Learning Predictive Models of Memory Landmarks

Eric Horvitz (horvitz@microsoft.com)
Microsoft Research, One Microsoft Way
Redmond, WA 98052 USA

Susan Dumais (sdumais@microsoft.com)
Microsoft Research, One Microsoft Way
Redmond, WA 98052 USA

Paul Koch (paulkoch@microsoft.com)
Microsoft Research, One Microsoft Way
Redmond, WA 98052 USA

Abstract

We describe the construction of statistical models that provide inferences about the probability that subjects will consider events to be memory landmarks. We review methods and report results of experiments probing the classification accuracy and receiver-operator characteristics of the models. Then, we discuss opportunities for integrating models of memory landmarks into computing applications, and present a prototype time-line oriented content-browsing tool.

Introduction

Studies of memory support the assertion that people make use of special landmarks or anchor events for guiding recall (Shum, 1994; Smith, 1979; Smith, Glenberg & Bjork, 1978) and for remembering relationships among events (Davies & Thomson, 1988; Huttenlocher & Prohaska, 1997). Such landmarks include both public and autobiographical events. More generally, there has been significant study and modeling of episodic memory, where memories are considered to be organized by *episodes* of significant events, including such information as the location of an event, attendees, and information about events that occurred before, during, and after each memorable event (Tulving, 1983; Tulving & Thomson, 1980). Memory has been shown to also depend on the reinstatement of not only item-specific contexts, but also on more general context capturing the situation surrounding events.

We believe that automated inferences about important memory landmarks could provide the basis for new kinds of personalized computer applications and services. Rather than focusing on specific models for recall (*e.g.*, Raaijmakers & Shiffrin, 2002; Shiffrin & Steyvers, 1997), we investigate the feasibility of directly learning models of memory landmarks via supervised learning. We focus here on the construction, testing, and application of predictive models of memory landmarks, based on events drawn from users' online calendars.

We first review experiments with the construction of personalized models of memory landmarks. We describe how we construct models that can be used to infer the likelihood that events will serve as memory landmarks, reviewing the extraction of data from subjects' online calendars, the collection of assessments about landmarks with tools that enable subjects to label their calendar events, and the learning of models via Bayesian learning procedures. After reviewing the performance of the models, we describe, as a sample direction for the use of predictive models of memory landmarks in computing applications, a prototype, named MemoryLens Browser. MemoryLens Browser employs the inferences about landmarks in visualizations for browsing files and appointments. Finally, we review research directions aimed at enhancing coverage and discriminatory power of models of memory landmarks.

Accessing Events and Event Properties

We will focus on the construction of models of memory landmarks derived from users' online calendar information. Electronic encodings of calendars provide rich sources of data about events in users' lives. People who rely on electronic calendars, often encode multiple types of events in an online format. Such items include appointments, holidays, and periods of time marked to indicate such activities as travel and vacation. In large enterprises that rely on computer-based calendaring systems, appointments and events are typically managed via schemas that capture multiple properties of the events.

We developed a calendar event crawler that works with the Microsoft Outlook messaging and appointment management system. The crawler analyzes a user's online calendar to create a case library of events and properties associated with each event. The calendar crawler extracts approximately 30 properties for each event. Most of these properties are obtained directly from the online data and metadata stored for events. These properties include the *time of day* and *day of week* of events, *event duration*, *subject*, *location*, *organizer*, *number of invitees*, *relationships between the user and invitees*, the *role of the user* (*i.e.*, user was the organizer, a required invitee, or an optional invitee), *response status* of the user to appointment invitations (*i.e.*, user responded *yes*, responded *tentative*, *no response*, or *no response request made*), whether the meeting is *recurrent* or *not recurrent*, whether the time is marked *as busy or free* on the user's calendar, and the nature of the *inviting email alias* (*i.e.*, the alias used to send the meeting invitation).

In addition to properties in the database schema employed by Outlook, a subsystem of the crawler accesses the Microsoft Active Directory Service to identify organizational relationships among the user, the organizer, and the invitees, noting for example, whether the organizer and attendees are *organizational peers*, *direct reports*, *managers,* or *managers of the user's manager*.

Beyond the use of data from Outlook and Active Directory Services, we created several derived properties representing statistics about atypical situations, based on the intuition that rare contexts might be more memorable than common ones. In particular, we developed procedures for computing *atypical organizers, atypical attendees,* and *atypical locations.* We compute a measure of the rarity for these properties of events by considering the portion of all meetings over all events under consideration or for a fixed period of time (*e.g.,* events over a year) in which the property has the same value it has in the event at hand. For the studies reported here, we computed atypicality based on all events under consideration.

To compute the value of *location atypia* for events, we first compute the number of times each location has appeared in a user's calendar over a fixed period. The system then discretizes the *location atypia* variable into a set of states, capturing a range of percentiles, and the location atypia variable for each event acquires a particular value based on the rarity of the location associated with that event.

An analogous derivation is used for computing *organizer atypia* and *attendee atypia*. For these variables, all people attending all of the appointments for the fixed period under consideration are analyzed, and the portion of a subject's appointments attended or organized respectively by each attendee is noted. A meeting acquires the *organizer atypia* or *meeting atypia* value associated with the least frequent attendee or organizer of the meeting.

Building Models of Memory Landmarks

We recruited 5 participants from our organization for data collection and tagging. We asked the subjects to review a list of all of the appointments, holidays, and other annotations stored in their calendars, and to identify the subset of events that they viewed as serving as salient, memory landmarks. More specifically, we directed the subjects to do the following:

> *Please review the events on your calendar and identify those events that would serve as key memory landmarks on a timeline of events for such purposes as searching for files and appointments.*

Each subject downloaded event-collection software to automatically crawl their calendars and create a case library of labeled data. The cases typically spanned several years of presentations, trips, meetings, tasks, and holidays, and included several thousand items. We provided subjects with a memory-landmark assessment tool that lists events drawn from their online calendar within a scrollable window, ranked from most recent to most distant events. The tool provides fields, adjacent to each event, that subjects use to label items as landmark or non-landmark events. Table 1 shows the number of calendar events judged by each subject (from 1743 to 3864), and the date range of these events (from 3 to almost 10 years).

We constructed predictive models of memory landmarks from the supervised training data. We employed Bayesian-network learning methods because they enable us to visually inspect key probabilistic dependencies among variables and, in particular, to understand key variables and states of variables influencing the likelihood of events being called memory landmarks. Although detailed comparative testing of learning algorithms is not the focus of this work, we compared the Bayesian learning approach with another learning method, support vector machines (SVMs), using techniques developed by Platt (1999).

We partitioned the data into training and testing cases, with an 80/20 temporal split; that is, we built the models for each individual using the first 80% of their labeled data and evaluated the learned model on the remaining 20% of the labeled data. Because the data is temporally ordered, we used the natural temporal split rather than cross-validation on random splits. We employed a Bayesian structure-search procedure, developed by Chickering, Heckerman & Meek (1997), to build Bayesian-network models for event landmarks for each subject The procedure employs a greedy search through a large space of dependency structures and computes, for each plausible dependency structure, an approximation for the likelihood of the data given the structure. A model score is computed as a function of this likelihood and a model-prior parameter that penalizes for complexity. The model with the highest score is selected. We optimized the model-prior parameter by splitting the training set 80/20 into subtraining and subtesting data sets, respectively, and identifying a soft peak in the Bayesian score. This value of the parameter at the soft peak was used to build the model from the full training set.

We inspected the predictive models constructed for each subject, noting dependencies among key variables, the discriminatory power of variables, and classification accuracy of the models at predicting the data held out from the training procedure.

Figure 1 displays a Bayesian network built from the data from one of the participants in the study (subject S1), showing all of the variables and the dependencies among them. A sensitivity analysis demonstrated that key influencing variables in this model for discriminating whether an event is a memory landmark are the *Subject, Location string, Meeting sender, Meeting organizer, Attendees,* and whether the meeting is *Recurrent*.

We explored the strength of dependencies for variables in the model for each subject and found similar influences of key variables across subjects. For subjects in our study, *atypically long durations*, *non-recurrence of events*, a user *flagging a meeting as busy* or *out of office*, and *atypical*

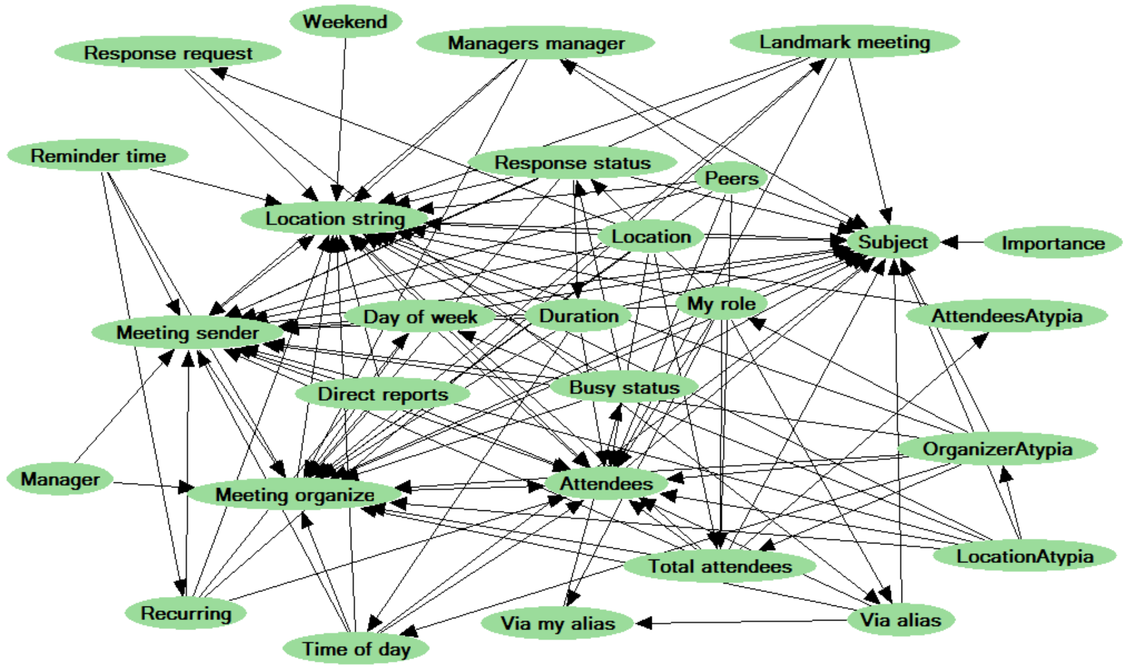

Figure 1: Bayesian network learned from online calendar data (subject S1) showing dependencies among event properties and likelihood that an event will be considered a memory landmark by a subject.

locations or *special locations* had significant influence on the inferred probability that events would be considered a landmark event. We found that meetings marked as *recurrent* meetings rarely served as memory landmarks.

Table 1 shows the classification accuracies of the learned models of landmarks. For each test case, the values of the properties of the event are identified (or computed for derived properties) and then input to the model which provides a probability that the event is a memory landmark. That is, we compute $p(\textit{Event selected as a memory landmark}|E)$, given evidence E—the multiple properties of associated with each event on the subject's calendar. The Bayesian network models range in classification accuracies for the five subjects from 0.78 to 0.95. The classification accuracies for the SVM model range from 0.71 to 0.94.

Table 1: Training data and classification accuracies for predictive models tested on hold-out data for five subjects.

	S1	S2	S3	S4	S5
Total events	3864	3740	2770	1743	1996
Train	3091	2992	2216	1394	1596
Test	773	748	554	349	400
Date range	9/1999-2/2004	8/1997-2/2004	10/1998-2/2004	1/2001-2/2004	6/1994-2/2004
SVM Accuracy	0.87	0.94	0.90	0.71	0.81
Bayes net accuracy	0.87	0.94	0.95	0.88	0.78
Bayes net accuracy - composite	0.87	0.96	0.97	0.89	0.72

We also computed the accuracy of *inter-subject* predictions. Inter-subject classification accuracy explores the potential for using models constructed from one subject's training data, or a composite model built from multiple subjects, to predict hold-out data from other participants. The last row in Table 1 shows the classification accuracy for a composite model built using the union of training data for all 5 subjects. For 4 of the 5 subjects, this composite model performs quite well, suggesting that a pre-

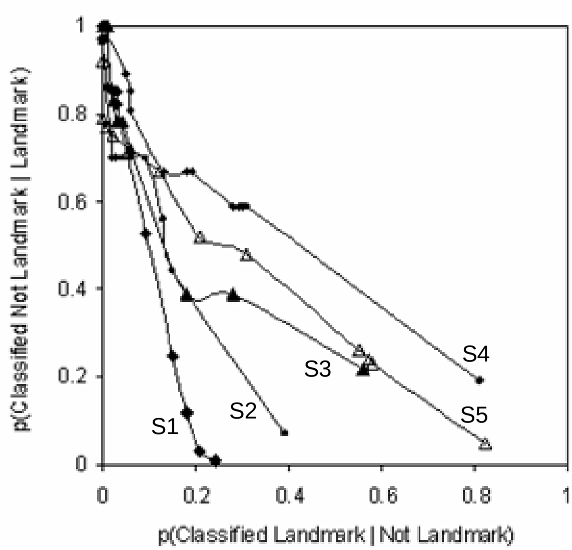

Figure 2: Receiver-operator curves showing the relationships of false negatives and false positives for five subjects at a range of thresholds on probabilities for admitting an event as a memory landmark.

trained "seed" model could be deployed. Such models could be augmented to consider highly personalized information such as variables containing specific text strings representing labels on meeting locations and subjects.

In addition to looking at overall classification accuracies, we swept out receiver-operator (ROC) curves to visualize the relationship between false negatives and false positives at different thresholds. The false-positive rate is varied by changing the threshold of the probability score that is required for classifying an event as a memorable landmark, and the corresponding false negative rate is noted. The curves for the subjects are displayed in Figure 2. We note that the ROC curves show a trend toward lower false positives and false negatives with increases in the size of the training sets.

The ROC curves are particularly important for understanding the value of employing predictive models of memory landmarks in computing applications. As we shall explore in the next section, one class of computing applications centers on the use of a user-controlled threshold on the probability of events used to identify landmark events. In such uses of predictive models of landmarks, users may be given the ability to define, *e.g.*, via a slider control, the subset of all events that should be admitted, say, for displaying within a rendering of a timeline of events. Such timelines could provide useful "memory backbones" when searching for content in a large personal store. Models for inferring the likelihood that events will serve as memory landmarks promise to endow such computing applications with the ability to minimize clutter by limiting the revelation of events to those which are likely to be useful landmarks. Moving beyond basic timelines for searching for desktop content, applications include the use of the inferential models for constructing hierarchical views of events for browsing large quantities of time-based content, such as autobiographical corpora. We now explore a sample application we have constructed to investigate prospects for harnessing statistical models of memory landmarks.

Applications of Models of Memory Landmarks

We developed a prototype, MemoryLens Browser, that demonstrates how such predictive models of memory landmarks can be used to augment computing applications. MemoryLens is in the spirit of recent work on developing tools to assist users in locating information from their personal stores (Adar, Karger & Stein, 1999; Dumais, Cutrell, Sarin, Cadiz & Jancke, 2003). A recent study by Ringel, Cutrell, Dumais & Horvitz, (2003) showed that memory landmarks can be used to help computer users find relevant search results. Users were faster at finding results when memory landmarks were added to a timeline of search results. That system employed informal, heuristic rules for selecting memory landmarks.

MemoryLens Browser provides users with a timeline of events to assist them in finding content on their computer, and uses learned models for selecting landmarks. The prototype allows users to train models of memory landmarks on events from their calendar as we described earlier. In use, the personal models of memory landmarks are used to infer the likelihood that each event drawn from the user's calendar will be considered a landmark, given multiple evidential properties extracted from unlabeled calendar items. These likelihoods are considered in generating a timeline of inferred landmarks adjacent to files gathered from across a user's file system. Since some calendar events are private and users might be reluctant to have them displayed in interfaces, we could add a capability for marking events as private and omitting these events from display in public settings.

A screenshot of the user interface of MemoryLens Browser is displayed in Figure 3. Thumbnails of file types are sorted in the right-hand column of the browser (Items), in a traditional time-sorted view. Within the left-hand column (Date), a list of relevant dates associated with the files are displayed. The middle column (Events) contains memory landmarks that have a landmark probability exceeding a user-set threshold. The titles of memory landmarks are displayed in the appropriate temporal location, adjacent to the files. The files are positioned along the timeline, based on the times that they were created or last modified.

An event-detail slider control provides users with a means of changing the threshold on the inferred likelihood of memory landmark that is required for displaying events. Only calendar items representing events that have a probability of being a landmark that is greater than a user-set threshold are displayed. As the slider is moved from "most memorable" to "least memorable," the required probability for display of events is lowered, thus bringing in greater numbers of events.

Figure 3 shows three different screenshots of the graphical interface of MemoryLens, each representing a different setting of the probability threshold for the same span of time. The view at the right is set to the highest probability threshold, thus revealing the fewest events. In this case, only the events representing two major conferences, for which the subject had to travel afar to attend, are displayed. As the threshold is lowered, a wedding, an editorial board meeting, a conference call, and a one-on-one meeting are included in the display. Further diminishing of the threshold for admitting events brings larger numbers of events into view. Beyond the use of thresholds for admitting events into the landmarks column, the saturation of color of the text used to title events is faded as the probability of memory landmark diminishes—providing an additional cue about the likelihood that the event would be viewed as a memory landmark.

We have been explored the ability of models with the discriminatory performance represented by the family of ROC curves displayed in Figure 2, to construct useful time-line views. Such views should contain recognizable landmarks, while bypassing the clutter associated with showing a great number of events, and should allow users to

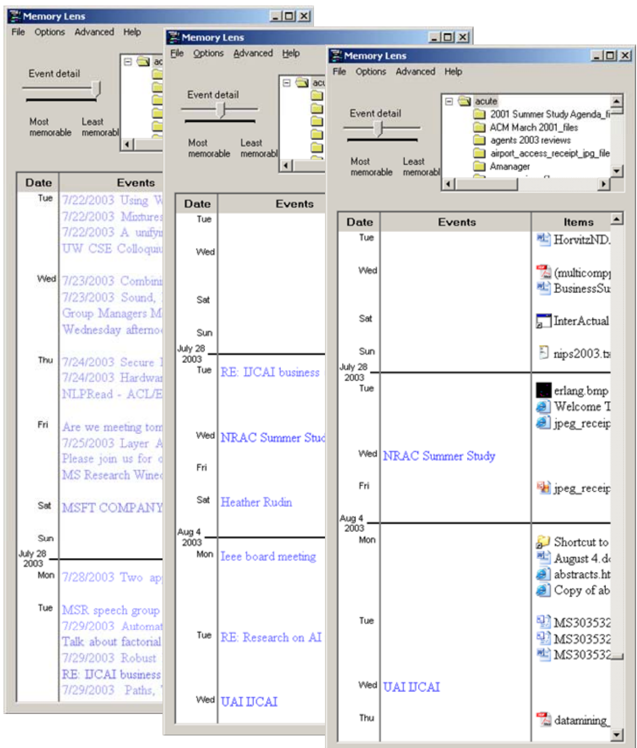

Figure 3: MemoryLens Browser with memory-landmark timeline displayed at three different settings of the threshold on the likelihood required for an event to be considered a memory landmark.

work with such models in an exploratory, interactive manner (Horvitz, 1999) with tools embodied in MemoryLens' controls and display.

To relay a qualitative feel for the quality of timelines constructed with the use of the predictive models that we have generated, consider the ROC curve for a model of subject S1. The curve tells us that, at a probability threshold for accepting events as landmarks where ninety percent of the events on the timeline are correctly identified as important landmarks, fifty percent of the important landmarks will not be displayed. Such precision and recall may be quite tolerable for navigating to target periods of time, given the overall density of landmarks for users; we found that subjects in our study typically showed 2-4 landmark events per week over the span of their assessments. A recall of half of these events would still tend to identify landmark events for every week.

Understanding the comprehensive value of providing users with selective views of landmark events on timelines will require detailed user studies of the use of specific prototypes and artifacts. We have recently distributed the prototype to a limited group of users within our organization and are pursuing studies of the value of specific designs built on predictive models of memory landmarks. Such studies promise to enhance our understanding of the sensitivity of particular features and services to the performance of the predictive models.

Research Directions

We have focused in this paper on the construction and performance of predictive models that can be used to infer the probability that events drawn from online calendars will be considered memory landmarks by users. We provided, as a motivating example, a prototype application to highlight potential applications. Although we did not dwelled on comprehensive evaluations of the value of the use of memory landmarks in such prototypes, we are nonetheless interested in pursuing a deeper understanding of the value to users of rendering memory landmarks of different types and in different settings. .

In addition to pursuing a better understanding of the value of memory landmarks for users performing search and retrieval in computing applications, we are exploring several avenues of opportunity with refining and extending models of memory landmarks.

New Classes of Evocative Features. We are exploring the value of adding new observations features to the modeling of memory landmarks. For example, we are interested in the value of introducing a consideration of observations that assist with inferences about the likelihood that a meeting has been attended, given desktop activity over time and the sensed location of systems. Prior work has demonstrated the feasibility of performing relatively accurate inferences about the likelihood that a meeting has been or will be attended, based on an analysis of meeting properties, including activity monitored during meetings (Horvitz, Jacobs, & Hovel, 1999; Horvitz, Koch, Kadie & Jacobs, 2002; Mynatt & Tullio, 2001). Information about the likelihood of meeting attendance promises to have influence on the probability that the meeting will be viewed as a memory landmark. Other factors include capture and analysis of acoustical energy during meetings, and preparatory or follow-up activity associated with appointments.

Beyond Calendar Events. Events captured on users' calendars are convenient, but represent only a small subset of "events" users may wish to have captured, reasoned about, and harnessed in computing applications. We are interested in building and refining predictive models for other items that could serve as additional memory landmarks or bolster event landmarks by providing richer context. As an example, we are pursuing, in a parallel project, the construction of predictive models that can identify the likelihood that images drawn from a large online personal photo library represent landmark events. We build on past research on photobrowsing tools that explored the use of image analysis, coupled with several heuristics, to select subsets of pictures from large photolibraries (Platt, 2000). In learning models of landmark images, we consider as observations sets of features derived from camera metadata and multiple image analyses.

In another realm, we are interested in learning from data predictive models that can automatically select the most important national and world developments, as captured by news events over time.

Beyond calendar-centric events, images and news, online interactions, communications, and patterns of interactions with computer-based content may serve as memory landmarks. For example, particular email exchanges, or documents associated with clusters of items that have been reviewed or created in periods of activity over time may provide an important source of landmark events.

Taken together, multiple models of memory landmarks may be used in conjunction to build rich, multi-source timelines, providing views at different scales of time and for different quantities of events, triaged by the likelihood that events will serve as memory landmarks.

Learning Models of Forgetting. Finally, we believe that there are opportunities for developing analogous statistical models of important events and tasks that will be forgotten without reminding. Recent longitudinal studies of office workers have identified classes of important events that are forgotten and have demonstrated the value of heuristics for ways to provide reminders about such events (Czerwinski & Horvitz, 2002). Beyond applications for people in good health, we see the feasibility of developing models for supporting people suffering with pathologies of memory associated with various forms of dementia. We foresee the value of developing such predictive models and joining them with decision-theoretic methods that can guide decisions about if, when, and how to remind people about things they are likely to forget, balancing the informational value and disruptiveness of such reminding actions (Horvitz, E., 1999).

Summary

We reviewed research highlighting prospects for developing and harnessing predictive models of events that will be viewed as landmarks. We focused in particular on the construction and evaluation of models that infer landmarks from events drawn from subjects' calendars. After reviewing the classification and ROC curves associated with training sets obtained from five subjects, we discussed the potential to employ predictive models of memory landmarks in computing applications. We described as an example, the MemoryLens Browser prototype. Before concluding, we touched on several current research directions, including opportunities to perform additional studies to evaluate the value of displaying memory landmarks in search tasks and developing models of landmark events for online images, news stories, and other items encountered or created by users in their daily lives that might be encoded as important landmarks in episodic memory.

Acknowledgments

We thank Johnson Apacible for assisting with data collection and analysis.

References

Adar, E., Karger, D. R. & Stein, L. (1999). Haystack: Per-user information environments. *Proceedings of CIKM'99*, 413-422.

Chickering, D.M., Heckerman, D. & Meek, C. (1997). A Bayesian approach to learning Bayesian networks with local structure. *Proceedings of UAI '97*, 80-89.

Czerwinski, M. & Horvitz, E. (2002). An investigation of memory for daily computing events. *Proceedings of HCI 2002*, 230-245.

Davies, G. & Thomson, D., eds. (1998). *Memory in Context: Context in Memory*. Wiley: Chichester, England.

Dumais, S., Cutrell, E., Sarin, R., Cadiz J. & Jancke, G. (2003). Stuff I've Seen: A System for personal information retrieval. *Proceedings of 26th International SIGIR 2003*. 72-79.

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. *Proceedings of the SIGCHI '99*, 159-166.

Horvitz, E., Jacobs, A. & Hovel, D. (1999). Attention-sensitive alerting. *Proceedings of UAI '99*, 305-313.

Horvitz, E., Koch, P, Kadie, K. & Jacobs, A. (2002). Coordinate: Probabilistic forecasting of presence and availability. *Proceedings of UAI '02*, 224-233.

Huttenlocher, J. & Prohaska, V. (1997). *Reconstructing the Times of Past Events. Memory for Everyday and Emotional Events.* Mahwah, NJ: Lawrence Erlbaum Associates, 165-179.

Mynatt, B. & Tullio, J. (2001). Inferring calendar event attendance. In: *Intelligent User Interfaces 2001*,. 121-128.

Platt, J. (1999). Fast training of support vector machines using sequential minimal optimization. In Smola, A., Bartlett, P., Scholkopf, B. & Schuurmans, D. (Eds.) *Advanced in Large Margin Classifiers*, 158-208.

Platt, J. (2000). AutoAlbum: Clustering digital photographs using probabilistic model merging. *IEEE Workshop on Content-Based Access of Image and Video Libraries 2000*, 96-100.

Raaijmakers, J. G. W. & Shiffrin, R. M. (2002). Models of memory. In Pashler, H. & Medin, D. (Eds.) *Stevens Handbook of Psychology 3rd Edition*, *Vol. 2: Memory and Cognitive Processes*, 43-76.

Ringel, M., Cutrell, E., Dumais, S. & Horvitz, E. (2003). Milestones in time: The value of landmarks in retrieving information from personal stores. *Interact 2003, Ninth IFIP Interact 2003*, 184-191.

Shiffrin, R.M. & Steyvers, M. (1997). A model for recognition memory: REM retrieving effectively from memory. *Psychonomic Bulletin & Review*, 4(2):145 166.

Shum, M. (1994). The role of temporal landmarks in autobiographical memory processes. *Psychological Bulletin*, *124*, 423-442.

Smith, S. M. (1979). Remembering in and out of context. *Journal of Experimental Psychology: Human Learning and Memory*, *5*: 460-471.

Smith, S. M., Glenberg, A. & Bjork, R. A. (1978). Environmental context and human memory. *Memory & Cognition, 6(4)*, 342-353.

Tulving, E. (1983). *Elements of Episodic Memory*. Oxford University Press.

Tulving, E. & Thomson, D. (1980). Encoding specificity and retrieval processes in episodic memory. *Psychological Review 80*, 352-373.