang and Yan, 1997)

Probabilistic ICI (pICI) functions have a combination function that is probabilistic rather than deterministic (Zagorecki, et al, 2006). An example is an *average* function. In this case, the value assigned to a state is the number of states with that value divided by the total number of states.

3.3 Automated Construction

Object Oriented Bayesian Networks (Pfeffer and Koller, 1997) and Multi-Entity Bayesian Networks (Laskey, 2006) propose approaches for automatically constructing Bayesian networks from a knowledge base of network fragments.

4 ABSTRACT MODELS

The first subsection presents the specific requirements for the problem. The second describes the abstract process network for modeling the steps at each level of granularity while the last section presents an aggregation strategy.

4.1 Specific Requirements

<u>Request specific</u>: First a model needs to be responsive to a customer's specific request. These requests, while often similar to past requests, are usually unique.

<u>Capability-based</u>: The organization needs to indicate whether current capabilities are adequate to respond to a request. These capabilities estimates indicate where in the process weaknesses lie. Besides giving an explanation of why the organization can/cannot answer a specific request, they also provide management with insight on what needs to be changed in order to improve the organization's ability to respond to future requests.

<u>Drill-down</u>: Furthermore, these capability estimates need to be provided at different levels of granularity to support detailed explanations for why or why not a request may be satisfied.

<u>Ordinal</u>: Without more specific historical data, the model is not expected to provide realistic probability estimates. Rather, the probabilities will serve as ordinal numbers for comparing existing capabilities and the relative probability of producing a given report.

4.2 Abstract Core Model

The network in Figure 1 shows the abstract model for the capability and process model. It includes two types of variables:
(1) Variables that represent the capabilities for each step of the research process, and
(2) variables that represent the steps of the process.

To make an instance of the abstract model, one simply specifies a context. Following our example, context is a combination of the set of artists, set of years, set of collections.

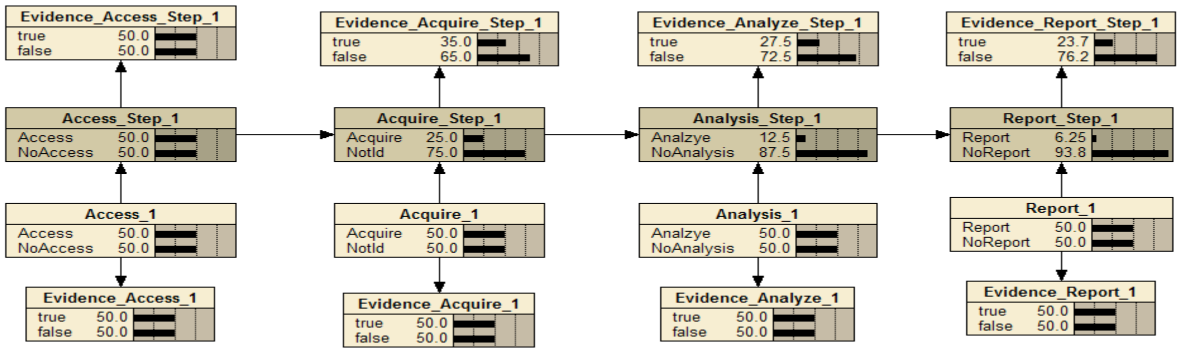

Figure 1 Core Model with Notional Evidence Nodes

At its most granular the context is a single artist, single decade and single collection.

The semantics for the abstract model are as follows:

- A capability node represents the probability that the organization will be able to perform the specific capability for the context. Each capability node has a uniform distribution representing a lack of prior knowledge. See the third row of nodes in Figure 1
- A process step node represents the probability that the organization will perform the specific step given a request to do so. These are deterministic nodes in the second row of Figure 1. A step occurs when both its associated capability is present and the previous step has been accomplished. Otherwise, the step does not occur.

Note that the relationships between steps combined with the priors of 0.5 for the capabilities forces the probability of each step to be half that of the prior step.

All evidence is specific to a level of granularity. For example, a report about a specific artist's work in a certain decade would be applied to an instance of the abstract core model with that context.

Figure 1 shows notional evidence nodes attached to both process and capability nodes. These serve as placeholders or stubs for more complex sets of evidence nodes. In this abstract model, the evidence nodes apply to exactly one process or capability node. This is a preferred structure, but not required.

The Bayesian network of the figure shows two different categories of evidence. The first is evidence of past performance, the top row of nodes in Figure 1, while the second is evidence of current capabilities, the bottom row of nodes in Figure 1.

Evidence of past performance is associated with process nodes. This evidence generally affects all of the other nodes in the network. For example, evidence of a report means that the prior processes have been accomplished sometime in the past for that context. Therefore, the associated capabilities must have been present sometime in the past and are therefore more likely to be present today.

Evidence of current capability could be that the organization has blanket permission to access artworks possessed by a given museum or the organization has analysts with specific talents under contract.

4.3 Basic Aggregation Model

A customer request can be viewed as a combination of low level requests. In order to answer how well and whether the organization can respond to a specific request, one may aggregate a set of low level process networks into a process network representing the customer request.

Capabilities associated with the low level processes are aggregated into higher level capabilities because it is ultimately the capabilities of the organization that determine whether or not it can respond to a specific request. The structure of the aggregation is shown in Figure 2.

This structure aggregates over binary variables. Binary variables allow one to readily order variables according to the posterior probabilities of the capability states. However, the approach is not limited to binary variables.

Several different aggregation strategies permits the user to choose the aggregation strategy that best fits the nature of the capability and the purpose of the aggregation. For example, if one needs access to all collections in which certain works of art reside, then the All aggregation strategy would be appropriate for the Access capability. On the other hand, if the organization may assess its chances of gaining access to a particular museum by simply taking an average of its experience in accessing works of art in that museum. In another case, even though a particular research request specifies four different low-level processes, the customer may be satisfied if only three of four of those low-level processes are successful.

Figure 2 shows five different possible aggregation strategies.

- All: All capabilities are required to satisfy the requirement.
- Any: Having any one of the capabilities will satisfy the requirement.
- Average: Calculate the average capability.
- At Least Half: Of these four, two or more will satisfy the requirement.
- Exactly Three: Exactly three capabilities are required.

5 QUERY-SPECIFIC NETWORKS

The probability that a customer's research request may be fulfilled can be answered by constructing a query-specific Bayesian network from instances of the abstract core model combined with instances of the basic aggregation model.

5.1 Query-specific network examples

In its simplest form, a query-specific network is composed of abstract core model instances for each element of the customer's request and an abstract core model instance for the request itself. These are connected via a set of aggregation structures for each of the capabilities. Evidence nodes for these core model instances is also attached to the query-specific network. Once the network is

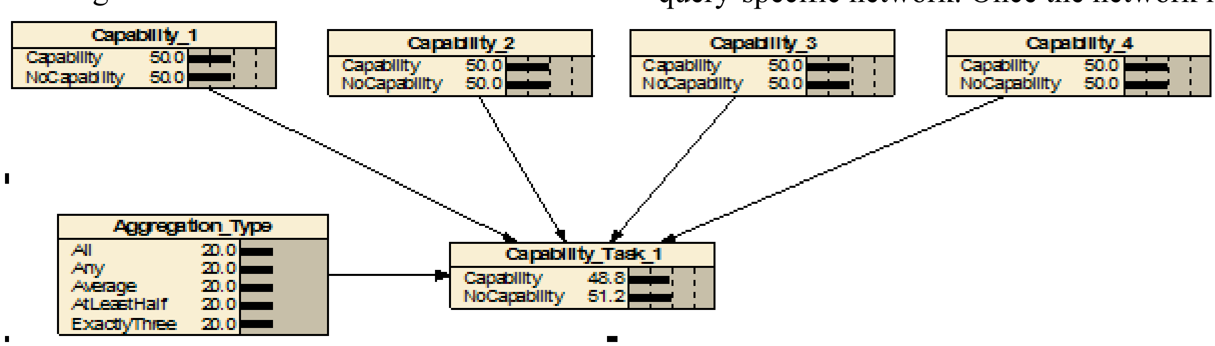

Figure 2 Basic Aggregation Model

constructed, evidence is applied and the inference algorithm is applied.

For example, suppose that the customer's request concerns artist A2's work during Decade D1. Assume that just four different collections hold examples of such work. The circled network of Figure 3 shows the simplest version of a query-specific network constructed in response to the customer's request. Each node in this network represents an instance of the abstract core model with evidence attached.

Suppose some previous report has been prepared comparing artist's A1's work with that of artist A2. Finding such a report in the evidential database, causes us to add its core model instance along with its associated low level core models to the initial query-specific network. Figure 3 shows the resulting query-specific network.

Essentially, the network construction rule in this case is "If there is evidence for a different request whose low level core model instances overlap those of the simplest query-specific network for the current request, then instantiate abstract core model fragments for that request and its low level core model instances." There are two key elements to the rule. The overlap requirement simply means that only relevant requests are considered. The there is evidence requirement is there because if there is no evidence for what could be a relevant request, then modeling that request will make no difference in the inferred posteriors. Because all of the nodes in the process model representing a high level request are descendents of nodes in low level process models, they make no difference in the computation if there is no evidence.

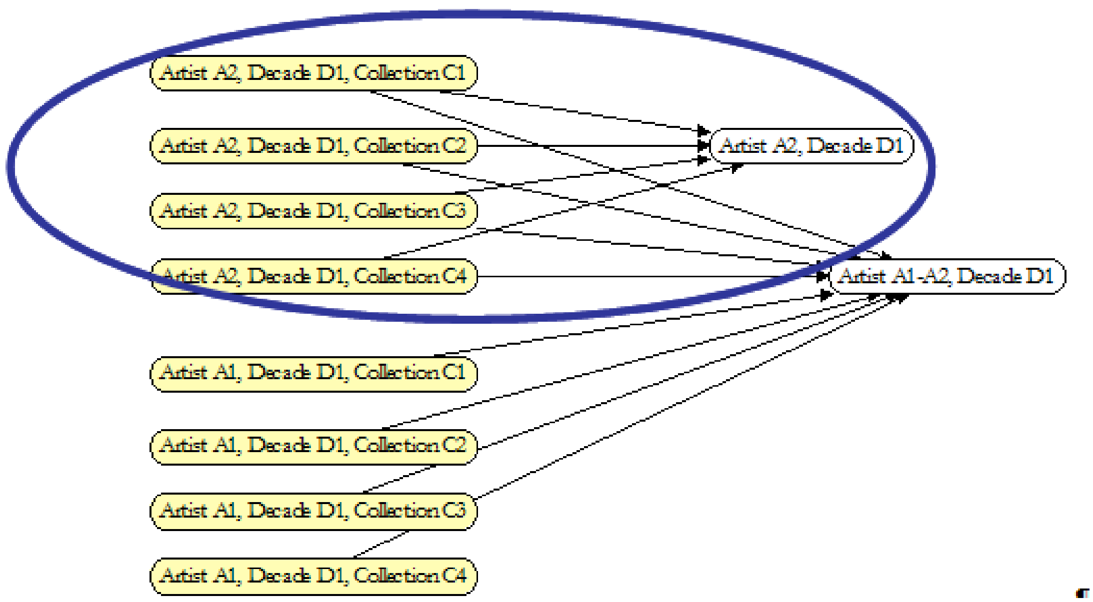

Figure 3 Aggregating to meet Customer Request

Now, consider a more complex example. A customer wants to know about artist A1's activities during decades 1 and 2. There are several 'obviously' relevant reports. They include the reports about the artist during each of those two decades. Even a report comparing artist 1 with artist 2 during one of the decades appears to be relevant. If there is evidence for a report comparing artists A1 and A2 for that decade, then the report about artist A2 is relevant within the context of the Bayesian network. The question is how important is that relevance for determining

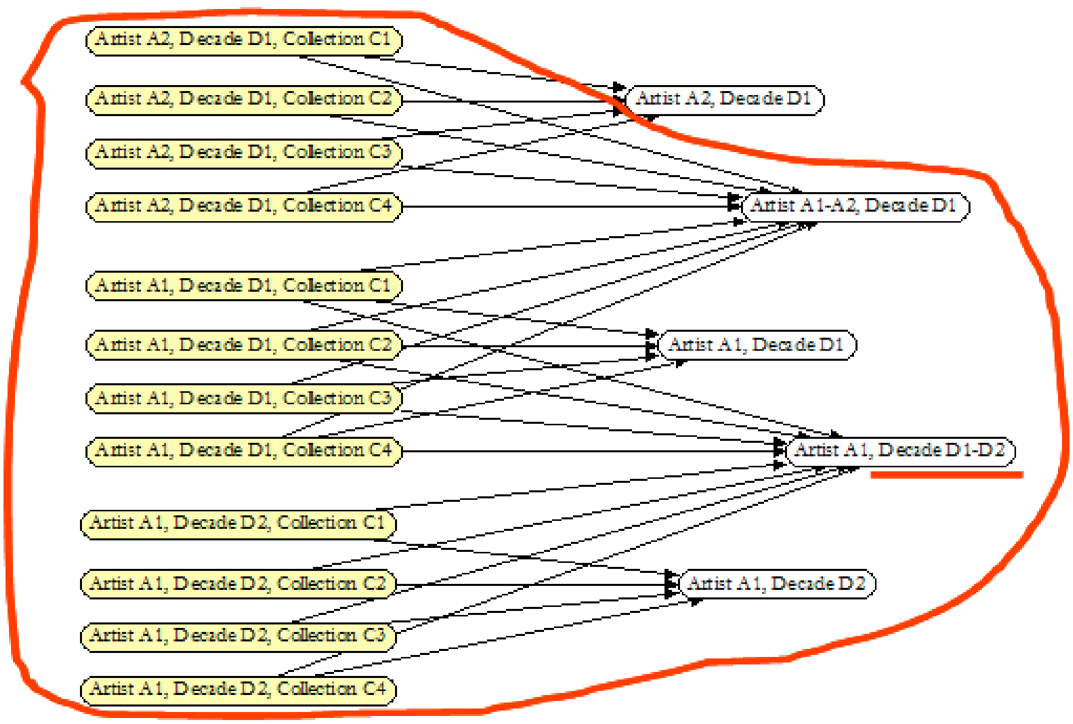

Figure 4. A More Complex Example

whether the organization can respond to the specific request about artist A1. In this case, the organization may decide that the report is not relevant enough. See Figure 4.

This suggests that the query-specific construction function requires stopping criteria so that the network built in response to a request remains tractable.

5.2 Automated Construction

The steps for constructing the simplest model tailored to a customer request are:

1. Instantiate a core model whose context is the level of granularity of the request.
2. Instantiate a set of lowest level core models that combined satisfy the request.
3. Aggregate these lowest level networks to request the level network using one or more versions of the basic aggregation model. Specific aggregation strategies may be specified by the customer and/or the organization.
4. Apply available evidence for both the lowest level and request level networks.

The additional steps for constructing a <u>highly relevant</u> query-specific network tailored to a customer request are shown below. If evidence exists for a high level request whose low level process models <u>substantially overlap</u> the low level models of this request, then:

5. Instantiate that high level request
6. Instantiate the low level models associated with that high level request.
7. Aggregate low level process models to the high level request model.
8. Apply available evidence

Of course, the issue here is to decide a policy for what substantially overlap means. This may depend upon the customer as well as the organization.

5.3 Inference Considerations

The presented abstract core model is singly connected. If the evidence is only attached to one capability or process node, the network remains singly connected. So, individually, the process models are computationally tractable. (Pearl, 1988)

The aggregation of capability nodes into higher levels has the pattern shown in Figures 3 and 4. The network is now no longer singly connected. However, as long as evidence is only collected at the lowest level of the process model, the network is tractable and higher level capability posteriors can be readily computed. In this case, the lack of evidence for higher level processes d-separates the network, so that when one is calculating high level capability posteriors, the network is computationally equivalent to the structure of the simplest network for the request.

Constructed query-specific networks will not usually be in their simplest form. To meet customer requests, rapidly computing the posteriors of a query-specific network is desirable. So, it falls upon the automated network construction software to limit the constructed model to one that is computationally tractable. Trade-offs between precision and inference efficiency need to be made.

6 SUMMARY

This paper shows how abstract core and basic aggregation models may be instantiated and combined with automated construction algorithms to form interconnected Bayesian networks of high and low level core model instances. Any subset of the low level instances may be combined into a high level model.

This flexibility supports customers by being responsive to any request and permits a request to be broken down across different dimensions. For example, the customer who is interested in two different artists may obtain an estimate for the artists individually.

At the same time, it gives the organization a large number of ways in which to combine its available data to answer high level questions about capabilities.

Some issues that need to be resolved include:

- To what extent should the customer be able to specify aggregation strategies. The default seems to be that customers want everything, but may be satisfied with part. This also goes to how the organization measures its performance regarding customer satisfaction.
- A second issue is how to measure relevance. To what extent should the customer control the relevancy and how does one present that to the customer?

References:

Deventer R., B. Kahles and H. Niemann (1999) Process modelling using Bayesian networks, *Book of abstracts, Workshop on Conditional Independence Structures and Graphical Models*

Di_ez, F.J. (1993). Parameter adjustment in Bayes networks. The generalized noisy OR gate. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93),* 99-105. San Francisco, CA: Morgan Kaufmann Publishers

Heckerman, D. & Breese, J.S. (1994). A new look at causal independence. In *Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence (UAI-94),* 286-292. San Francisco, CA: Morgan Kaufmann Publishers

Heckerman, D. & Breese, J.S. (1996). Causal independence for probability assessment and inference using Bayesian networks. *IEEE, Systems, Man and Cybernetics, 26*, 826-831.

Koller, D. and A. Pfeffer (1997) Object-Oriented Bayesian Networks In Geiger, D. and Shenoy, P. (eds) *Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference*, San Francisco, CA: Morgan Kaufmann. pp. 302-313.

Laskey, K.B. (2006) MEBN: A Logic for Open-World Probabilistic Reasoning George Mason University C4I Center Technical Report C4I06-01

Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann, San Mateo, CA.

Srinivas, S. (1993) A Generalization of the Noisy-Or Model. In *Proceedings of the Ninth Annual Conference on Uncertainty in Artificial Intelligence (UAI-93),* 208-215. San Francisco, CA: Morgan Kaufmann Publishers

Weidl, G., A. Madsen and E. Dahlquist, (2003) Object oriented Bayesian network for industrial process operation, First Bayesian Applications Modeling Workshop, http://www.intel.com/ research/events/UAI03_workshop/

Wolbrecht, E., B. D'Ambrosio, R. Paasch, D. Kirby (2000) Monitoring and diagnosis of a multistage manufacturing process using Bayesian networks *Artificial Intelligence for Engineering Design, Analysis and Manufacturing* Volume 14 , Issue 1 (January 2000) pp. 53-67

Zagorecki A. & Druzdzel, M.J. (2006). Knowledge engineering for Bayesian networks: How common are noisy-MAX distributions in practice? In *Proceedings of the Seventeenth European Conference on Artificial Intelligence (ECAI-06)*, G. Brewka, S. Coradeschi, A. Perini & P. Traverso (eds.), pp. 482-489. Amsterdam: IOS Press.

Zagorecki, A., Voortman, M., & Druzdzel, M.J. (2006). Decomposing local probability distributions in Bayesian networks for improved inference and parameter learning. In FLAIRS Conference.

Zhang, N., & Yan, L. (1997). Independence of causal influence and clique tree propagation. In *Proceedings of the 13th Annual Conference on Uncertainty in Artificial* Query-Specific *Intelligence (UAI-97),* 481-488. San Francisco, CA: Morgan Kaufmann Publishers.