interspeech 2008 incorporating SST 08

Towards the Integration of Automatic Speech Recognition and Information Retrieval for Spoken Query Processing

A. Moreno-Daniel[1], *J. Wilpon*[2], *B.-H. Juang*[1], *S. Parthasarathy*[2,*]

[1]Georgia Institute of Technology, Atlanta GA, USA
[2]AT&T Labs, Florham Park NJ, USA
{antonio,juang}@ece.gatech.edu, {jgw,sps}@research.att.com

Abstract

Spoken query processing (SQP) is the task of fulfilling an information need, inferred from a spoken query, by listing a set of ranked relevant documents. The two main sources of uncertainty in SQP lay on the realization of the speech waveform and on the realization of the observed document. The proposed integration models these uncertainties under a single probabilistic framework. A case study on movie title retrieval by voice is presented to illustrate the proposed methodology. By allowing an ontology inlet, a 14% relative gain in the model convergence was achieved. An improved mean reciprocal rank and mean inclusion rate of the retrieval outcome was obtained using the proposed framework.

Index Terms: spoken query processing, voice search, automatic speech recognition, information retrieval

1. Introduction

Spoken query processing (SQP) is the task of fulfilling an information need, inferred from a spoken query, by listing a set of ranked relevant documents. SQP is a particularly challenging problem because the length of the queries is considerably shorter than the length of the documents. SQP gains relevance in environments where the presence of a keyboard is cumbersome or infeasible, such as in home theaters or media rooms, on small portable devices, or in places where hands/eyes-free operation is needed.

A family of problems related to SQP have already been approached in the past with relative success. Spoken document/utterance retrieval (SDR/SUR) [1, 2] and keyword spotting [3] perform a search in a speech corpus from a text query (opposite to SQP). On the other hand, a spectrum of tasks that perform search on a text corpus from a spoken query can be spanned by the degree of syntactic mismatch between the queries and the listing entries. In directory assistance (DA) [4] canned queries match the listing entries, with optional field re-ordering/skipping (e.g. *"Kung Fu Panda, Jack Black"*). In voice search (VS) [5, 6] queries differ from the listing names due to the multiple syntactic formulations (e.g. *AT&T Labs, AT&T Shannon Laboratories* for the listing: AT&T Shannon Labs, Inc.-Research). In SQP, queries are open to natural speech phrases that match semantically (not necessarily syntactically) the intended document (e.g. *"Show me an animated film with a fat panda and martial arts"*).

The two main components in SQP are: automatic speech recognition (ASR) and information retrieval (IR), two techniques that have evolved and improved independently for their own cause; however, when placed to co-exist and pursue a single endeavor, such as SQP, the operation is suboptimal. Our hypothesis is that a joint re-design of ASR and IR that allows a broader bridge for interaction will result in an improved SQP performance. The work being presented is a step in this direction.

Recent work has attempted to utilize a richer representation of the recognition output from the ASR system at the IR stage, rather than the traditional plain *best-path* text sequence converted from the spoken input. For example, under a vector-space IR paradigm, a *bag of words* can be built from the entire word recognition lattice, and weigh each word with its *a posteriori* probability [7], or with a lattice-depth independent weight based on the best-path [8]. Although improved results over their unweighted counterpart were found, suggesting that an ASR-IR integration can improve SQP, there still exists a mismatch between the ASR and the IR frameworks.

SQP operates in two steps: 1) *pre-processing* (offline), where the document corpus is indexed and a language model is designed, and 2) *processing* (online), where the user's need is inferred and relevant documents are retrieved.

A common performance measure in ASR is word error rate (WER), defined as:

$$\mathrm{WER} = (S + I + D)/N, \tag{1}$$

where S, I and D are the number of substitutions, insertions and deletions, respectively, in a word string of length N. In SQP, the *tokenization* stage that performs morphological transformations (stop-word removal, word stemming and collocations) converts this measure into term error rate (TER), defined as:

$$\mathrm{TER} = \left[(S - \epsilon_S) + (I - \epsilon_I) + (D - \epsilon_D)\right]/\tilde{N}, \tag{2}$$

where ϵ_S is the number of substitutions that are ultimately merged into the same term (including null), ϵ_I is the number of inserted words that become null (dropped), and ϵ_D is the number of deleted words that would have been set to null in any case. $\tilde{N}$ is the number of terms in the string after tokenization. In general $\tilde{N} \ll N$, thus often $\mathrm{TER} > \mathrm{WER}$.

SQP is a relatively new paradigm that is still to be explored by users [9] and further studied by the human computer interface community. The formulation of queries in a casual style is certainly more seemly for spoken queries than for typed ones, which helps ASR because it can take advantage of the word inter-dependence (context).

The notation used along this document for random variables denotes single terms as w, term-sequence as $\mathbf{w}$ and its length as $|\mathbf{w}|$, the acoustic features sequence as $\mathbf{X}$, and a particular document as d.

*Now with Yahoo! Labs.

2. Methodology

There are several approaches to broaden the interaction between ASR and IR. Some are driven by common sense, others rely on heuristics, and yet others are inspired by methods that have been successful in similar tasks. Some of these methods are listed in Sect 2.1. Given that the acoustic and language models in ASR are probabilistic, the proposed framework embraces a probabilistic scheme for IR (Sect. 2.2). This proposed framework can be seen as an extension to IR to handle queries with an additional layer of uncertainty (acoustic in our case, but also applicable to queries entered with a soft-keyboard). Additionally, an inlet for exploiting ontology is also proposed (Sect. 2.3).

2.1. Broadening the ASR-IR interaction

The recognition lattice $\mathbf{R}$ is a compact yet rich representation of the ASR outcome representing the multiple hypotheses in the recognition network within a threshold in the search space. For the sake of SQP, this lattice ought to be "trimmed" in ways such that information conveying the user's need is kept while other is discarded.

- β^1: *best word string*. This is the simplest and most common method: $\mathbf{R} \rightarrow \mathbf{R}^{\beta_1} = \mathbf{w}_1$, where $\mathbf{w}_n$ is the n-th best word-string. The performance of this approach is optimum for those cases where the $\mathrm{TER} = 0$, even if the $\mathrm{WER} > 0$.
- β^N: *N-best word strings*. Common sense suggests that the most consistent words across the N-best word-strings might be correct, therefore the recognition lattice can be transformed to:
  $$\mathbf{R} \rightarrow \mathbf{R}^{\beta_N} = \bigcup_{n=1}^{N} \mathbf{w}_n.$$
  The value of N needs to be selected (how deep the lattice is explored).
- $\mathcal{G}^N$: *N-best word graph*. Additionally to the word-strings $\mathbf{w}_n$, $\mathbf{R}$ provides a measure of "responsibility" (or negative cost) for every word-string. $c(\mathbf{w}_n) = p(X|\mathbf{w}_n)p(\mathbf{w}_n)$, $\forall \mathbf{w}_n \in \mathbf{R}$,
  $$\mathbf{R} \rightarrow \mathbf{R}^{\mathcal{G}_N} = \bigcup_{n=1}^{N} \left\{ \mathbf{w}_n : c(\mathbf{w}_n) / \textstyle\sum_k c_k \right\}.$$
  The value of N needs to be selected.
- $\mathcal{G}^w$: *word graph*. An approach commonly used in SDR is to transform $\mathbf{R}$ into a bag of weighted words:
  $$\mathbf{R} \rightarrow \mathbf{R}^{\mathcal{G}_w} = \bigcup_{w} \left\{ w : \textstyle\sum_{\mathbf{w} \ni w} c(\mathbf{w}) / \sum_k c(\mathbf{w}_k) \right\}.$$
  In practice, the number of words (or terms) w is truncated to a reasonable number (e.g. a value proportional to $|\mathbf{w}_1|$).

2.2. Probabilistic framework for SQP

A fundamental difference between SQP and IR alone is the intrinsic uncertainty of the query due to its acoustic realization. In IR, text queries and documents are merely a limited length realization of a description, each in its own way, of a complex concept; therefore, there exists a component of uncertainty in *understanding* that depends on the concept being described and the number of terms (words) needed/provided to fully register it.

A number of similarity measures have been proposed in literature including cosine, Jaccard and intersection similarities, however no single one is successful in all domains. In this work, following the probabilistic formulation, relevance of a document to a spoken query was defined as the probability: $p(d|\mathbf{X})$, therefore the most relevant documents are: $\tilde{d} = \mathrm{argmax}_d \; p(\mathbf{X}, d)$.

Let us first expand this joint probability with the nuisance variable $\mathbf{w}$ that represents any arbitrary term-sequence (or word-sequence): $\mathbf{w} = (w_1, \ldots, w_L)$ allowed by the language model,

$$p(\mathbf{X}, d) = \sum_{\mathbf{w}} p(\mathbf{X}, d, \mathbf{w}), \tag{3}$$

Since the acoustic realization $\mathbf{X}$ and the document d are assumed to be conditionally independent given the term-sequence $\mathbf{w}$, the Eq. 3 becomes:

$$p(\mathbf{X}, d) = \sum_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})\, p(\mathbf{w}, d). \tag{4}$$

Notice that factor $p(\mathbf{w}, d)$ models the understanding uncertainty as the joint probability of observing $\mathbf{w}$ and d.

Probabilistic latent semantic analysis (PLSA) [10] is an aspect modeling method that assumes terms and documents are drawn from a limited set of unobserved aspects (or topics). PLSA finds the term-document joint probability distribution: $p(w, d)$ that minimizes the Kullback-Leibler (KL) divergence w.r.t. the empirical distribution, subject to K latent aspects, and assuming that terms and documents are independent conditioned to an aspect, denoted by the variable z, thus: $p(w, d|z) = p(w|z)\, p(d|z)$. Since $\mathbf{w}$ in Eq. 4 is a term-sequence, we defined $p(\mathbf{w}, d) = \left[\prod_{w \in \mathbf{w}} p(w, d)\right]^{1/|\mathbf{w}|}$ (mean log-probability), which neglects the order of occurrence.

Therefore, extending PLSA to queries carrying acoustic uncertainty, Eq. 4 becomes:

$$p(\mathbf{X}, d) = p(d) \sum_{\mathbf{w}} p(\mathbf{X}|\mathbf{w}) \left[\prod_{w \in \mathbf{w}} \sum_{z} p(w|z)\, p(z|d) \right]^{1/|\mathbf{w}|}, \tag{5}$$

where $p(d)$ is the document's prior (popularity) and $p(z|d)$ is PLSA's representation (indexing) of document d in the aspect simplex space.

The factor $p(\mathbf{X}|\mathbf{w})$ in Eq. 5 models the acoustic uncertainty as the likelihood of observing $\mathbf{X}$ when the terms $\mathbf{w}$ are present in the spoken query. Since $\mathbf{w}$, the nuisance variable, can arbitrarily take any term-sequence, a reasonable subset is the N-best from $\mathbf{R}$ as these will have the largest $p(\mathbf{X}|\mathbf{w})$. Therefore, among the described forms for broadening the ASR-IR interaction in Sect. 2.1, $\mathcal{G}^N$ suits the framework. Experiments with the other representations that illustrate how $\mathcal{G}^N$ performs w.r.t. to other approaches are examined in Sect. 3.

2.3. Inducing ontology

The two immediate advantages obtained by exploiting ontology are: 1) the latent aspects have a "real-world" meaning attached to them; and 2) the information provided by the ontology helps to place related documents closer together, a task difficult to accomplish with sparse and limited data alone.

For clarity, this scheme will be illustrated with an example. Let each text document be the storyline or synopsis of a movie, and the ontology be the movie genres (25 types) denoted as $\mathcal{Z}_1, \ldots, \mathcal{Z}_{25}$. The steps are the following:

- Assume uniform distribution for $p(\mathcal{Z}|d)$, $\forall \mathcal{Z} \ni d$ (the distribution of genres in a document).
- Divide each genre $\mathcal{Z}$ into $K_{\mathcal{Z}}$ aspects, where $K_{\mathcal{Z}} \propto \sum_d 1(d \in \mathcal{Z})$ (i.e. number of documents in a particular genre), thus $K_{\mathrm{drama}} > K_{\mathrm{documentary}}$.

- For every genre $\mathcal{Z}$, use K-means to cluster the empirical term-distributions $p(w|d),\ d \in \mathcal{Z}$ and set $p(w|z), \forall z \in \mathcal{Z}$ to the centroids found. These centroids are the initial aspects z.
- Let $p(z|d) = \frac{1}{K_{\mathcal{Z}}} p(\mathcal{Z}|d)$, i.e. divide $p(\mathcal{Z}|d)$ evenly into $p(z|d), \forall z \in \mathcal{Z}$.
- Dither $p(w|z)$ and $p(z|d)$ to prevent numerical singularities.
- Use $p(w|z)$ and $p(z|d)$ to seed PLSA's maximum likelihood estimation.

The log-likelihood of the observed data is maximized via EM (expectation maximization) algorithm, where the E-step is:

$$p(z|w,d) = \frac{p(w|z)p(z|d)}{\sum_z p(w|z)p(z|d)}, \tag{6}$$

which follows Bayes theorem, and the M-step is:

$$p(z|d) = \frac{\sum_w n(w,d)\, p(z|w,d)}{\sum_w n(w,d)}, \tag{7}$$

$$p(w|z) = \frac{\sum_d n(w,d)\, p(z|w,d)}{\sum_w \sum_d n(w,d) p(z|w,d)}. \tag{8}$$

where $n(w,d)$ represents the empirical term-document joint distribution (from observed co-occurrences). The outcome of this optimization for PLSA is an estimate of $p(z|d)$ and $p(w|z)$, which form part of our SQP scheme in Eq. 5.

3. Experiments

3.1. Case study: movie retrieval

The case study selected is a movie title retrieval system. Speakers were instructed to provide natural spoken queries requesting broad information in movies. For example: "*I'd like to see, uh, something about the American civil war*", "*do you have anything about aviation or planes?*", were typical natural spoken queries. A set of 130 spoken queries was obtained from 30 users (telephone quality speech) using the system: *SpeakFlicks*[1]. The spoken queries had an average of 7.2 words and 3.0 terms.

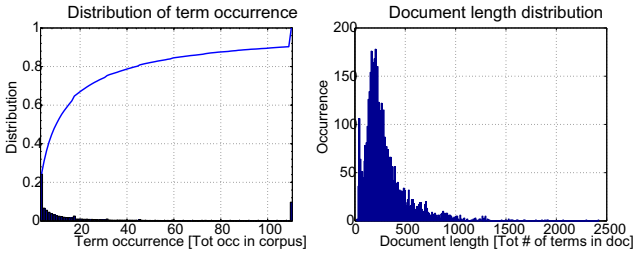
The text corpus contains 4.5K documents assembled from synopses of popular movies, collected from up to four different publicly available Internet sources, adding up to a mean document length of 300 terms, and a vocabulary of 25K terms (after tokenization). The authors of these synopses are anonymous and assumed to be diverse. The distribution of the term-occurrence within the corpus evidences the sparseness as 50% of the terms occur ten times or less (see Figure 1).

3.2. Induced ontology

The performance metric used to assess this method is the perplexity of $p(w,d)$ (PLSA's term-document models) for documents d in a validation set (10% of text documents), this measures how well the model anticipates the validation documents, where $p(w|z)$ was learned from the training documents and $p(z|d \in \text{validation})$ was obtained by *folding* d into the aspect simplex space [10] via EM algorithm.

Figure 2 shows the convergence for these initialization schemes with $K = 60, 260, 360, 460$ hidden aspects. As expected, the use of ontology provides an advantage in the early iterations and clustering further improves it. The curve labeled as *Ont* skips the clustering step setting $p(w|z) = p(w|d \in \mathcal{Z})$.

[1] SpeakFlicks is an experimental system developed by Georgia Tech available at speakflicks.com for development and data collection.



(a) (b)

Figure 1: Histogram and cumulative distribution of the term occurrences is shown in (a). Half of the terms occur ten times or less. The document length (number of term occurrences) is shown in (b). The small peak around 80 are documents (movies) with only one Internet source.

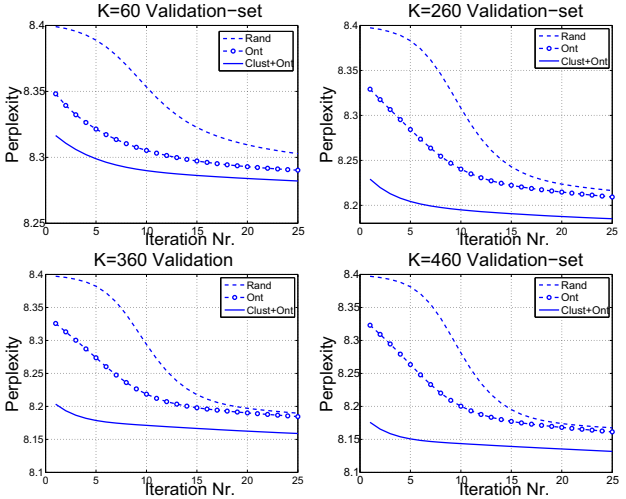


Figure 2: Convergence of the perplexity in a validation set for three initialization schemes: *Rand* (traditional), *Ont* (enforced ontology without clustering), and *Clust+Ont* (proposed).

Ont converges to an slightly lower value of perplexity than *Rand* particularly for models with large values of K. *Clust+Ont*, described in Sect. 2.3, outperforms the other two. Assuming the perplexity of a random $p(w,d)$ as reference point, an overall relative improvement of 14% w.r.t. the conventional *Rand* initialization is achieved with the proposed method. Multiple realizations with different random seeds led to qualitatively equivalent results.

3.3. Acoustic uncertainty

A statistical language model (SLM) is used to drive the recognition engine. Building such an SLM from a database of transcribed queries is ideal; however, a sufficiently large set of spoken queries (100K+) is not available at the moment. Nevertheless, since queries can be interpreted as distorted versions of documents, a distorted SLM can be created from the text document corpus itself. A 3-gram SLM is designed using the 4.5K documents, and a vocabulary of 75K words. Despite of this mismatch, a 0.44 WER, and a 0.50 TER were obtained, as shown in Table 1, while maintaining a real-time factor close to one.

The methodology for evaluating the retrieval outcome for

Words		Terms	
SER(W):	0.82	SER(T):	0.66
WER:	0.44	TER:	0.50
WIR-β^1:	0.58	TIR-β^1:	0.61
WIR-$\mathbf{R}$:	0.68	TIR-$\mathbf{R}$:	0.77

Table 1: SER: sentence error rates for words (W) and terms (T), WER/TER: word/term error rates. WIR/TIR: reference-words/terms inclusion rate in β^1 and $\mathbf{R}$ (higher is better).

SQP was the following. Noting that, unlike TREC [1], a manually annotated database for SQP providing pairs of spoken queries and their corresponding *target retrieval outcome* does not exist in the present time, an alternative form of evaluation had to be devised. If we consider that the human-transcribed spoken query is the ASR upper-ceiling, its retrieval outcome represents a reference that can be used as "ground" and be set as the target retrieval outcome. This scheme allows (for now) to perform an experimental analysis without any manual scoring. For the following experiments, *Clust+Ont* was used to obtain $p(w|z)$ and $p(z|d)$.

Two performance measures were used: 1) $\mathrm{MRR}(r)$: the mean reciprocal rank of the r-th target outcome entry in the retrieved outcome being evaluated, and 2) $\mathrm{MIR}_R(r)$: the mean inclusion rate of the r-th target outcome within the top R retrieved entries.

MIR is a more lenient metric than MRR because MIR only penalizes the absence of the target entry from the top R entries, not its actual rank within the outcome. The value of $R = 8$ was chosen assuming this many entries would fit on the screen of a movie-request interface (such as SpeakFlicks). Table 2 shows the $\mathrm{MRR}(r)$ and Table 3 shows the $\mathrm{MIR}_R(r)$ for the methods described in Sect. 2.1 denoted as: β^1, β^N, $\mathcal{G}^N$, $\mathcal{G}^w$. The larger the MRR and MIR, the better the performance.

r:	1	2	3	4	5	6	7	8
Orcl	1.00	0.50	0.33	0.25	0.20	0.16	0.14	0.12
β^1	0.42	0.21	0.15	0.11	0.09	0.08	0.07	0.06
β^N	0.38	0.19	0.13	0.09	0.08	0.07	0.07	0.06
$\mathcal{G}^N$	0.43	0.24	0.16	0.11	0.09	0.07	0.08	0.07
$\mathcal{G}^w$	0.33	0.18	0.13	0.10	0.07	0.06	0.05	0.04

Table 2: Mean reciprocal rank (MRR) for different ranks: $r = 1, \ldots, 8$. β^1 represents the traditional scheme, $\mathcal{G}^N$ the proposed one and Orcl: the performance of the oracle (ideal) outcome.

r:	1	2	3	4	5	6	7	8
Orcl	1.	1.	1.	1.	1.	1.	1.	1.
β^1	0.45	0.44	0.44	0.40	0.40	0.38	0.36	0.34
β^N	0.40	0.40	0.39	0.36	0.37	0.34	0.35	0.31
$\mathcal{G}^N$	0.47	0.50	0.44	0.41	0.40	0.39	0.39	0.33
$\mathcal{G}^w$	0.42	0.42	0.35	0.34	0.27	0.23	0.23	0.13

Table 3: Mean inclusion rate at $R = 8$. Orcl: is the performance of the oracle (ideal) outcome.

The method β^N naively combines the N-best term-strings (a value of $N = 2$ was selected because larger values decreased the performance further). This result evidences how important it is to combine the probabilities systematically. A method often used in disciplines such as SDR is $\mathcal{G}^w$, which is based on a measure of the posterior probability for a set of words in the lattice. This scheme does not consider the "joint occurrence" of terms as it rather takes each term separately. The traditional method β^1 has a very competitive performance, which is not surprising if we notice that 34% of the β^1 term-sentences have a $TER = 0$ and that 61% of the reference-terms are present in the best-path (see Table 1). The method $\mathcal{G}^N$ consistently performs well. Solid improvement is found in MRR and MIR particularly for top ranks (the most relevant). This simple evaluation, based on the target-retrieval outcome, encourages the pursuit of a more elaborated end-to-end evaluation for SQP based on hand-labeled data.

4. Conclusions

A framework for integrating ASR and IR for SQP was proposed based on PLSA. This framework provided an inlet for ontology into PLSA's initialization that lead to improving the convergence point of the aspect model by 14%, and it also extended PLSA to handle queries with acoustic uncertainty (spoken queries) achieving a consistent gain in the MRR and MIR w.r.t. to the traditional method. Future work will explore a more comprehensive evaluation method and define a suitable data set for this task. Future research will also take advantage of a larger data set (text documents and spoken queries) as it is currently being accumulated.

5. Acknowledgments

The first author thanks S. Byers and B. Holister for helping in the collection and transcription of data, and T. Wada for the useful discussions.

6. References

[1] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: a success story," *Proceedings of TREC-8*, pp. 107–130, Apr. 2000.

[2] M. Saraclar and R. Sproat, "Lattice-based search for spoken utterance retrieval," *Proceedings of the HLT-NAACL*, pp. 129–136, 2004.

[3] J. G. Wilpon, L. R. Rabiner, C. H. Lee, and E. R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden markov models," *Acoustics, Speech, and Signal Processing, IEEE Transactions on*, vol. 38, no. 11, pp. 1870–1878, 1990.

[4] S. Parthasarathy and A. Moreno-Daniel, "Directory retrieval using voice form-filling," *Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. IV:161–165, Apr. 2007.

[5] D. Yu, Y.-C. Ju, Y.-Y. Wang, G. Zweig, and A. Acero, "Automated directory assistance system - from theory to practice," *Proceedings of Interspeech*, pp. 2709–2712, August 2007.

[6] Y.-Y. Wang, D. Yu, Y.-C. Ju, and A. Acero, "An introduction to voice search," *IEEE Signal Processing Magazine*, pp. 29–38, May 2008.

[7] P. Wolf and B. Raj, "The MERL SpokenQuery information retrieval system," *Proceedings of ICME*, vol. 2, pp. 317–320, Aug. 2002.

[8] A. Moreno-Daniel, S. Parthasarathy, B.-H. Juang, and J. G. Wilpon, "Spoken query processing for information retrieval," *Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. IV:121–125, Apr. 2007.

[9] F. Crestani and H. Du, "Written versus spoken queries: A qualitative and quantitative comparative analysis," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 7, pp. 881–890, May 2006.

[10] T. Hofmann, "Probabilistic latent semantic analysis," *Proceedings of Uncertainty in Artificial Intelligence*, pp. 289–299, 1999.