Haystack: Per-User Information Environments
MIT9904-08

Progress Report: July 1, 2000—December 31, 2000

David Karger and Lynn Andrea Stein

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.

Progress Through December 2000

Our progress in the past two months has been twofold. First, we have carried out a complete redesign of the low-level storage and communication layer of the Haystack system. Second, we have recruited a number of PhD students, familiarized them with the Haystack project, and begun working with them on the higher level user-interface, machine learning, and collaborative parts of the Haystack research project.

Regarding the low level system, we concluded that Haystack's data model (a close cousin of RDF) and the system used to store it was quite general and plausibly useful in a wider context. We made the decision to separate the storage of our data model from the higher-level haystack applications that would make use of that information.

Besides improving the modularity of the Haystack system, this creates the opportunity to let applications other than Haystack store other information in the same storage module. At the limit, one might consider a semistructured data repository taking over as the "file system" for a large family of applications that wish to share information, saving the work of writing translation tools to pass information from one application to the other. The various data-extraction services which we have written for Haystack (such as a tool for identifying titles and authors in postscript documents) can run as independent applications, communicating only through their accesses to the data model.

With an eye towards the information-navigation goals of Haystack, we have augmented the basic storage role with two capabilities. To allow for the detection and correction of invalid data in the repository, the storage module maintains information about the source of every piece of information from outside the repository, as well as chains of deduction that led to particular conclusions. This allows us to use the repository as a "truth maintenance system" that updates its deductions to reflect changes in the underlying system.

In the past few months, we have enrolled four new PhD students to join the one already working with us on Haystack. After a time spent becoming familiar with the goals of the project and learning their way around the system, these students have identified particular aspects of the project that they wish to pursue. With this additional manpower, we are finally able to begin serious development of the higher level components of the haystack system-—learning modules, collaborative filters, and user interfaces—that reside above our low-level data model.

Research Plan for the Next Six Months

Our plans for the next sixth months are as follows:

• Complete the implementation of the now-designed low-level storage module. Two Masters students graduating in June are tasked with this project.

• Begin to develop new probabilistic document models to be used to drive information retrieval and machine learning algorithms within our system. Develop mechanisms to allow machine learning, which is typically applied only to textual information retrieval, to also be used in retrieving the semistructured information stored in the Haystack data model. Two PhD students are tasked with this project.

• Begin to develop tools for seeking useful information in widely distributed web sites, as a preliminary step towards aggregating information from a large community of Haystacks. Important issues include the identification of web sites likely to have information relevant to the query (using machine learning, based on the user's past interactions with those web sites), and interpreting the results produced by those web sites to decide what to show to the user. One PhD student is tasked with this project.

• Begin to develop user interfaces to Haystack that are optimized for the kind of learning and personalization that we wish to explore. As a preliminary step, we are beginning a study of the way people work with their own information. The key questions include how people structure their own information given the tools available to them (directories, book-marks, personal web pages), and how they use these structures to retrieve information that they themselves have stored in the past. Two PhD students are tasked with this project. Mark Ackerman, and expert in Human Computer Interaction, is helping to lead this effort.