Learning user information interests through the extraction of semantically significant phrases

Bruce Krulwich and Chad Burkey
Center for Strategic Technology Research
Andersen Consulting LLP
3773 Willow Drive, Northbrook, IL 60062
Internet: {krulwich, burkey} @ cstar.ac.com

Abstract

InformationFinder is an intelligent agent that learns user information interests from sets of messages or other on-line documents that users have classified. While this problem has been addressed by a number of recent research initiatives, InformationFinder's approach is innovative in a number of ways. First, the agent uses heuristics to extract significant phrases from documents for learning rather than use standard mathematical techniques. This enables it to learn highly general search criteria based on a small number of sample documents. Second, the agent learns standard decision trees for each user category. These decision trees are easily transformed into search query strings for standard search systems rather than requiring specialized search engines.

1. Large-scale on-line information systems

A growing number of businesses and institutions are using distributed information repositories to store large numbers of documents of various types. The growth of Internet services such as the World Wide Web and Gopher, the continued increase in use of Usenet bulletin boards, and the emergence on the market of distributed database platforms such as Lotus Notes™ all enable organizations of any size to collect and organize large heterogeneous collections of documents ranging from working notes, memos and electronic mail to complete reports, proposals, design documentation, and databases. However, traditional techniques for identifying and gathering relevant documents become unmanageable when the organizations and document collections get very large.

This problem exists outside of corporate information repositories as well. On the Internet's World Wide Web, for instance, it is impossible to even attempt to see all pages that may be of interest. It is equally impossible to simply scan all of the news media (such as newspaper and magazine articles) that are becoming available on the Web. The same is true of other information systems based on the Internet and other world-wide networks, such as Usenet bulletin boards

This paper describes an intelligent agent developed to address this problem similar to research systems under development for similar tasks [Holte and Drummond, 1994; Knoblock and Arens, 1994; Levy *et. al.*, 1994; Pazzani *et. al.*, 1995] or for other tasks such as e-mail filtering or Usenet message filtering. The agent learns a search query string for each of the user's interest categories, and searches nightly for new documents that match these interests to send to the user. Our most significant finding is that effective results depend largely on extracting high-quality indicator phrases from the documents for input to the learning algorithm and less on the particular induction algorithms employed.

We present our solution in the context of a Lotus Notes system, consisting of electronic mail, bulletin boards, news services, and databases, but our approach is equally applicable to both the World Wide Web and Usenet. We are planning to make our InformationFinder publicly available for these systems in the near future.

2. Learning user interests

Figure 1 shows a user reading a document about Java, a language for Internet development. Upon reading this document, the user decides that it is representative of his interest in Java. To indicate this to InfoFinder the user selects the "smiley face" icon in the upper right corner. The agent asks the user to categorize his interest in the document, which he gives as "Java." These categories are fully user-specified and need not be given names representative of the content: they are used simply for grouping of documents (e.g., [Gil, 1994; Lieberman, 1994]) and communication with the user. The document is copied into a collection of sample documents for subsequent processing.

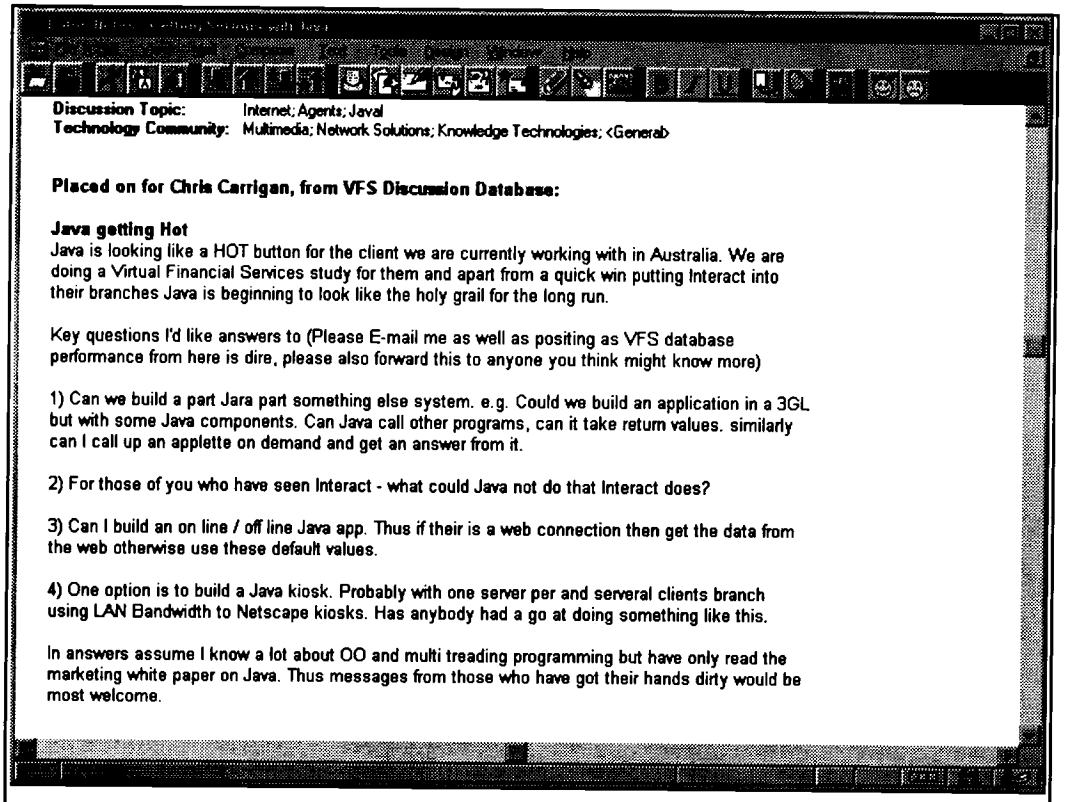

Figure 1: A document that matches the user's interest in "Java"

After the user has selected a number of documents as being relevant or not relevant to a particular category, InformationFinder will use these sample sets to learn search query strings for each category. These search strings are then used to send the user new messages on a routine basis that match one of his interests. The agent does this using a three step process:

1. Extract semantically significant phrases from each document.
2. Learn decision trees for each category based on the extracted phrases.
3. Transform each decision tree into a boolean search query string.

The first step in processing sample documents, and the step in which InformationFinder's approach is unique, is to use heuristics to extract significant phrases from the document text. These heuristics are based on the observation that document authors tend to use syntactic methods to delineate key phrases or ideas in documents, such as putting them in italics, identifying them with acronyms, or the like. These heuristics are discussed in detail elsewhere [Krulwich and Burkey, 1995a, 1995b].

In our example, InfoFinder extracts a number of topic phrases from the document, such as Java, HotJava, LAN development, Multimedia, and NetScape. These and the other phrases shown in Figure 2 appeared in the document in a syntactically notable form. The heuristics that InfoFinder uses are designed to be quite liberal in extracting phrases on the assumption that phrases that do not represent document content will be discarded in the process of induction.

After the user has selected a number of sample documents, InfoFinder uses them to learn a general description of the user's category interests. This is handled separately for each of the user's categories, enabling the system to focus on learning a characterization of presumably similar documents. Because each sample document has been effectively reduced to a set of significant phrases, InfoFinder can carry out its next task, induction of a decision tree, using a straightforward variant of ID3 [Quinlan, 1983]. While this has been shown previously to be less effective than other learning algorithms (e.g., [Pazzani *et. al.*, 1995]), the effective extraction of significant phrases appears to compensate and enables highly effective with more straightforward induction algorithms.

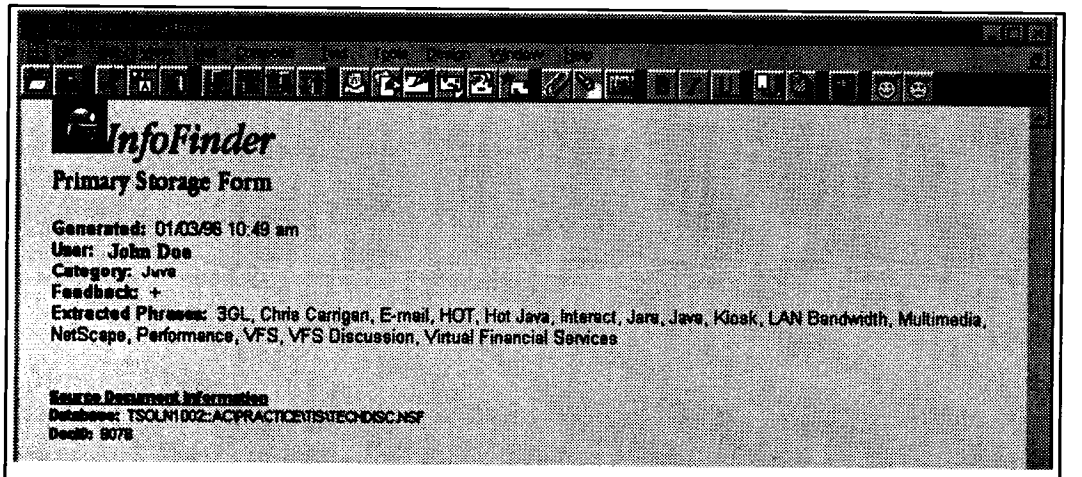

Figure 2: Phrases extracted from the sample document in Figure 2

A learned decision tree for the Java category is shown in Figure 3. It is clear from the tree that InfoFinder has been able to focus primarily on the significant text in the documents without being sidetracked to spurious but common text. Additionally, this learning is based on only 14 sample documents rather than the much larger numbers necessary in other approaches.

Once InfoFinder has induced a decision tree, it transforms it into a boolean query string for standard search engines. As is evident from the query string in Figure 3, this transformation is quite straightforward. The learned query strings are then used to search new documents for those of interest to the users, and these matching documents are sent to the user routinely. It is important to note that this process must certainly be an iterative one, with the user giving incremental feedback and further instruction to the agent and thereby converging to an effective search criterion.

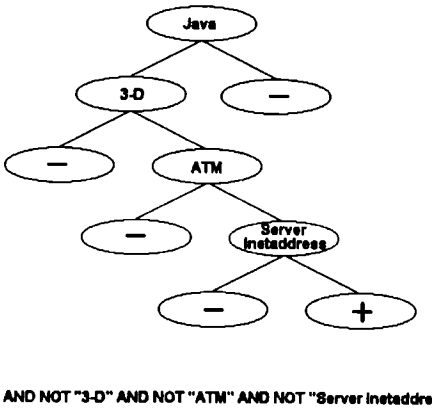

Figure 3: A learned decision tree and search query for category "Java"

3. Summary and future work

We have described a system under development that learns user preferences in dynamic message systems. While most previous research in this area has focused on induction algorithms, we have instead focused on extracting high-quality learning inputs from the documents. This allows us to get high quality results that better match the user's expectations with less computational cost.

Future research will focus on four areas. First, better phrase extraction heuristics are necessary for handling formatted text. Second, the system metaphor should be extended to support working groups. Third, more powerful induction methods will be considered and applied. Fourth, the approach will be applied to other information sources, such as World Wide Web documents or Usenet bulletin boards.

More generally, our approach raises the question of what other agent functionality can be achieved using document processing techniques such as significant phrase extraction, inductive learning, and document search. We are beginning development of several agents based on this techniques. We are also investigating the application of other core document processing techniques, such as complex schema matching and message sequence modeling, to intelligent agent tasks.