Haystack: Per-User Information Environments
MIT9904-08

Progress Report: July 1, 2002—December 31, 2002

David Karger

Project Overview

The Haystack project is investigating the ways in which electronic infrastructure can be used to triangulate among the different knowledges of an individual, of collegial communities to which we belong, and of the world at large. Its infrastructure consists of a community of independent, interacting information repositories, each customized to its individual user. It provides automated data gathering (through active observation of user activity), customized information collection, and adaptation to individual query needs. It also facilities inter-haystack collaboration, exposing (public subsets of) the tremendous wealth of individual knowledge that each of us currently has locked up in our personal information spaces. The Haystack-NTT project involves augmenting its customization, learning and adaptation, and inter-haystack communication.

The Haystack project stands upon three research pillars. We study semi-structured databases and semantic-web ontologies as tools for representing all of the knowledge useful to an individual. We explore user interfaces that present this rich information to the user in an effective way, letting them search, navigate, and manipulate their information. And we investigate deductive agents and machine learning as tools to reduce an individuals information management burden.

Progress Through December 2002

This past six months have been devoted to turning Haystack into a deployable system, by performing preliminary user studies to get a sense of what individuals are going to need to make haystack a useful tool for them, and by fleshing out all the components needed to make create that useful system. We are in the final stages of preparing that deployable tool.

We carried out a number of user studies. One “ecological” study involved close observation of 30 individuals as they used standard information management tools over a week, looking for patterns, habits and problems that would influence the way we design haystack. A second study exposed a few users to a haystack prototype and gathered information about what they found useful and what they found confusing about the system. Results were promising; for example, our use of “context menus” that let the user right click on any object in the system and see a list of all actions that make sense for that object, was very well received. Certain confusions about the user interface were generally due to “lowered expectations”---the users were used to certain actions being quite difficult to perform under standard user interfaces, and tried to do things the “hard way” when haystack offered a much simpler approach. A third, lower level study, evaluated a prototype tool for letting users manage collections of information. Instead of dragging items into folder hierarchies, the typical approach in current interfaces, we let users specify overlapping (rather than hierarchical) categories using check boxes. We then offered a natural browsing framework to let users navigate their categorized data hierarchically by invoking or eliminating certain category restrictions. Our experiment showed that users were better able to retrieve using our categorization mechanism than using standard folder hierarchies.

On the deployment front, we have chosen, as our first motivation application, a tool for managing bibliographic references. This activity is quite common within the laboratory, as any student writing a research paper needs to collect a bibliography for it. Bibliography management exercises many of the interesting features of Haystack: the objects being managed have rich metadata; users have collections of them; they tend to build new subcollections by navigating their collections and selecting relevant items, and they need to incorporate external information when they find it.

We have also undertaken to exploit haystack's special capabilities in the management of email. We have developed machine learning tools that watch a user categorize their email and learn over time how to do that categorization automatically. The most obvious use is to block spam email, but there is additional benefit in, for example, auto-categorizing information incoming email into "work" and "recreational" categories to be read at different times.

Our work on this deployable system highlighted performance limitations in our underlying RDF database; thus we redeveloped (for the fourth and hopefully last time) a higher performance RDF database to support our system. Our experience has suggested some ideas of more general relevance to the development of RDF databases.

Research Plan for the Next Six Months

Our current focus is on deploying a system. We continue to fix bugs, complete user interface elements, and add functionality we expect will motivate users to use the system (in addition to the bibliography tool, we are attempting to provide standard PIM functionality, calendar and meeting tools, and email management). When the system is deployed, we will begin long term user studies. Our hypothesis, that users benefit from a customized information retrieval system, will be tested by examining, over time, how users customize their haystacks. If no customization occurs, this will be evidence that our hypothesis is incorrect. If haystacks do become customized over time, we will mine the customization information to find out what types of customizations are most useful to users, and continue developing haystack to make such customization capabilities more effective.