Haystack: Per-user Information Environments
MIT9904-08
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.

Planned work for next year.

Shortly after June, we intend to have a robust implementation of the low-level storage layer of the Haystack system, as well as a basic user interface. With that in place, we can rapidly advance our work on adaptation, user interfaces, and distributed/collaborative aspects of the Haystack system.

As I discussed in the most recent progress report, we came to realize that our semistructured, RDF-like data model was a natural storage model not just for Haystack, but also for many of the other applications being developed in the Oxygen project. We've been participating in an ongoing discussion with other Oxygen researchers to make concrete the specific needs of these projects. In parallel, we have been implementing our own low-level storage module with these needs in mind, hoping to offer a general-purpose tool. Letting many groups share one storage module will simplify inter-project sharing of data. In particular, it should easily broaden the data available to the Haystack system for retrieval and adaptation purposes.

In parallel, Mark Ackerman has been completing the development of a basic user interface that lets users navigate the information in a semistructured repository such as ours. This display handles general semistructured information but can also be configured to display specialized structures (such as those corresponding to documents) in specialized ways.

Finally, over the past semester, we have been bringing up to speed a number of new graduate students. By June, each will be intimately familiar with the goals and infrastructure of Haystack and will be ready to embark on a well defined project that advances the Haystack framework.

With this background, we plan to embark on the following activities over the next year.

• With Mark Ackerman, Christine Alvarado and Jaime Teevan, we have begun planning a lengthy user study to explore the ways people organize their own personally-held information (as they plan for future retreival) and

the way they retreive it at need. Our goal is to identify the commonly used contextual cues that help people remember where they put things. This information will influence the design of the more powerful user interface we intend to layer atop Ackerman's basic data browser. It will also affect the design of Haystack's automatic data-gathering tools, which it uses to autonomously create the context that will help the user find their information later.

• With Kai Shih, we will develop tools that let Haystack gather information from outside sources. If Haystack knows so much about its user, it makes sense to let Haystack mediate the user's interaction with the world of information on the web. To avoid being swamped by the complexity of this problem, we have chosen to focus on the problem of collecting "news"—to adaptively provide the user with their own daily gleaning of what is interesting to them on the web.

We have developed an architecture divided into a "gatherer" that collects information from the external web, a "presenter" that displays this information to the user, and a learning module that decides what to collect and how to present it, based on observed user behavior. The presenter and learner are, of course, already parts of the haystack system, but they will be specialized for the news application. For example, the learner must be able to handle the selection of sites likely to contain interesting information for this user, and site-specific queries that will find this information if it is available. We must also give serious attention to the gatherer. The basic idea goal of this component is for it to decompose a web site into logical, semistructured information that can then be integrated with the information coming from many other web sites. This element is likely to highlight the fact that content providers cannot expect to own the "eyeballs" of their users (and thus call into question the advertising model that supports much of the web). Rather, they will provide data elements which are gathered by automated agents and assembled for the benefit of the user.

Given the wide variety of web sites, it is impossible to hard-code disasembly rules for all of them. Thus we are developing machine learning tools that can disasemble new sites based on web-site structures studied at old sites. In addition, we are developing interfaces that let a user show the system by example how to disasemble a site of interest.

• If the presentation of content in human-readable web pages is a passing phenomenon, we must think about the way content-provision sites will provide information to content-gathering agents. The semantic web project at the World Wide Web consortium aims to study this problem. And individual user Haystacks serve as natural content-providing sites on which to experiment. Dennis Quan will be studying the question of how Haystacks can exchange information with each other. Using semantic web techniques, we want a Haystack to be able to explain what information it has and why it likely to be of interest to another user. Since the first Haystack structures information according to the idiosyncratic needs of its own user, there is a "language barrier" that must be crossed to convey information to a different user. The semantic web provides a framework for doing this. The Haystack project will provide an important test of this framework and suggest ways that it needs to be extended.

Should Kai's work on gathering and Dennis' work on the semantic web be successful, it will be natural to combine them: Kai's gatherer can be seen as a wrapper for web sites that turns them into semantic-web enabled sites, which can then be consumed into Haystacks using Dennis' tools.

• Nick Matsakis is studying machine learning methods in information retrieval. Once the haystack data store is available, he will begin studying the Haystack specific problem of machine learning over a structured repository. While a great deal has been done with machine learning in text repositories, the semistructured model raises new questions (such as a bluring of the distinction between data item and feature) that will have to be addressed in new ways.

State of the Collaboration.

There is little collaboration information to report. We exchanged some early emails with our NTT contact, but things have not progressed beyond that. We have not made any visits, nor received any proposals for visits here (though we would be happy to host some). No joint publications have occurred.