

Learning Probabilistic User Profiles

Applications for Finding Interesting Web Sites, Notifying Users of Relevant Changes to Web Pages, and Locating Grant Opportunities

Mark Ackerman, Daniel Billsus, Scott Gaffney, Seth Hettich, Gordon Khoo, Dong Joon Kim, Ray Klefstad, Charles Lowe, Alexius Ludeman, Jack Muramatsu, Kazuo Omori, Michael J. Pazzani, Douglas Semler, Brian Starr, and Paul Yap

■ This article describes three agents that help a user locate useful or interesting information on the World Wide Web. The agents learn a probabilistic profile to find, classify, or rank other sources of information that are likely to interest the user.

The World Wide Web (WWW) contains a vast amount of information potentially relevant to a user's goals or interests. We describe three agents that help a user locate useful or interesting information. The agents share a common core approach to learning a probabilistic profile from a user's feedback. This profile is then used to find, classify, or rank other sources of information that are likely to be of interest to the user. The agents described in this article are *SYSKILL & WEBERT*, *DICA*, and *GRANT LEARNER*.

SYSKILL & WEBERT is designed to help a user with long-term information-seeking goals, such as finding previously unvisited web sites on a particular technical topic. *DICA* is designed to monitor user-specified web pages and notify the user of important changes to these pages. The user provides feedback on which changes are important, and *DICA* learns

to notify the user only for significant changes. *GRANT LEARNER* notifies an individual of new research grant opportunities that meet the learned profile of the individual's research interests.

These systems are available for use at www.ics.uci.edu/~pazzani/Agents.html. Although the representation and learning of user profiles is the same in the three systems, they differ along several dimensions, including what information is processed by the system, how additional information is located by the system, and how recommendations are presented to the user. We begin our discussion with *SYSKILL & WEBERT* because it was the first of the three agents and has undergone the most experimentation and user testing.

SYSKILL & WEBERT

Web search engines are designed to assist a user in finding information relevant to a short-term goal, such as a web page with information on a product one is considering buying. In contrast, *SYSKILL & WEBERT* is designed to assist a user in pursuing long-term goals, such as keeping track of interesting de-

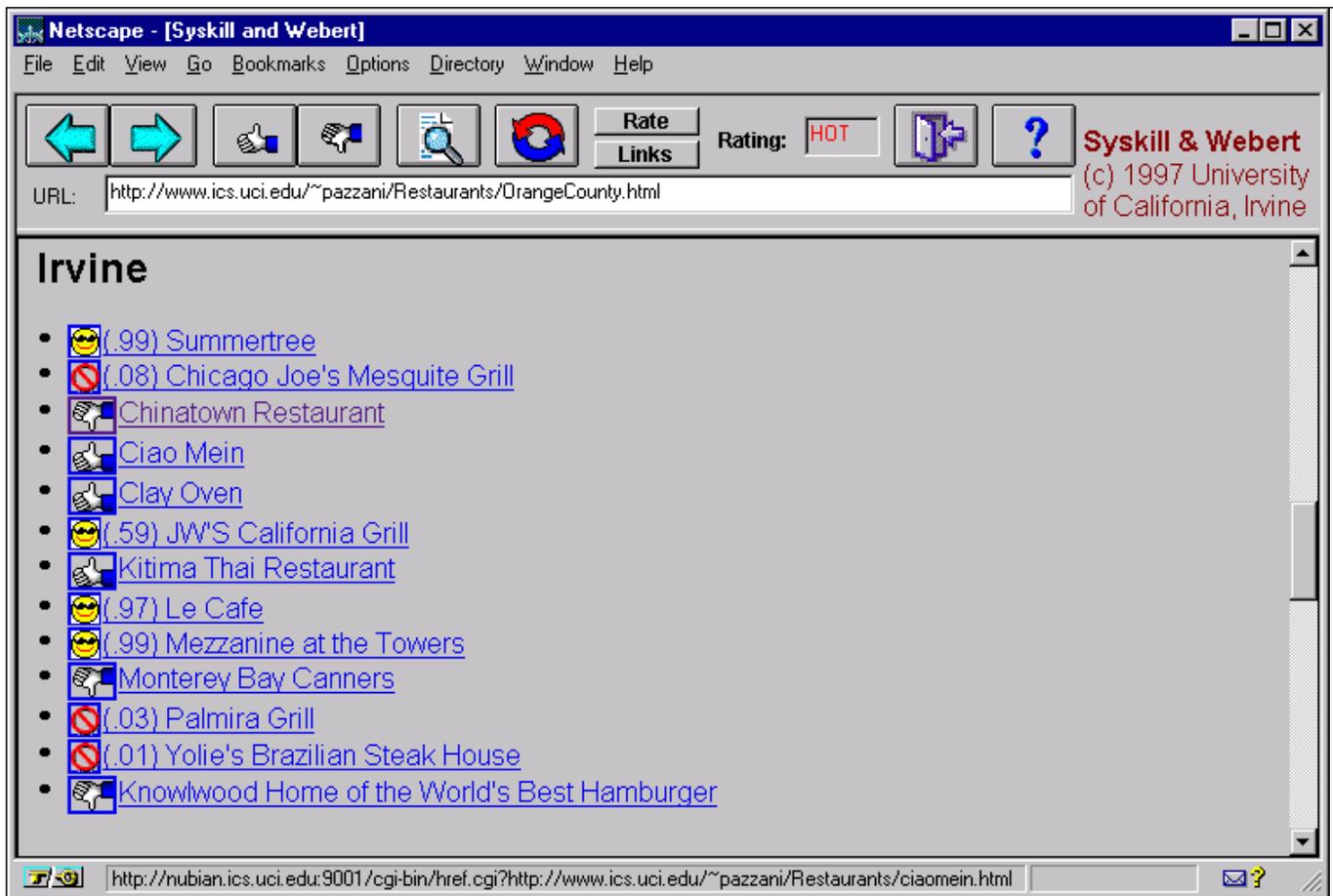


Figure 1. The SYSKILL & WEBERT Interface for Rating Web Pages Adds Additional Commands to the Web Browser.

velopments or information resources in a particular field. The current version of SYSKILL & WEBERT is implemented as a plug-in and collection of JAVA classes that add extra functions to a web browser. A user of SYSKILL & WEBERT first declares a new topic (or reuses an existing topic). Topics of current users include biomedical computing, independent recording artists, and raising goats. The user can optionally associate with each topic one or more *topic pages*, which are typically pages created by a person listing web pages available on the topic. For example, golgli.harvard.edu/biopages/all.html contains links to more than 2000 web pages on biosciences. Although the maintainers of this page have done a remarkable job of collecting information, the usefulness of the resource would be enhanced if there were an assistant watching the user explore this resource, learning his/her interests, and recommending which unvisited links should be explored next. SYSKILL & WEBERT is designed to provide such assistance.

Rather than inferring the user's preferences from behavior (Lieberman 1995), SYSKILL & WEBERT allows the user to provide feedback on the interestingness or usefulness of a web page. The user can click on hot or cold buttons that are added to the browser window (figure 1).¹ SYSKILL & WEBERT records these ratings to learn a profile of the user's likes and dislikes. A simple Bayesian (Duda and Hart 1973) classifier creates a user profile from the user's preferences. The profile can be used to calculate the probability that any web page is interesting to the user. SYSKILL & WEBERT can locate potentially interesting pages in two ways: First, it can locate pages directly accessible from the current page. Second, it can construct a query of a search engine such as LYCOS (Mauldin and Leavitt 1994). Two types of word are used to build a query: First, a query contains words that occur in the most number of pages that have been rated hot (ignoring common English words and all hypertext markup language [HTML] commands). Second, a query contains words whose presence

Syskill & Webert- Lycos Search

**Lycos search: GRANTS CONTROL WUSTL DATA
GENOME CDC INFECTIOUS UNIVERSITY
RESEARCH PHARMACY HEALTH JOURNAL
BIOLOGY MEDICAL**

Lycos Nov 15, 1995 catalog, 11646653 unique URLs

1)  **.008 Research & Data** [1.0000, 5 of 14 terms, adj 1.0]

Abstract: Research & Data HHS makes a substantial investment in improving understanding of **health** and social services. An increasing amount of the information generated by HHS organizations will be made available on the Internet and through this page. Program evaluation and policy **research** * Abstracts of HHS program evaluation studies * The 1994 Green Book: **Data** and information on selected social welfare programs * Office of the Assistant Secretary for Planning and Evaluation **Biomedical research** * National Institutes of **Health** (NIH) * National Library of <http://www.as.dhhs.gov/resdata.html> (3k)

2)  **1.0 Online Journals** [0.9902, 6 of 14 terms]

Abstract: American Chemical Society Publications DOE Whitepaper on Bio-Informatics Emerging Infectious Diseases (EID) - CDC European Molecular Biology Network Newsletter Human Genome Newsletter GCG Transcript Journal of Biological Chemistry Journal of Clinical Monitoring Journal of Computer-Aided Molecular Design Mathematics and Biology Protein Science Recombinant DNA/Protein Resource Facility News Report: NCHGR GESTEC Director's Meeting on Genome Informatics Informatics Center, Vanderbilt University Medical Center Last Modified: Wednesday, 04 October 1995
webmaster@www.mc.vanderbilt.edu
<http://vmcilib.mc.vanderbilt.edu/resources/interests/journal.html> (3k)

3)  **.996 Selected Internet Resources Starting Points Subject Specific**
Site Institution [0.9950, 5 of 14 terms, adj 1.0]

Figure 2. An Example of a Page Annotated by SYSKILL & WEBERT.

Here, SYSKILL & WEBERT constructed a LYCOS query from a user profile for the biomedical topic.

in an HTML file helps discriminate pages that are rated hot from cold using mutual information. Because LYCOS cannot accept long queries, we use the seven most informative words that are found in a higher proportion of hot pages than all pages and the seven most commonly occurring words as a query.

Once a search engine returns a page with links to potentially relevant pages, SYSKILL & WEBERT retrieves each page and analyzes the HTML to produce its recommendations. Although this process is time consuming,² the page ratings of SYSKILL & WEBERT can reduce the time spent by a user manually searching

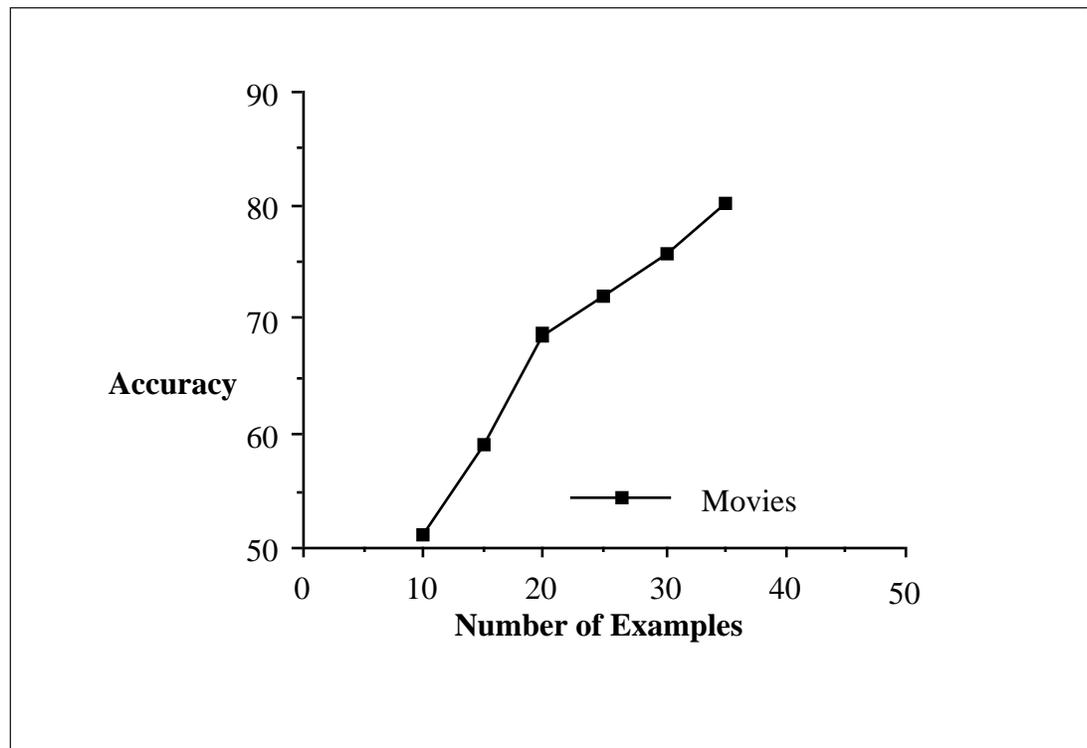


Figure 3. *SYSKILL & WEBERT's Accuracy at Predicting the User's Preference for Movies Increases to More Than 80 Percent.*

the web. In the newest version of *SYSKILL & WEBERT*, the multithreading capabilities of *JAVA* are exploited so that page ratings are displayed interactively while other pages are being retrieved.

SYSKILL & WEBERT displays its recommendations by annotating links that point to web pages that have been rated. Figure 2 shows an example of a web page returned by *LYCOS* using a query constructed by *SYSKILL & WEBERT*. The annotations indicate whether the user is likely to be interested in the page and what the probability is that a page is interesting.

Pazzani, Muramatsu, and Billsus (1996) provide details on how *SYSKILL & WEBERT* learns and uses a profile. Here, we just summarize the most important details. *SYSKILL & WEBERT* represents each page rated by the user as a set of Boolean features. Each feature corresponds to the presence or absence of a particular word. *SYSKILL & WEBERT* selects the 96 most informative (Lewis 1992; Quinlan 1986) words³ as features (that is, words that occur in a greater proportion of hot training documents than cold, or vice versa). Once represented as a set of Boolean features, a variety of machine-learning or information-filtering algorithms might be used to create a user profile.

The publicly available version of *SYSKILL & WEBERT* uses a standard machine-learning algorithm, a Bayesian classifier to create a probabilistic profile. The profile contains the estimated prior probability that a page is hot (and cold) and, for each informative feature, the conditional probability that a page contains (and doesn't contain) a word given that the page is hot (and cold). This algorithm was selected because it was experimentally found to be one of the most accurate at this task, it has desirable time and space complexity (linear in the number of examples and features), and it estimates the probability that a page would be interesting to the user for a particular topic. This probability can be used to rank the unseen pages. The probability can also be used to decide whether to recommend the page to the user. By selecting a threshold for the probability, it is possible to trade precision for recall. We use 0.5 as the default threshold in *SYSKILL & WEBERT*.

We have run several experiments to evaluate how often *SYSKILL & WEBERT's* prediction agrees with the user's rating of a page. Here, we report on one such experiment that was conducted using 48 web pages rated by a user on the topic of films that were available at

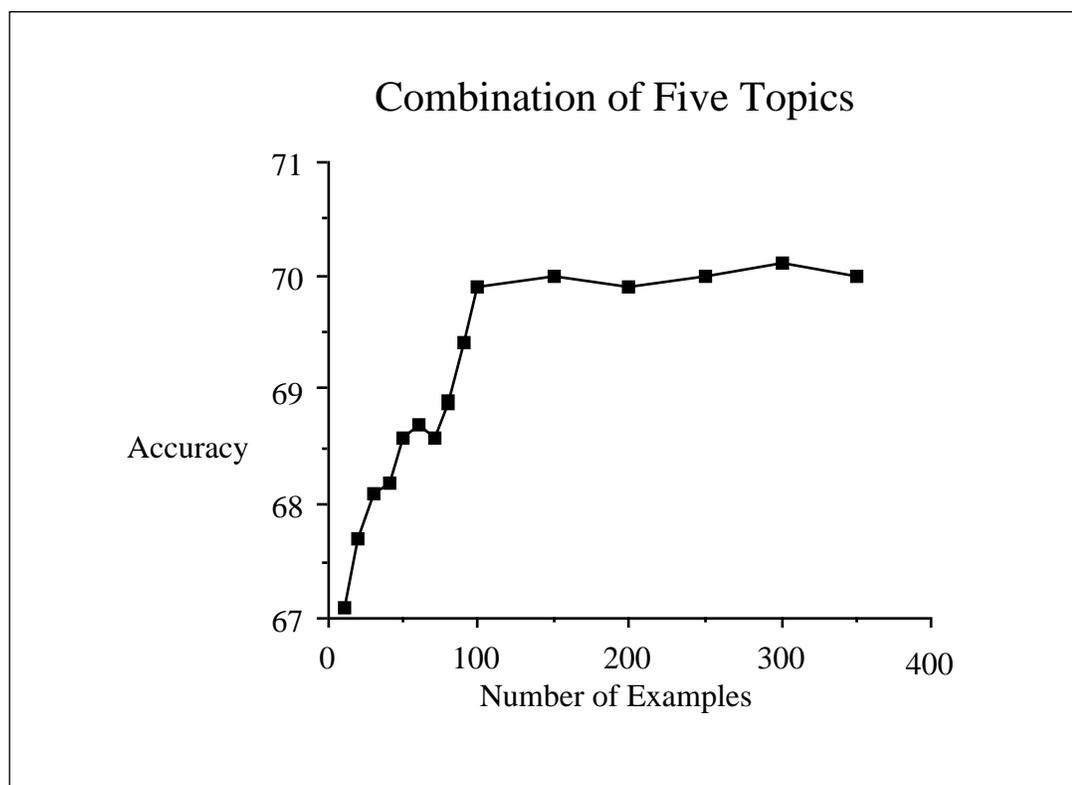


Figure 4. *The Bias of the Bayesian Classifier Is Not Suitable for Representing a User Profile When There Are Several Different Types of Page That Are Liked by the User.*

rte66.com/Movies/blurb.html. In this case, the user rated the pages according to whether the user liked the film that was described by the page. Each page had information on the people involved with the film and a summary of the film. Twenty-five of the 48 pages were rated hot by the user. To test the ability of SYSKILL & WEBERT to learn the user's preferences, it was trained on a subset of the rated pages and evaluated on pages that were not in the training set. Because we are interested in seeing how many pages a user would have to rate to learn a useful profile, we increased the number of training pages from 10 to 35 in increments of 5. Figure 3 shows the average accuracy at predicting unseen pages as a function of the number of training examples. Each point represents the average of 40 trials. On each trial, n pages were randomly selected for training, and a new profile was learned for these pages. The accuracy of the test consisting of all pages not in the training set was computed. The figure shows that with 10 rated pages, SYSKILL & WEBERT performs at near chance levels. However, as the number of rated pages increases, the accuracy of its predictions increases. When trained on 35 pages,

SYSKILL & WEBERT agreed with the user's rating more than 80 percent of the time.

The user profile provides relatively fine-grained probability estimates that can also be used to order the exploration of new pages. For example, on the biomedical domain, we used a leave-one-out testing methodology to predict the probability that a user would be interested in a page. The 10 pages with the highest probability were all correctly classified as interesting, and the 10 pages with the lowest probability were all correctly classified as uninteresting. There were 21 pages whose probability of being interesting was above 0.9, and 19 of these were rated as interesting by the user. Sixty-four pages had a probability of being interesting that was below 0.1, and only 1 was rated as interesting by the user.

The decision to learn a separate profile for each topic in SYSKILL & WEBERT was based on our intuition that restricting the scope would simplify the learner's task. In principle, an intelligent agent, such as a close colleague, should be able to fairly accurately decide whether any web page would interest a user without regard to whether the topic is films

In principle, an intelligent agent, such as a close colleague, should be able to fairly decide whether any web page would interest a user without regard to whether the topic is films or biomedical research.

or biomedical research. The experiment we report next investigated whether SYSKILL & WEBERT would be able to accurately learn a user's preferences when data from several topics are combined. In particular, we pooled pages from 5 topics rated by a single user and created a collection of more than 450 documents. The user rated 69.5 percent of these documents as cold. The results of an experiment on these data, run in the same manner as the previous experiment, are reported in figure 4. Here, we see that the Bayesian classifier does not reach an accuracy level that differs substantially from simply guessing that the user would not like any page.

The problem arises for two reasons: First, words that are informative in one domain (for example, recording artists) are irrelevant in another (for example, biomedical applications). Furthermore, within a particular single domain, it is possible that a simple decision boundary exists that separates the interesting pages from the others. However, it is likely that in several domains, there are a variety of complex clusters of interesting articles, and the Bayesian classifier is unable to represent this decision boundary. One might suspect that a more complex learning algorithm might perform better at this more general problem. However, neither decision tree learners nor multilayer neural nets were more accurate than the Bayesian classifier because of the relatively small number of examples available for the complex high-dimensional concept needed to make accurate predictions across domains. Therefore, we conclude that asking the user to provide a topic and learning a separate profile for each topic is an essential, and not particularly burdensome, requirement for an agent such as SYSKILL & WEBERT.

In summary, SYSKILL & WEBERT collects a user's ratings of web pages, learns a profile of the user's preferences for pages of a given topic, retrieves and rates unseen web pages, and annotates links pointed to rated pages with a rating of the page.

DICA

The DICA (do-I-care agent) (Starr, Ackerman, and Pazzani 1996) Project explores a naturally reoccurring problem with the web. Even after one has discovered web resources, a problem remains: How do you know when to revisit these resources for new material? Consider some common postdiscovery situations:

First, you know the page where a colleague has his publications. You would like to obtain

new papers as they are posted. Furthermore, you would like to know only about papers on a particular topic rather than all papers.

Second, you have obtained the page for a low-cost clothes merchandiser. You would like to know when there is a sale on jeans or Hawaiian shirts.

These are examples of change events that occur somewhat unpredictably and offer no explicit notification. Today, interested parties must occasionally check by hand. This process can be burdensome and tedious, and one might forget to do it on a timely basis. Furthermore, we would like an agent to notify us only of changes that are significant. DICA creates and maintains a web page to report its findings to a user. DICA agents of other users can monitor such a web page to be notified when another user's agent has found an interesting change to the web.

DICA explores the nature of collaborative resource discovery and information sharing on the web. Its goal is to augment certain types of computer-supported cooperative work situations. Collaborative efforts to discover internet resources, such as interesting web pages, require work by users. Unfortunately, users often will not do the extra, altruistic work, leading to adoption failures for social-filtering systems. DICA assumes that agents are especially useful in situations where the benefits would be marginal if the task had to be done by hand and that cooperating agents are especially useful in situations where individuals want to use the agents for their own purposes anyway.

To track interesting changes, an agent needs to know what changes are interesting and where to find them. DICA collaborates with both users and their peers to identify potentially interesting changes. The agent works by soliciting a user's opinions about changes it finds to train the user model; it is in the users' best interests to keep their personal agents informed about relevant pages and the quality of reported changes. Without extra work, the agent can then share these opinions and findings with agents of like-minded users. Individuals engage in resource discovery for themselves, but they can share this effort without any extra work.

Because different changes can be interesting for different reasons, users can maintain multiple DICA agents specialized on particular topics or kinds of web page. Each DICA agent must periodically visit a user-defined list of target pages, identify any changes since the last time the agent visited, decide whether the changes were interesting, notify the user if the change

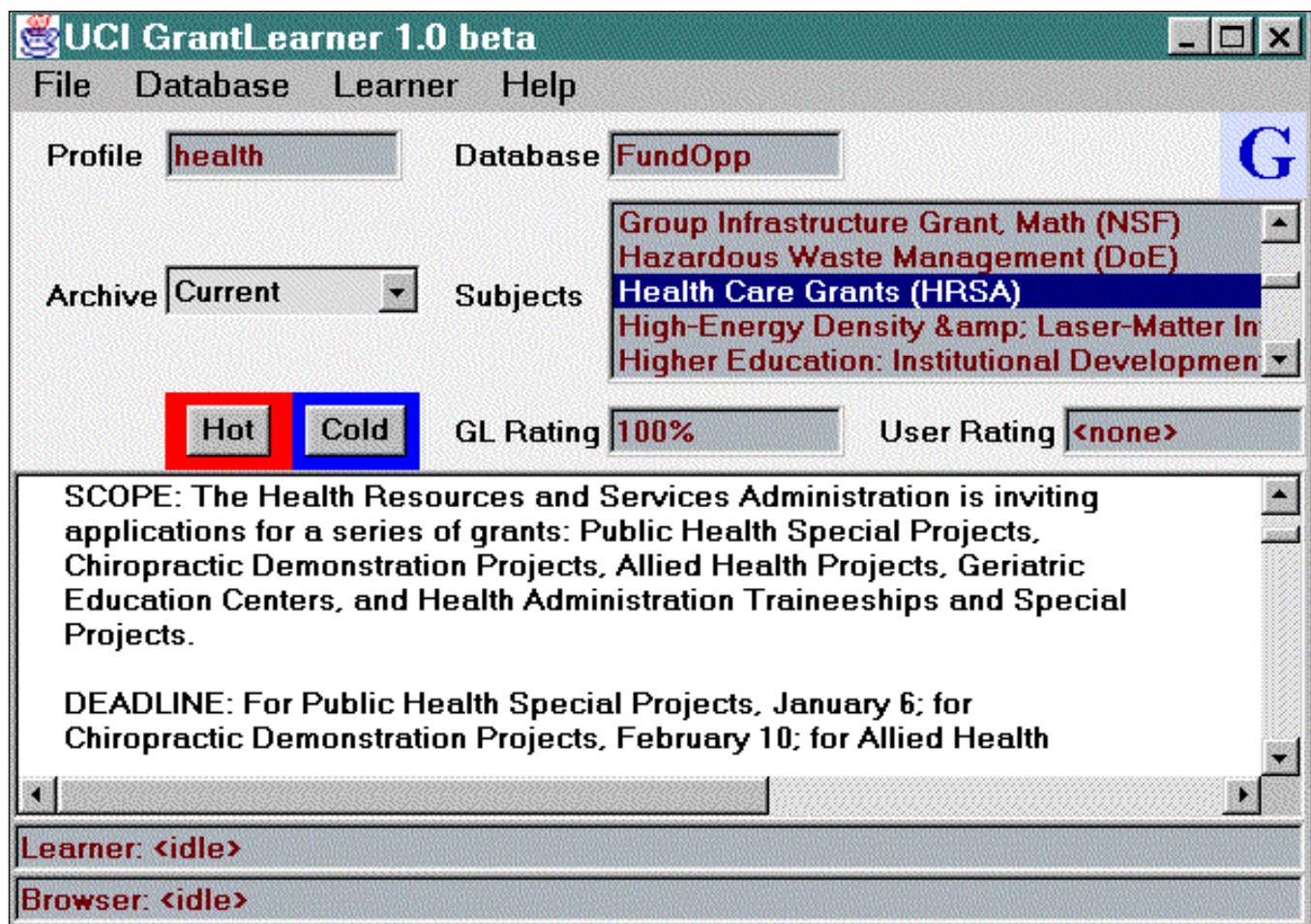


Figure 5. The University of California at Irvine GRANT LEARNER Interface for Finding Grant Opportunities for Funded Research.

was interesting, accept relevance feedback (Rocchio 1971) on the interestingness of the change and the timeliness of the notification, and facilitate information sharing and collaboration between individuals and groups.

DICA maintains a list of web pages for each user's topic that have interesting information. By limiting the search to a small subset of web pages specifically selected because they are likely to contain interesting information, we can greatly improve precision. By accepting or rejecting DICA's suggestions, the user refines the types of change that are interesting. Like SYSKILL & WEBERT, a Bayesian classifier learns a profile. The primary difference to the learning algorithm between SYSKILL & WEBERT and DICA is that SYSKILL & WEBERT extracts features from a web page to determine the interestingness of the page, but DICA extracts features from the difference between the current and an earlier version of a web page (computed using the UNIX diff command).

Once an agent spots an interesting change, the user is notified by e-mail, and the change is appended to the agent's associated web page. This web page is also used for relevance feedback and collaborative activity. DICA addresses two post-discovery problems: (1) when a user should revisit a known site for new information and (2) how a user shares new information with others who might be interested. These post-discovery chores are laborious, suffer from inequitable distribution of costs and benefits, and are not both addressed by existing systems. By providing a simple technical mechanism, we have found a collaboration mechanism that provides for group effort through individual incentives.

GRANT LEARNER

The profiles learned by the Bayesian classifier are in no way limited to web pages and can

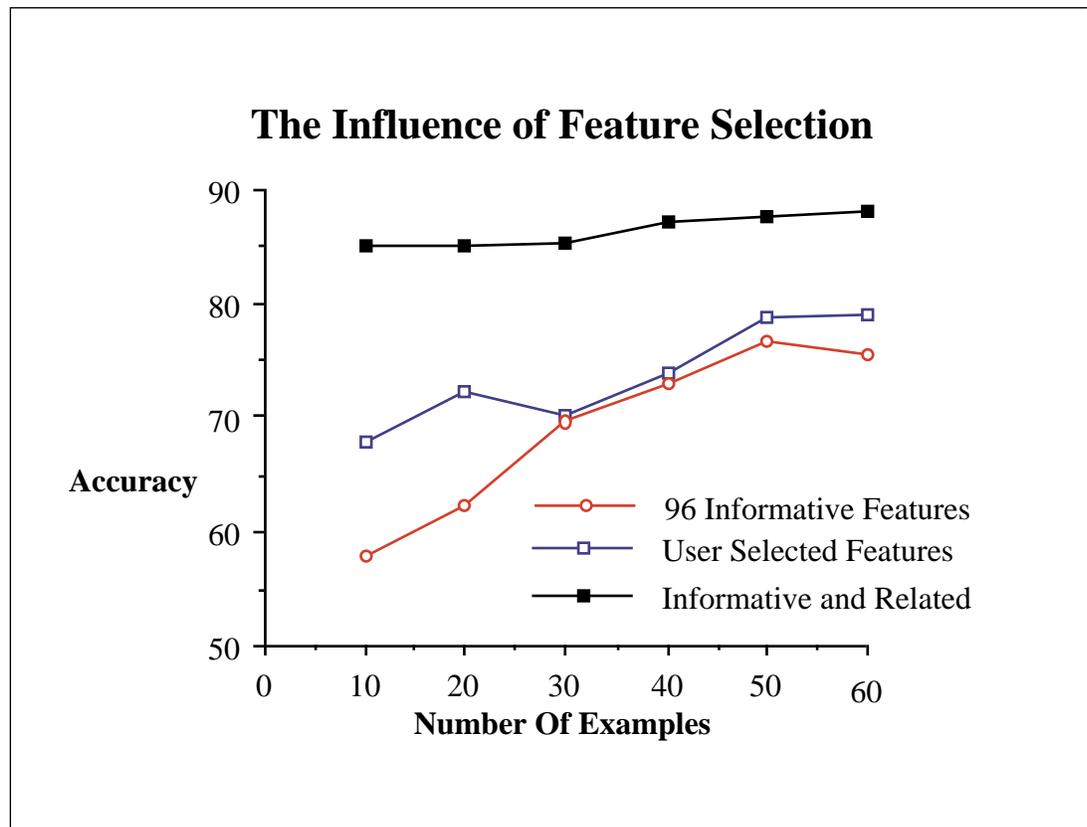


Figure 6. Improving on the Bayesian Classifier by Automatically or Manually Selecting Relevant Features.

be applied to all kinds of text document available on the internet. In a different project, we are using the described learning techniques to classify descriptions of funding opportunities. The goal of the University of California at Irvine GRANT LEARNER Project is to provide a system that learns to distinguish between interesting and uninteresting funding opportunities based on a user's ratings of these descriptions.

Our current implementation of GRANT LEARNER is a JAVA stand-alone application that runs on any JAVA-compatible platform. It connects to a regularly updated grant database maintained at UCI and provides a user interface to browse this database. Figure 5 shows a screen shot of the current version of GRANT LEARNER. The system displays a list of all grant subjects available in the database. By clicking on a subject line, the user can retrieve a corresponding full-text description. Grant descriptions usually provide information about the grant's subject area, funds, eligibility, and deadlines. Using the hot and cold buttons, the user can indicate whether a grant description is related to his/her interests. Once the user has rated a

minimum number of grants (10 in the current implementation), GRANT LEARNER uses the previously described feature-extraction and learning algorithms to learn a profile of the user's interests. The learned profile is then used to process the rest of the database, that is, all the grant descriptions remaining to be looked at by the user. The result of this process is a list of grant descriptions sorted according to GRANT LEARNER's relevance prediction. A learned user profile can be saved and reused at a later time. For example, users can train GRANT LEARNER with a set of rated descriptions once and reuse the learned classifier at a later time to get a ranked list of database entries that were added recently. First experiments and feedback from different users suggest that GRANT LEARNER is a helpful tool to locate potentially interesting grant descriptions.

Future versions of GRANT LEARNER will allow the use of the same learned profile to filter different grant databases (for example, IRIS and the National Science Foundation grant database). We are also planning to implement an e-mail notification system, such that users are automatically notified when a grant rele-

| | | | | |
|--------------------|-------------------|------------------|-----------------|-------------------|
| pygmy | return | production | cashmere | management |
| milk | animal | angora | feed | spring |
| library | breeding | feeding | cheese | other |
| program | computer | fair | fiber | green |
| however | health | dairy | time | summer |
| took | quality | early | normal | farm |

Table 1. *The Stricken Words Are Excluded as Features by Using WORDNET.*

vant to their interests is added to a database. In addition, we believe that exploiting the structure of descriptions, for example, deadline or eligibility sections, is useful to construct a set of special-purpose features that help prevent the system from suggesting grants for which one is ineligible or whose deadlines have passed. More information about the UCI GRANT LEARNER can be found at www.ics.uci.edu/~dbillsus/grantlearner.html.

Current Research Work

Our current research (for example, Billsus and Pazzani [1996]) involves exploiting different forms of knowledge or information to improve the accuracy of an agent's recommendations. The goal is to reduce the overall effort required by a user to get useful results from the agent. Our efforts have focused on finding additional information to determine which words should be features in the profile. Although implemented in an experimental version of SYSKILL & WEBERT, they are of more general applicability. We have explored two ways of addressing this problem:

First is asking the user for words to use as features for each topic. SYSKILL & WEBERT can take advantage of the user's knowledge of the domain to identify key terms.

Second is using additional lexical knowledge to filter the informative features. We have experimented with using WORDNET (Miller 1994), a lexical database containing the relationships between approximately 30,000 commonly occurring English words. When there is no relationship between a word and the topic (for example, goat), SYSKILL & WEBERT can eliminate the word from consideration as a feature.

Figure 6 shows an experiment (run in the same manner as the previous experiments)

on the goat topic. The experiment uses profiles learned by a Bayesian classifier in which we compare the accuracy of representing documents by the 96 most informative words, the words that a user indicates are most useful in determining whether a page is interesting, and the subset of the 96 most informative features that are related to the word *goat* according to WORDNET (by following a series of hypernym, antonym, member-holonym, or part-holonym links between the word and the word *goat*). The results show that both approaches to selecting features improve on the use of just the statistical information available in the training data. On this problem, the user-selected features only provide a benefit when there are few training examples. Using the additional lexical knowledge encoded in WORDNET to only retain the subset of the 96 informative features that are related to the topic provides a substantial improvement in accuracy on this problem.

Table 1 lists some informative features found from one training set and strikes out those for which no relationship exists between the word and the topic. Although there is room for improvement, WORDNET does remove many words that are unrelated to the topic (and a few, such as *farm*, that are related). Most of the words that are not eliminated are related to the topic (with a few exceptions, for example, *computer*).

Our future plans also include a significant enhancement to the SYSKILL & WEBERT interface that allows users to share their profiles with other users. In many fields, communities of interest already exist; that is, people know who else is working in their respective fields or who is to be considered an expert in a certain area. We envision a useful information-filtering framework that allows people to publish their specialized filtering profiles electronically.

These profiles could then be used by other people interested in someone's area of expertise and be utilized as intelligent filtering tools to view large information spaces, such as the web, "through the eyes of an expert." We also believe that a similar approach might be useful in educational settings. Instructors train a profile based on their judgment about the relevance of items to a class they are teaching. These profiles could then be given to students, who obtain an "automated information guide" for their information needs with respect to the class they are taking.

Summary

We described three examples of intelligent agents under development at UCI. All make use of a simple Bayesian classifier for learning user-probabilistic profiles. Experience with the agents has shown that this classifier provides a useful foundation for creating agents that learn from user feedback and recommends information to the user.

Notes

1. The terminology *hot* and *cold* comes from an early version of *SYSKILL & WEBERT* that maintained a *hot list* for the user's favorite pages and a *cold list* for pages the user did not find interesting.
2. If there were a closer integration between the search engine and *SYSKILL & WEBERT*, the considerable overhead of retrieving a page to rank it would effectively be eliminated.
3. In our experiments, having fewer than 50 features or more than 250 reduces the accuracy of predications, but there was little difference among intermediate values.

References

- Billsus, D., and Pazzani, M. 1996. Revising User Profiles: The Search for Interesting Web Sites. Presented at the Third International Workshop on Multistrategy Learning (MSL '96), 23-25 May, Harpers Ferry, West Virginia.
- Duda, R., and Hart, P. 1973. *Pattern Classification and Scene Analysis*. New York: Wiley.
- Lewis, D. 1992. Representation and Learning in Information Retrieval. Ph.D. thesis, Computer Science Department, University of Massachusetts.
- Lieberman, H. 1995. LETIZIA: An Agent That Assists Web Browsing. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, 924-929. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.
- Mauldin, M., and Leavitt, J. 1994. Web Agent-Related Research at the Center for Machine Translation. In Proceedings of the ACM Special Interest Group on Networked Information Discovery and Retrieval. New York: Association of Computing Machinery.

- Miller, G. 1991. WORDNET: A Lexical Database for English. *Communications of the ACM* 38(11): 39-41.
- Pazzani, M.; Muramatsu J.; and Billsus, D. 1996. *SYSKILL & WEBERT: Identifying Interesting Web Sites*. In Proceedings of the Thirteenth National Conference on Artificial Intelligence, 54-61. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Quinlan, J. R. 1986. Induction of Decision Trees. *Machine Learning* 1:81-106.
- Rocchio, J. 1971. Relevance Feedback Information Retrieval. In *The SMART Retrieval System—Experiments in Automated Document Processing*, ed. Gerald Salton, 313-323. Englewood Cliffs, N.J.: Prentice-Hall.
- Starr, B.; Ackerman, M.; and Pazzani, M. 1996. Do-I-Care: A Collaborative Web Agent. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI'96), 273-274. New York: Association of Computing Machinery.



Scott Gaffney, Seth Hettich, Gordon Khoo, Dong Joon Kim, Alexius Ludeman, Kazuo Omori, Douglas Semler, and Paul Yap are undergraduate students in information and computer science who worked on *SYSKILL & WEBERT* in the summer of 1996 as part of a National Science Foundation Research Experiences for Undergraduates grant.

Daniel Billsus, Jack Muramatsu, and Brian Starr are Ph.D. students in information and computer science conducting research on various aspects of intelligent agents.

Mark Ackerman received his Ph.D. in information technology in 1993 from the Massachusetts Institute of Technology. He is an assistant professor in the Computing, Organizations, Policy, and Society Group at the University of California at Irvine.

Charles Lowe is an administrative analyst in research and graduate studies at the University of California at Irvine.

Michael Pazzani received his Ph.D. in 1987 from the Computer Science Department at the University of California at Los Angeles. He is a professor and chair of the Information and Computer Science Department at the University of California at Irvine. His e-mail address is pazzani@ics.uci.edu.