# The ContactFinder agent: Answering bulletin board questions with referrals

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

## Abstract

ContactFinder is an intelligent agent whose approach to assisting users is valuable and innovative in the following four ways. First, ContactFinder operates proactively in reading and responding to messages on electronic bulletin boards rather than acting in response to user queries. Second, ContactFinder assists users by referring them to other people who can help them, rather than attempting to find information that directly answers the user's specific question. Third, ContactFinder categorizes messages and extracts their topic areas using a set of heuristics that are very efficient and demonstrably highly effective. Fourth, ContactFinder posts its referrals back to the bulletin boards rather than simply communicating with specific users, to increase the information density and connectivity of the system. This paper discusses these aspects of the system and demonstrates their effectiveness in over six months of use on a large-scale internal bulletin board.

## 1. Electronic information systems

The explosive growth of the Internet by individuals and corporations, and the growing use of corporate information repositories based on Internet technology or systems such as Lotus Notes™, has led to an unprecedented number of people using network-based systems for finding solutions to problems. In addition to document browsing systems such as the Internet's World Wide Web, a continuing interest remains in electronic bulletin board systems that allow large numbers of distributed users discuss issues, ask questions, and give answers.

Unfortunately, a number of operational problems make it difficult to get high quality, fast answers to questions on bulletin boards, or to search bulletin boards for previous messages that can help solve a problem. First, the explosion in the number of bulletin boards makes it less likely that true experts will read any single bulletin board on a very frequent basis. Second, the increased volume of messages on these systems leads to more frequent routine deletion or migration of messages. For a user trying to find a quick solution to a problem, bulletin boards can be as frustrating as they are useful.

This paper describes an intelligent agent called ContactFinder, that has been developed to address this problem. ContactFinder is similar to intelligent agents under development for question answering [Hammond *et. al.,* 1995], e-mail filtering [Maes and Kozierok, 1993; Lashkari *et. al.,* 1994], Usenet message filtering [Sheth, 1994], or other information search and retrieval domains [Holte and Drummond, 1994; Knoblock and Arens, 1994; Levy *et. al.,* 1994; Pazzani *et. al.,* 1996; Krulwich and Burkey, 1996]. Like these other systems, ContactFinder extracts information from a large number of documents in order to present it to users in a more focused and productive fashion.

Unlike these previous approaches, however, ContactFinder's task is not to present the user with a subset of the information that can be used directly in problem solving. The agent instead keeps track of people who are key contacts in various topic areas, and helps question-askers by referring them to the appropriate key contact. More specifically, ContactFinder monitors the bulletin board for indications of message authors who are key contacts in some specific area, and stores these contacts for later use. The agent simultaneously watches for questions, and responds to the questions with a referral.

This is a very valuable function for an intelligent agent to perform for several reasons. First, an agent that attempts to provide information that is directly relevant to the user's goals will always be limited by the information that is available. While this is not a problem in solving problems that are very basic or frequently asked [Hammond *et. al.,* 1995], it may make it difficult to be helpful in novel or very focused situations. In such a situation, however, a referral to a human contact will be available more often [Kantz and Selman, 1996]. As we discuss in detail in section 5, ContactFinder has been able to make a referral for over 13% of the questions in a large-scale technology-related bulletin board, most of which were for very specific questions. Second, extracting contacts and facilitating human expertise transfer fits very well into current work styles and facilitates good learning from bulletin boards, which make the system easier to apply and test.

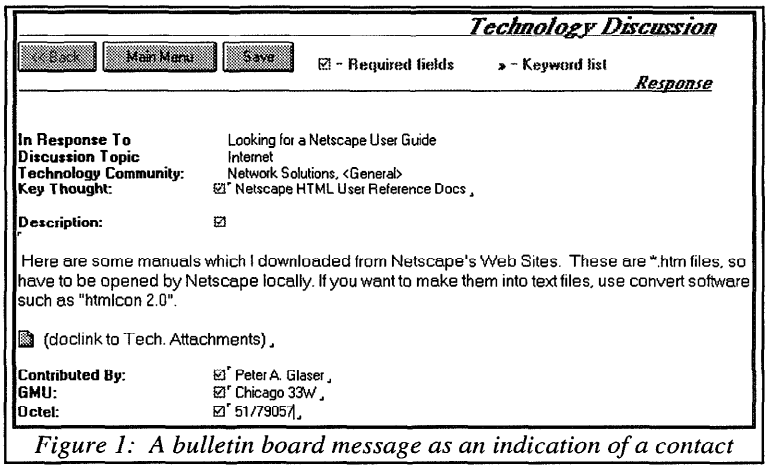


*Figure 1: A bulletin board message as an indication of a contact*

ContactFinder's processing happens in two phases. The first phase scans the new documents in the information repositories and searches for indications of key contacts in any technical area. It extracts the contacts and their technical areas, and stores them in its own database. In the second phase, ContactFinder scans on-line discussions for questions. It extracts the topics of the questions and checks if it has a contact to give as a referral on those topic areas. If it does, it responds to the question with a referral.[1] This referral gives the name and contact information, along with a reference to the previous documents that served as the basis for the referral.

## 2. Extracting key contacts

Figure 1 shows a bulletin board message that can provide indications of a key contact. A number of issues arise in looking at messages such as these as sources of contact information. First, while it is clear to people reading the message that it is a response to a question, and given the content is probably a good indication that the author is an expert or key contact in the relevant topic areas, it is not trivial to determine this in an automated fashion. Second, while this particular bulletin board system supports user-specified keywords, they do not contain enough detail to effectively classify the author's areas of knowledge.

ContactFinder approaches the problem of extracting key contacts from text messages by using heuristics that are specifically designed for extracting topic areas and contact information from text documents. Rather that attempt to process the document in a general fashion, it simply searches for indications of an appropriate contact, and looks locally at that point in the document for a name and contact information. This approach, very focused information extraction instead of general document understanding, has proven to be highly effective, as we discuss in section 5.

The most critical and difficult task that ContactFinder carries out in phase one is to extract topic areas from each document which serve as a description of the content areas for the extracted contact. The process used to achieve this in both phases of ContactFinder, for extracted contacts and for subsequent questions, is to look for phrases that are delimited by syntactic devices as central to the meaning of the message [Swaminathan, 1993; Krulwich, 1995]. These methods are discussed in detail in section 4.

Figure 2 shows ContactFinder extracting contact information from the discussion document like the one in Figure 1.[2] The top of the screen shows the document information and its contents. The bottom shows the contact that was extracted and the topic areas, along with the heuristic basis for the extraction. In this case, the document in Figure 1 was a response to a message that ContactFinder examined and classified as a question, and the response itself was not a further question. Other bases for contact extractions include offers to be contacted and explicit referrals to third parties.

## 3. Answering questions with referrals

ContactFinder's second phase is to find questions in the bulletin board, extract their topic areas, and search for previously-extracted contacts to give as referrals. Figure 3 shows such a question in a discussion group which asks about the same product mentioned in Figure 1. ContactFinder finds this document, determines heuristically that it is a question, extracts the topic areas, and searches its database for an appropriate referral. In this case it finds the expert extracted in the previous section.
For the message in Figure 3, ContactFinder realizes that this is a question based on the phrase "Does anyone" with the question mark at the end of the sentence. It first extracts the topic indicators using the same methods as are used in

[1] For our discussion in this paper we are omitting logistical details, such as human confirmation of contact accuracy and topic area prior to public referral.

[2] Note that the display shown in Figure 2 will never be seen by a user, since the process is run on the information repositories in background. This display is used for explanation and demonstration only.

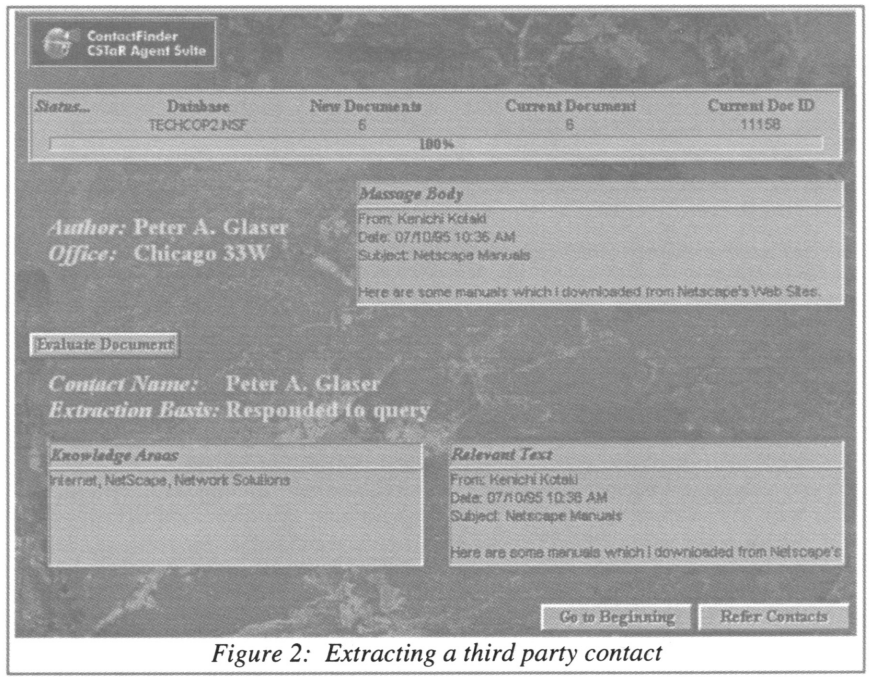


*Figure 2: Extracting a third party contact*

phase one (discussed in detail in section 4). It then proceeds to search its database of key contacts, as shown in Figure 4. It finds the contact which was extracted as a contact previously and sends the referral shown in Figure 5.

A key point of this research initiative is that the documents that provide the original indication of expertise in phase one need not actually address the questions that are handled in the second phase. All that is required is that the topic areas that are extracted match close enough to make it likely that the contact person would be able to help the question asker or at least be able to refer the question asker to a third person. Because of this, our primary focus has been on the extraction of topic areas rather than on the details of the questions being asked or the expertise being provided.

## 4. Extracting topic indicators

The most important step in both phases of the process described above is the extraction of semantically significant phrases. Previous research has attempted to perform document comparison using most or all of the words in a document (e.g., [Sheth, 1994; Hammond *et. al.,* 1995; Pazzani, *et. al*, 1996]), but we are avoiding this approach for two reasons. First, very few of the words in a document reflect the underlying meaning and importance of the text, and moreover the distribution of words does not reflect the words or phrases that best characterize the document. Second, processing the entire text of a document is extremely costly in computational terms, and can be prohibitive for very large sample sets. Extracting semantically significant phrases and processing them is quite tractable.

The critical step for extracting high quality phrases for documents is the set of heuristics for processing blocks of text. This is especially true for highly unstructured documents, which don't have many structured fields or keyword classifications. Even if a set of documents does have categorization keywords associated with each document, it is necessary to augment them with other significant phrases that the authors include in the document text.

To accomplish this we are in the process of integrating and building upon the heuristics found in previous related research [Swaminathan, 1993; Krulwich, 1995] for extracting visually significant features from documents. This approach is built upon the observation that document authors typically use a variety of visual techniques to convey significant pieces of information to readers. Some examples are key points, lists of significant items, document structure, synopses, logical progression, and so on. Recognizing some of these visual patterns allows our agent to extract semantically meaningful phrases from bodies of text.

For example, a simple heuristic is to extract any single word that is fully capitalized. Such a word is most likely an acronym or in some cases a proper technical name. In addition, there are a number of ways to find a definition of an acronym, such as looking for a parenthesized or quoted phrase immediately after the acronym, or at the words before the acronym if the acronym is itself in parentheses, or in the sentence or two preceding the acronym if neither is in parentheses.

Another type of heuristic extracts sequences of multiple word that when taken as a unit can be used as a topic

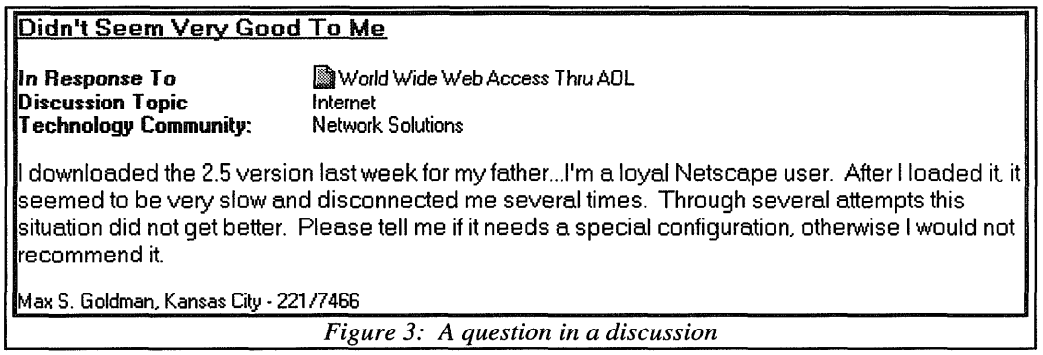


*Figure 3: A question in a discussion*

indicator. For example, a series of capitalized words will most often be a stronger indicator of topic than the constituent words treated independently. Phrase-level heuristics of this sort enable ContactFinder to operate on phrases that more closely resemble human usage, rather than on approximations such as word correlations.

Another simple heuristic is to extract any short phrase, of 1-5 words, which appears in a different format from surrounding text and which is not a complete sentence. This heuristic takes advantage of the convention of italicizing or underlining significant phrases the first time that they're used, or of capitalizing the first letters of proper names.

A further condition for both of these heuristics to be applicable is that the phrase not appear on a fixed list of non-significant words and phrases. For example, the first heuristic should not extract the acronym TM that may follow a product name and the second should not extract words such as "not" or "certainly," which are often italicized for emphasis.

Other heuristics of this sort include recognition of lists of items (with bullet points or numbers), section headings, diagram labels, row and column headers in tables, and heavily repeated phrases. We are in the process of exploring heuristics such as extracting compound noun phrases (made up of three or more nouns in a row), which are frequently domain-specific phrases. Additionally, we are investigating the integration of a thesaurus, either commercial or domain-specific, to allow the agent to recognize that two words or phrases that have been extracted should be treated as equivalent.

A key aspect of these heuristics is that they are completely free of domain and contextual knowledge, and rather focus entirely on the syntactic structure of the text. This allows them to be widely applicable, without relying on background knowledge, and to be computationally efficient. There will be situations, however, in which such knowledge is necessary to perform effectively. Types of knowledge that could be added include topic areas and their relationships. The next section discusses a particular situation in which this is necessary and undoubtedly more such cases will be uncovered as experimentation progresses. For the most part, however, ContactFinder will operate using knowledge-free heuristics of the sort described in this section.

## 5. Experimental results

To date, ContactFinder has operated for several months on an internal bulletin board for discussion of technical issues. Out of 2893 total documents processed from the bulletin board, ContactFinder extracted 762 key contacts on various topics. This reflects our desire that ContactFinder operate relatively conservatively and error on the side of false negative contact extractions (failing to extract contacts) rather than false positives (extracting people as contacts who in fact are not). Many messages that may be indications of expertise, such as those that are top-level (not responses) but are not questions, are skipped by ContactFinder for lack of certainty. Exploration of other heuristics for identifying contacts will improve this rate.

Out of the same total set of messages the system extracted 611 questions for which it found 83 potential referrals. This rate of success (13.6%) reflects a number of aspects of the system's operation. First, the system will never post a referral to someone who has already responded to the question, which will often be the case during early operation when most of the system's set of key contacts have been extracted from the same set of messages. Second, the system requires a fairly strong match between topic areas of the question and the contact (90%) before considering a referral.

Of these 83 referrals, 3 related to a particular technical topic (a system named SAP) that posed difficulties for ContactFinder's approach. SAP is a very large system composed of many sub-systems and for the most part any individual will only work with a small number of these sub-systems. It's necessary, therefore, for ContactFinder to correctly determine the relevant sub-systems for every contact and question. Unfortunately, this has proven difficult for a number of reasons. First, the sub-systems are often named as two-letter acronyms, without the use of punctuation as separators, such as SD, FL, HP, MM, PS, DB, and PP. Many of these two-letter names are also used in English messages for other purposes, such as PP being used for page number references or FL being the postal

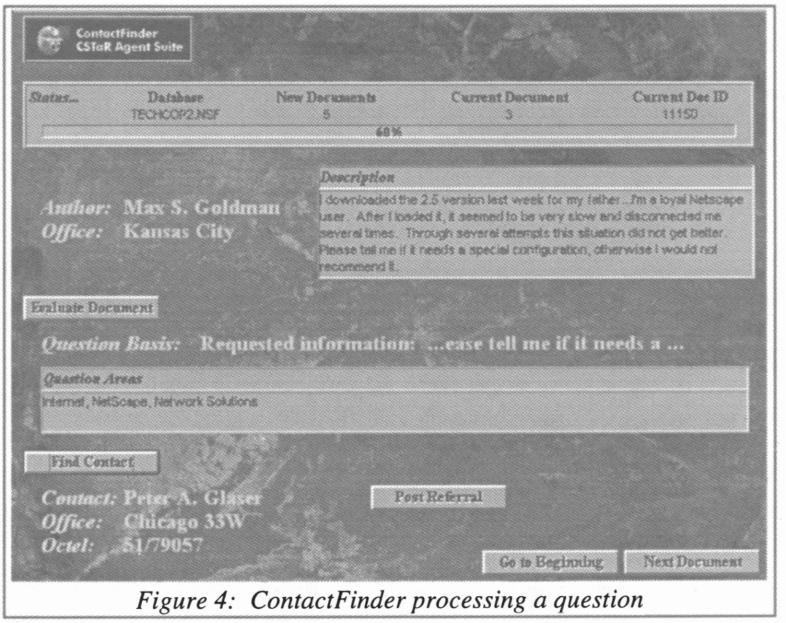


*Figure 4: ContactFinder processing a question*

code for Florida. For this reason it has been impossible for ContactFinder to extract these topic indicators in a knowledge-free and context-free fashion. Second, these sub-systems are sometimes referred to by their expanded names, requiring that ContactFinder know that the two-letter codes are synonyms for their expansions.

In general, this problem is with the knowledge-free nature of ContactFinder's topic extraction heuristics. Were ContactFinder to have specific knowledge of SAP and its sub-systems, it could look for the two-letter names only in the context of SAP and could know the relevant synonyms. While we have in fact included knowledge of some synonyms in ContactFinder, we have not yet explored broader domain knowledge such as system components and sub-systems. Future research will determine the degree to which knowledge of this sort is necessary.

Out of the 80 remaining referrals, 39 of them have been approved by the contacts themselves and 11 of them have been refused, giving us a 78% success rate (after excluding SAP-related referrals). Continuing testing of the system will determine how this rate holds up over larger numbers of documents.

Anecdotal feedback concerning ContactFinder has been very positive. Some bulletin board users have feared that the system will reduce the number of on-line responses and move the flow of knowledge off-line as people call contacts directly instead of waiting for them to respond on-line. In practice, however, it appears that just the opposite is true. In several cases the contacts referred by ContactFinder have posted information on-line having not seen the questions until ContactFinder contacted them. If this trend continues, it appears that ContactFinder will in fact increase the amount of information on-line as people who do not have a chance to read the bulletin board regularly are encouraged to respond to particular messages that relate to their areas of expertise.

## 6. Summary and discussion

We have described an intelligent agent prototype that mines a heterogeneous information repository for key contacts in specific subject areas. This approach allows the agent to assist people seeking information without requiring deep understanding of the information source documents. It also allows the agent to fit well with typical work styles by facilitating transfer of expertise between people. Lastly, the advice and the reasoning behind it is very easily understood by the people involved because the referral can include a reference to the document that provided the contact.

The agent has been designed to operate by responding to questions on discussion groups. This allows it to answer only those questions for which it has referrals and to operate in a background fashion appearing to users as simply another source of messages.

The system currently leaves open a number of issues that will serve as the basis for our continuing research. How can a large variety of types of documents be successfully mined for indications of contacts? How can documents consisting of plain formatted text be processed effectively to extract contacts, questions, and indications of subject area? What types of background will be needed to operate effectively in a variety of domain areas?

More generally, our approach raises the question of what other intelligent agent functionality can be achieved using document processing techniques such as significant phrase extraction, inductive learning, and document search. We are currently developing of several agents based on these techniques, such as an agent that learns the information interests of various users along with how to find new

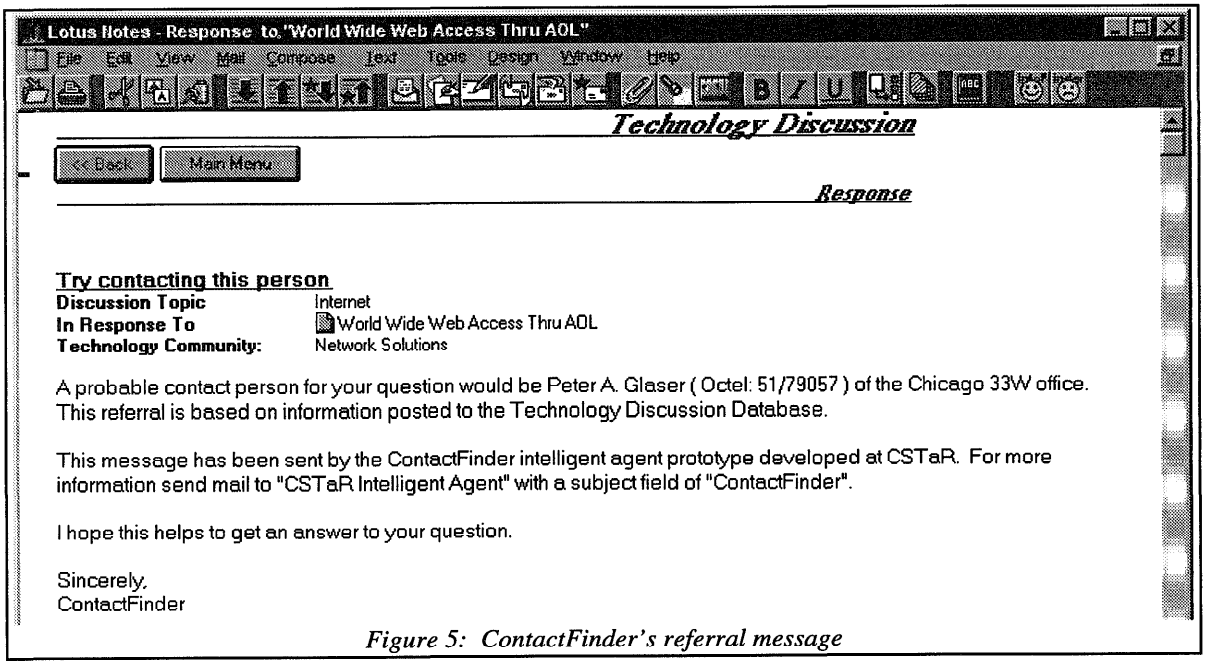


*Figure 5: ContactFinder's referral message*

documents matching those interests [Krulwich and Burkey, 1996], an agent that interacts with on-line Internet services, and an agent that browses on-line documents to extract summary information. We are also investigating the application of other core document processing techniques, such as schema matching and message sequence modeling, to intelligent agent tasks. Future research will determine the range and effectiveness of intelligent agents that can be built on core document processing techniques such as these.

## References


Hammond, K., Burke, R., Martin, C., and Lytenin, S., 1995. FAQ Finder: A case-based approach to knowledge navigation. In *Working Notes of the 1995 AAAI Spring Symposium on Information Gathering in Distributed Environments,* Palo Alto, CA.

Holte, R. and Drummond, C., 1994. A learning apprentice for browsing. In *Working Notes of the 1994 AAAI Spring Symposium on Software Agents,* Stanford, CA, pp. 37-42.

Kantz, H. and Selman, B., 1996. Agent Amplified Communication. In *Proceedings of the 1996 National Conference on Artificial Intelligence,* Portland, OR.

Knoblock, C. and Arens, Y., 1994. An architecture for information retrieval agents. In *Working Notes of the 1994 AAAI Spring Symposium on Software Agents,* Stanford, CA, pp. 49-56.

Krulwich, B., 1995. Learning user interests across heterogeneous document databases. In *Working Notes of the 1995 AAAI Spring Symposium on Information Gathering in Distributed Environments,* Palo Alto, CA.

Krulwich, B. and Burkey, C., 1996. Learning user information interests through the extraction of semantically significant phrases. In *Working Notes of the 1996 AAAI Spring Symposium on Machine Learning in Information Access.*

Lashkari, Y., Metral, M., and Maes, P., 1994. Collaborative interface agents. In *Proceedings of the 1994 AAAI Conference,* Seattle, WA, pp. 444-449.

Levy, A., Sagiv, Y., and Srivastava, D., 1994. Towards efficient information gathering agents. In *Working Notes of the 1994 AAAI Spring Symposium on Software Agents,* Stanford, CA, pp. 64-70.

Maes, P. and Kozierok, R., 1993. Learning interface agents. In *Proceedings of the 1993 AAAI Conference,* Washington, DC, pp. 459-465.

Pazzani, M., Muramatsu, J., and Billsus, D., 1996. Syskill and Webert: Identifying Interesting Web Sites. In *Proceedings of the 1996 National Conference on Artificial Intelligence,* Portland, OR.

Sheth, B., 1994. *A learning approach to personalized information filtering.* M.S. Thesis, EECS Department, MIT.

Swaminathan, K., 1993. *Tau: A domain-independent approach to information extraction from natural language documents.* DARPA workshop on document management, Palo Alto.