Patterns of Search: Analyzing and Modeling Web Query Refinement

Tessa Lau[1] and Eric Horvitz[2]

[1] Department of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[2] Decision Theory & Adaptive Systems, Microsoft Research, Redmond, WA, USA

Abstract. We discuss the construction of probabilistic models centering on temporal patterns of query refinement. Our analyses are derived from a large corpus of Web search queries extracted from server logs recorded by a popular Internet search service. We frame the modeling task in terms of pursuing an understanding of probabilistic relationships among temporal patterns of activity, informational goals, and classes of query refinement. We construct Bayesian networks that predict search behavior, with a focus on the progression of queries over time. We review a methodology for abstracting and tagging user queries. After presenting key statistics on query length, query frequency, and informational goals, we describe user models that capture the dynamics of query refinement.

1 Introduction

The evolution of the World Wide Web has provided rich opportunities for gathering and analyzing anonymous log data generated by user interactions with network-based services. Web-based search engines such as Excite, AltaVista, and Lycos provide search services by crawling and indexing large portions of the Web. In a typical session, a user enters a string of words into the search engine's input field and receives an HTML page containing a list of web documents matching the user's query. This list may include hundreds of documents. Web search engines rank this list and present the results in small groups. The user may follow one or more of the returned links, request additional results, or refine and resubmit a query.

We review our work in developing models of users' search behaviors from log data. Our motivation is to enhance information retrieval by developing models with the ability to diagnose a user's informational goals. In this paper, we describe probabilistic models that are used to infer a probability distribution over user's goals from the time-stamped data available in server logs. More specifically, we elucidate probabilistic relationships between user goals and temporal patterns of query refinement activity. In contrast with the work of Maglio and Barrett (1997), who study several users' complete web searches, we consider only interactions with the search service itself. In distinction to the work of Silverstein et al. (1998), which reported statistics over a large corpus of unprocessed log data, we mine a corpus that has been hand-tagged to provide information about the refinements and goals associated with queries.

We shall first describe the initial data set and our methodology for transforming the data into a representation of user behavior with richer semantics. We review a set of definitions that abstract queries into classes of query refinement and informational goals. We then present key statistics of our corpus. Finally, we describe the construction of Bayesian network models that capture dependencies among variables of interest.

2 Server Log Corpus

We analyzed a data set derived from the server logs generated by the Excite Internet search engine. The server log data was made available for research purposes by Excite. The server logs capture an unspecified portion of all queries to the Excite search engine over a twenty-four hour time period on Tuesday, September 16, 1997.

Each entry in the initial log file records a single query, the time the query was input, and a globally unique identifier (GUID), which uniquely identifies each client using the search service. The total size of the data set is 48 megabytes, representing approximately one million queries. We extracted a 200-kilobyte portion of the corpus, representing 4,690 queries. The unprocessed logs contain data of the following form:

```
8A563CBE26CA77A9   970916144332   rhubarb
8A563CBE26CA77A9   970916144534   rhubarb pie
8A563CBE26CA77A9   970916144836   rhubarb pie
B04ABFA483164552   970916080514   trac right
5F5338040B2A4285   970916225207   peace, adrian paul fan club
```

The first column contains the GUID. The second column shows the time at which the query was made in YYMMDDHHMMSS format (year, month, day, hour, minutes, seconds). The remainder of the line contains the user's query.

3 Enriching the Semantics of Server Data

This server data is limited in that it shows only users' interactions with the Excite search engine. No information is provided about the user's selection of links offered by the search engine or about navigation to content beyond the search engine. We are also limited in our ability to assess the ultimate success or failure of users to find what they were seeking. Nevertheless, informational goals and query actions can be inferred via deliberate inspection of queries. We have tagged the server logs by hand to extend the data set with a human interpretation of search actions and informational goals of users.

3.1 Assigning Query Refinement Classes

In an initial phase of hand tagging, we partitioned queries into classes representing different search actions. We focused in particular on the inferred refinement strategy of query sequences. We abstracted search actions into a set of mutually exclusive refinement classes, where the refinement class of a query represents a user's intent relative to his prior query. Refinement classes include:

- New: A query for a topic not previously searched for by this user within the scope of the dataset (twenty-four hours).
- Generalization: A query on the same topic as the previous query, but seeking more general information than the previous query.
- Specialization: A query on the same topic as the previous query, but seeking more specific information than the previous query.

- Reformulation: A query on the same topic that can be viewed as neither a generalization nor a specialization, but a reformulation of the prior query.
- Interruption: A query on a topic searched on earlier by a user that has been interrupted by a search on another topic.
- Request for Additional Results: A request for another set of results on the same query from the search service. Duplicate queries appear in the data when a person requests another set of results for the query, as detailed by Spencer (1998).
- Blank queries: Log entries containing no query. These entries arise when a user clicks on the search button with no query specified or when a query by example is performed, as explained by Spencer (1998). We removed blank queries from consideration.

Our original set of refinement classes did not include interruptions. However, we found a small proportion of query sequences in which users had two distinct goals (such as lawnmowers and pornography), and interleaved queries on both of these topics over the span of our study. We introduced the *Interruption* refinement class to capture this type of behavior.

While annotating queries, we noted several opportunities for automating the assignment of refinement class. For example, a request for additional results appears in the logs as a duplicate of the previous query. Generalizations and specializations can often be identified by query contractions or extensions, or by the addition of new terms with the use of Boolean connectives such as *and* and *or*. Although complete automation of the tagging of refinement classes is likely infeasible, there are opportunities for recognizing a subset of refinements in an unsupervised manner.

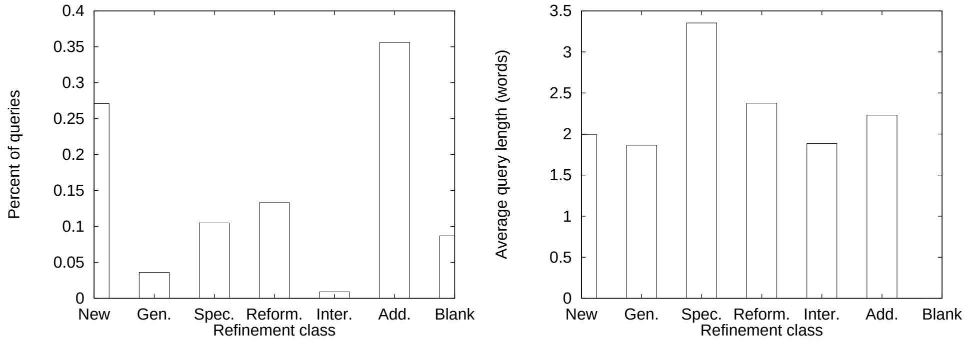
3.2 Assigning Informational Goals

A second phase of annotation of the data set focused on the classification of each query in terms of our best guess about the user's goal. We created a broad ontology of fifteen informational goals. To support our inferences about a user's goals, we reviewed the hits returned by Excite on the queries and carefully examined the other queries made by the same user. We implemented an editorial tool for assigning informational goals that allowed the editor to have immediate access to web pages recommended by a later version of the Excite service for each query in the data set.

Our ontology of the informational goals of users included *Current Events, Weather, People, Health Information, Products and Services, Recreation and Sports, Entertainment, Business, Adult Content, Science and Technical, Places, Education, Career Opportunities,* and *Non-Scientific Reference*.

We were unable to classify some queries into any of the above categories and labeled these queries as *Unclassifiable*. Unclassifiable queries included foreign-language queries (`"consorzio lecole"` or `ljuba vrtovec pribec`), malformed queries for nonexistent web sites (`www.hahoo .com`), and short or cryptic queries with little information content such as `http` or `hello`.

Many queries were difficult to classify because of the small amount of context available or the editor's unfamiliarity with the query topics. In some cases, review of a sequence of queries revealed the nature of a query. For example, a query of `dr-511` was only classifiable after noting the query `pioneer 24x cd` in the same session. In this context, the editor could infer that `dr-511` is likely to be the model number of a CDROM drive, and that the searcher was most likely looking for a product.



(a) (b)

Figure 1. (a) Distribution of user query behavior in 4690 queries analyzed. Fraction of queries is indicated on left axis. (b) Graph of average query length for each refinement class. Query length for additional results refers to active query when more results are requested.

Other queries, such as `harvest moon`, have several meanings. Harvest Moon is the name of a video game, the name of a furniture company, another name for the September moon, an album by musical artist Neil Young, the name of a natural foods company, and a poem by Ted Hughes. In this case, we classified `harvest moon` as a non-scientific reference. Because of the difficulty we encountered in assigning informational goals to queries, the prospect of fully automating this coding process is unlikely.

4 Key Statistics

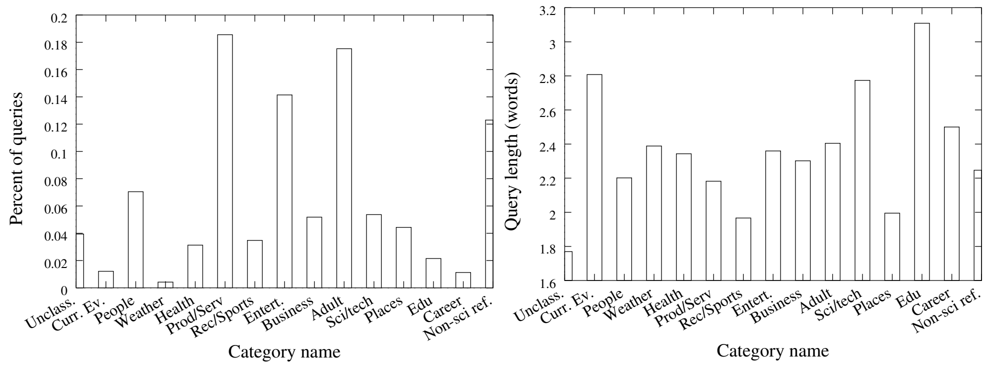
We shall first present several fundamental statistics we derived from the data set. Then we shall explore temporal trends. Finally, we will construct and exercise Bayesian network user models.

Users made on average 4.28 queries over the course of a day, with a standard deviation (sd) of 5.39. The average length of queries was 2.30 words (sd: 1.42). Users averaged 1.31 distinct informational goals per day (sd: 0.78), and performed 3.27 queries per goal (sd: 3.91).

4.1 Distribution and Query Length for Refinement Classes

The graph in Figure 1a displays the probability distribution over queries for the different query refinement classes. We found that most actions were either new queries or requests for additional information. Relatively few users refined their searches by specialization, generalization, or reformulation. A very small number of people interleaved searches on different topics (the *Interrupted* class of queries).

Figure 1b shows the breakdown of query length conditioned on different classes of query refinement. Specialization queries tended to contain more words than any other type of query; this was evidenced by the fact that people tended to add more words to queries in order to narrow the scope of their search. Reformulated queries tended to be longer than the initial queries they heralded from. These results highlight the opportunity for harnessing query length as an indicator of the refinement class of a query.



(a) (b)

Figure 2. (a) Distribution of informational goals in tagged data set. (b) Relationship between query length and information goal of query.

4.2 Distribution and Query Length for Informational Goals

We analyzed the distribution of informational goals in the tagged corpus (Figure 2a). We found that the largest category was *Products and Services*, describing 19% of all queries. The second largest category was *Adult Content* covering 18% of the queries. The smallest category was *Weather*, which occupied less than one percent of all queries.

We explored the influence of informational category on query length (Figure 2b). Compared to the overall average query length of 2.30 words, the longest queries were in the *Education* category, with a mean of over 3 words per query, followed by *Current Events* and *Scientific/Technical*. The shortest queries were in the *Unclassified* category. This is not surprising since shorter queries contain less information and are thus harder to classify into informational goals.

5 Temporal Dynamics of Query Behavior

The analyses we have described so far capture snapshots of a potentially complex process of formulation and iterative refinement of queries. We shall now review our work on generalizing the analyses of the tagged log data to probe temporal patterns of query behavior.

5.1 Inter-query Interval and Refinement Actions

We pursued an understanding of relationships between the time taken to browse or process the results of a search and the nature of refinement actions. As one approach to this problem, we examined pairs of adjacent queries from individual users, and assigned the pairs to buckets representing different inter-query time intervals according to a predefined discretization of intervals, ranging from the shortest interval of *0-10 seconds* to the largest of *Greater than 20 minutes*. For each inter-query interval, we computed the conditional probability that the next query would be in each refinement class. The results of this analysis are displayed in the graph in Figure 3. For each time bucket, the probabilities of all of the different refinement classes sum to 1.

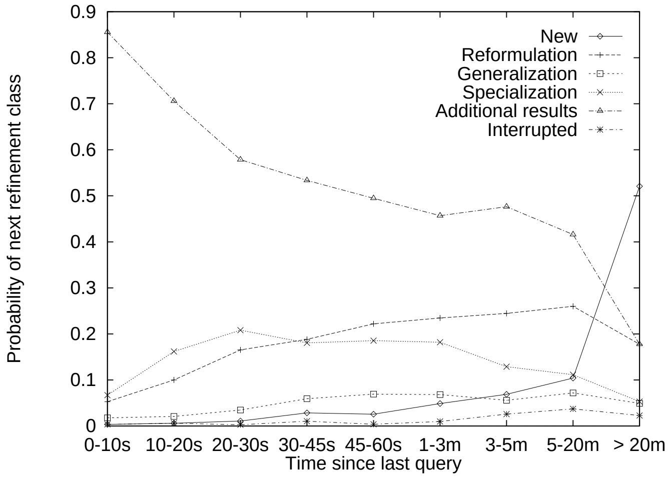


Figure 3. Relationships between the inter-query interval and next refinement class.

We discovered distinct relationships between the time interval separating adjacent query actions and the refinement class of the successive query. For example, the probability that a new query will be issued increases as the inter-query interval increases, growing to over a 0.5 probability when the inter-query interval is greater than 20 minutes. The probability of requesting additional results on a previous query is greater than 0.8 for the 0-10 second interval, but decreases to approximately 0.5 for the 1-3 minute interval, where it remains relatively stable until the interval becomes greater than 20 minutes. The probability that a user will specialize a previous query rises to a maximum in the 20-30 second interval, and holds steady at a probability of nearly 0.2 until it begins to diminish when the inter-query interval grows to 1-3 minutes. The probability of a generalization is always a small fraction of the probability of a specialization; this probability grows with the interval size, peaking in the 5-20 minute interval before diminishing. The probability of seeing a reformulation grows with the interval duration, peaking at 5-20 minutes before beginning to diminish.

The dynamics of the probabilities of alternative actions likely represent common behavioral patterns. For example, we believe the larger amount of time without interaction with the Excite service tends to indicate that the precursory query pointed the user towards a region of the Internet that provided either the desired information or a path to relevant information.

6 Constructing Probabilistic User Models

Our interest in characterizing—and predicting—a user's web search behavior under uncertainty led us to pursue the construction of general probabilistic dependency models that could take into consideration multiple observations including the timing of subsequent interactions, the pattern of query refinements, the length of a user's query, and prior probabilities of informational goals.

We have constructed Bayesian network models and parameterized the models with data drawn from the tagged Excite logs. A Bayesian network is a directed acyclic graph (DAG) representation of the joint probability distribution for a set of random variables Horvitz et al.; Pearl (1988; 1991). Nodes in Bayesian networks represent random variables and arcs represent probabilistic dependencies among pairs of variables. Several user modeling applications have benefited from Bayesian networks: Charniak and Goldman (1993) have used Bayesian networks for plan understanding; Horvitz and Barry (1995) and Jameson (1995) apply them to time-critical decision support and utility directed display of information; Conati et al. (1997), Horvitz et al. (1998) and Jameson (1995) diagnose a user's goals and needs; and Albrecht et al. (1997) model actions in a game setting using Bayesian networks.

6.1 Considering Inter-query Interval and Adjacent Actions

We constructed a Bayesian-network model for modeling a user's search behavior that considers the probabilistic relationships between inter-query intervals and the refinement classes of adjacent queries. As displayed by the Bayesian network in Figure 4, we consider relationships among the variables *User Search Action*, representing the first of two adjacent search actions, *Time Interval*, representing the inter-query interval, and *Next Search Action*, representing the next action taken with the search service. Arcs represent the assumption that the current search action and inter-query time interval may both influence the probability distribution over the next search action. As evidenced by the arc from *User Search Action* to *Time Interval*, we also allow for the possibility that the current search action also directly influences the delay before the next interaction with the service. The states for each variable in the model are displayed adjacent to the nodes.

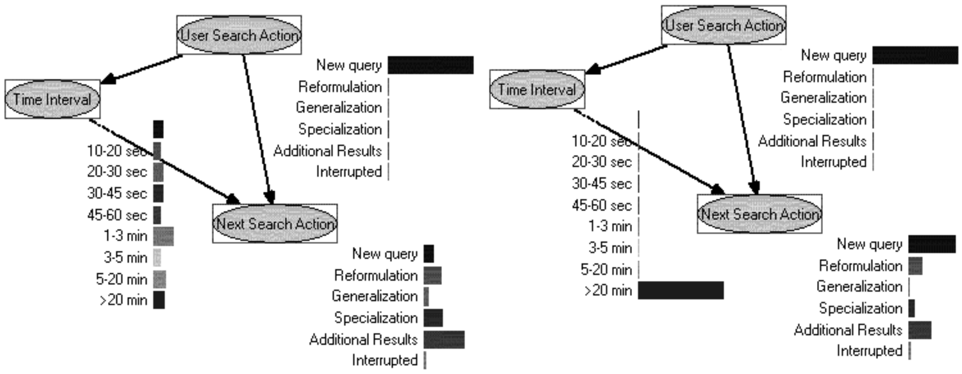
After constructing the overall dependency model, we generated from the initial data set conditional probability tables with probabilities of the states of each variable, conditioned on combinations of the states of nodes that are its immediate parents in the directed graph. The completed Bayesian network can be harnessed to infer probability distributions over the states of all variables given the explicit setting of the value of the states of one or more variables. The probabilities computed for each state are displayed as the length of bars adjacent to the nodes representing network variables.

Figure 4a displays the inferred probability distributions for the inter-query time interval and for the next search action in the situation where we input only the information that a new query was performed. The explicit setting of the variable *User Search Action* to the state *New Query* is indicated by that state having probability 1.0 and the other states having probability zero. Following a new query, the maximum likelihood delay before the next action taken with the search service is 1-3 minutes (p=0.22), and the action with highest likelihood is a request for additional results (p=0.46).

Figure 4b highlights the ability to make inferences about the probability of the next action in response to the current query and the delay before the next action occurs. Suppose we know that no action has occurred in the 20 minutes since a new query was input. Probabilities are inferred over the next action. In this case, the maximum likelihood action is a new query (p=0.54), followed by a request for an additional page of results (p=0.25).

6.2 Extending the User Model to Consider Informational Goals

We extended the Bayesian-network introduced in section 6.1 by introducing additional context in the form of the inferred informational goals of users. We were interested in predictive power



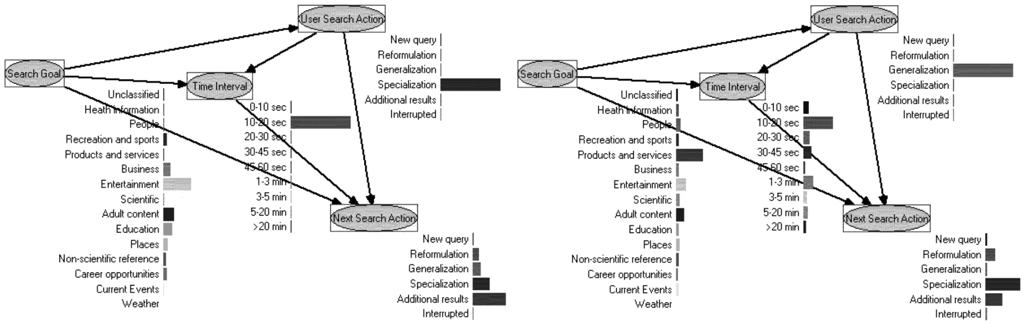
(a) (b)

Figure 4. A Bayesian-network model representing dependencies among a user's prior action, next action, and inter-query interval. The states of each variable are listed next to the variables, and probabilities of each state are captured by the length of the adjacent bars. (a) The probability distributions for *Next Search Action* and *Time Interval* show the implications of setting the variable *User Search Action* to *New query*. (b) Display of the inferred probabilities for the next search action in the case where the user composes a new query and then does not perform another search action within 20 minutes.

associated with including the context of a user's goals in the model, and in the feasibility of inferring information goals from search action and inter-query interval information. As portrayed in Figure 5, we conditioned all of the variables in the model introduced in Section 6.1 on the variable *Search Goal*, and derived conditional probabilities for the model from the tagged data set.

Figure 5a considers the specific case where a specialization of an earlier query has occurred followed by delay of 10-20 seconds before the user takes another action with the search service. We set the initial query and time interval to these states and infer probability distributions over the states of the other variables. The inferred probability distribution over the next search action shows us that exploring additional results is the most likely action (p=0.53), with additional specialization of the query occurring at about half that likelihood (p=0.26), trailed by generalization (p=0.12) and reformulation of the query (p=0.09). The inference over goals shows that a specialization followed by another action at 10-20 seconds implies that a user is most likely searching for entertainment-related information, and then, with diminishing likelihood, for adult-related content, education, business, and places.

Figure 5b shows the probability distributions resulting from a single generalization action. Given this action, the most likely informational goal is a search for products and services, the next search action will occur within 10-20 seconds, and it will most likely be a specialization. This result evinces the common pattern of a generalization followed quickly by a specialization, perhaps because the generalization resulted in too many matching results.



(a) (b)

Figure 5. Bayesian network conditioning adjacent search actions and inter-query interval on the informational goals of users. (a) Probabilistic implications of a specialization followed by another action within 10-20 seconds. (b) Probabilistic implications of performing a generalization.

6.3 Opportunities for Leveraging Models of Search Dynamics

We are impressed by the power of the Bayesian network models to identify the most likely informational goals of users and the timing and nature of the next search actions given extremely limited observations of activity. Experimentation with the models reveals that informative distributions about the informational goals of users can be generated by simply considering the timing between queries. Knowledge about the status of a user's current query and the timing of the next search action often can be used to generate relatively peaked distributions over the type of refinement that will occur next.

The ability of the probabilistic models to predict user activity from such simple observations as timing highlights opportunities for enhancing the user search experience. Future search services may leverage such models to assist a user in forming queries. For example, a search service could calculate the probabilities of next search actions based on the length of the delay noted since the last search. If the user has not performed an action within two minutes, the search service might conclude that the user is likely to desire a reformulation of his query and suggest an alternate query using synonyms of the original query—or invoke a larger help system targeted at reformulation of the query.

We believe it is feasible to detect adjacent refinements in an automated manner and to use such information in conjuction with the interquery interval to diagnose a user's informational goals with probabilistic user models. As we mentioned in Section 3.1, it is often possible to identify the refinement class of a query by examining how it differs from the previous query. An ability to identify likely user goals through detecting one or more query refinements could be exploited in a variety of ways to enhance search from the user's or provider's perspectives. Consider the likely informational goal of *Entertainment* inferred in the case displayed in Figure 5. Armed with this goal, a search service might highlight links to entertainment-related web pages. Such inferences can also be harnessed by search services to perform targeted advertising.

7 Conclusion

We have reviewed analyses of a large data set of log information capturing the interaction of users with an Internet search service. We described ontologies that we created to classify query actions and informational goals, and reviewed key statistics characterizing user web search activity. We then explored patterns of query refinement over time and reviewed several trends. We constructed Bayesian networks and demonstrated the power of these probabilistic user models for capturing relationships about the dynamics of users' search activities. We presented several illustrative examples of the use of the Bayesian networks to infer the probability of a user's next action, the time delay before taking the action, and the user's informational goal, based on a consideration of partial evidence about the status of a search. Our ongoing work is focusing on further generalizing the model to consider more complex temporal patterns of activity and exploring the variety of ways that such inferences might be leveraged to enhance the search experience for users in pursuit of specific information from large unstructured corpora.

References

Albrecht, D., Zukerman, I., Nicholson, A., and Bud, A. (1997). Towards a Bayesian model for keyhole plan recognition in large domains. In Jameson, A., Paris, C., and Tasso, C., eds., *Proceedings of the Sixth International Conference on User Modeling*. New York: Springer-Verlag. 365–376.

Charniak, E., and Goldman, R. (1993). A Bayesian model of plan recognition. *Artificial Intelligence* 64(1):53–79.

Conati, C., Gertner, A., VanLehn, K., and Druzdzel, M. (1997). Online student modeling for coached problem solving using Bayesian networks. In Jameson, A., Paris, C., , and Tasso, C., eds., *Proceedings of the Sixth International Conference on User Modeling*. New York: Springer-Verlag. 231–242.

Horvitz, E., and Barry, M. (1995). Display of information for time-critical decision making. In Besnard, P., and Hanks, S., eds., *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, 296–305. San Francisco: Morgan Kaufmann.

Horvitz, E., Breese, J., and Henrion, M. (1988). Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning, Special Issue on Uncertainty in Artificial Intelligence* 2:247–302.

Horvitz, E., Breese, J., Heckerman, D., Hovel, D., and Rommelse, D. (1998). The Lumiere Project: Bayesian User Modeling for Inferring the Goals and Needs of Software Users. In *Fourteenth Conference on Uncertainty in Artificial Intelligence*, 256–265. Morgan Kaufmann Publishers.

Jameson, A. (1995). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction* 5:193–251.

Maglio, P. P., and Barrett, R. (1997). How to Build Modeling Agents to Support Web Searchers. In Jameson, A., Paris, C., and Tasso, C., eds., *User Modeling: Proceedings of the Sixth International Conference, UM97*, 5–16. Vienna, New York: Springer Wien New York.

Pearl, J. (1991). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco: Morgan Kaufmann Publishers.

Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. (1998). Analysis of a Very Large AltaVista Query Log. Technical Report 1998-014, Digital Systems Research Center, Palo Alto, CA.

Spencer, G. (1998). Personal communication. Email correspondence between Eric Horvitz and Excite CTO, 8/30/98 and 9/9/98.