AI Magazine Volume 18 Number 2 (1997) (© AAAI)

LIFESTYLE FINDER

Intelligent User Profiling Using Large-Scale Demographic Data[1]

Bruce Krulwich

■ A number of approaches have been advanced for taking data about a user's likes and dislikes and generating a general profile of the user. These profiles can be used to retrieve documents matching user interests; recommend music, movies, or other similar products; or carry out other tasks in a specialized fashion. This article presents a fundamentally new method for generating user profiles that takes advantage of a large-scale database of demographic data. These data are used to generalize user-specified data along the patterns common across the population, including areas not represented in the user's original data. I describe the method in detail and present its implementation in the LIFESTYLE FINDER agent, an internet-based experiment testing our approach on more than 20,000 users worldwide.

Generation of user profiles from samples of user interests and characteristics has become a common task for AI research. The input data most often take the form of samples of the user's interests or preferences in a given area, and the profile is a generalization of these data that can be used generatively to carry out tasks on behalf of the user. A common application takes sample documents that a user finds interesting (or uninteresting) and generates profiles of the user's interests. These profiles are then used to find or recognize other documents that are likely to be of interest. Other common applications process input data such as movies or music albums, that the user likes and dislikes and use the resulting profiles to suggest new movies or albums to the user.

Generalizing User Profiles

Most previous methods approach this problem by reasoning from scratch about the generalization of the user data. These systems, particularly those that process documents, operate using the following two-step process (Krulwich and Burkey 1996; Pazzani, Muramatsu, and Billsus 1996; Krulwich 1995; Sheth 1994). First, they extract features from each item of user data, such as common or significant words or phrases in documents. Second, they use machine-learning techniques to develop profiles of the users. Although this approach has the attraction of generality and precision, and the resulting profiles can be used to assist the user in a variety of ways, it has a number of limitations: First, the profiles will only encompass direct generalizations of the input features. Second, users are typically required to specify a large number of samples. Third, the systems are forced to reason from first principles about each user rather than leverage commonalities between users.

A second approach to this task, commonly called *collaborative filtering*, takes a different approach (Lashkari, Metral, and Maes 1994). A user's profile consists simply of the data that the user has specified. These data are compared to those of other users to find overlaps in interests between users, and each user is recommended new items from the data of other users with overlapping interests. This approach requires less computation than the previous one because it doesn't have to reason about the user data, and it clearly leverages the commonalities between users. However, it has the drawbacks of requiring data from a large number of users before being effective, requiring a large amount of data from each user, and limiting its recommendations to the exact items specified by the population of users.

Each of these methods has different

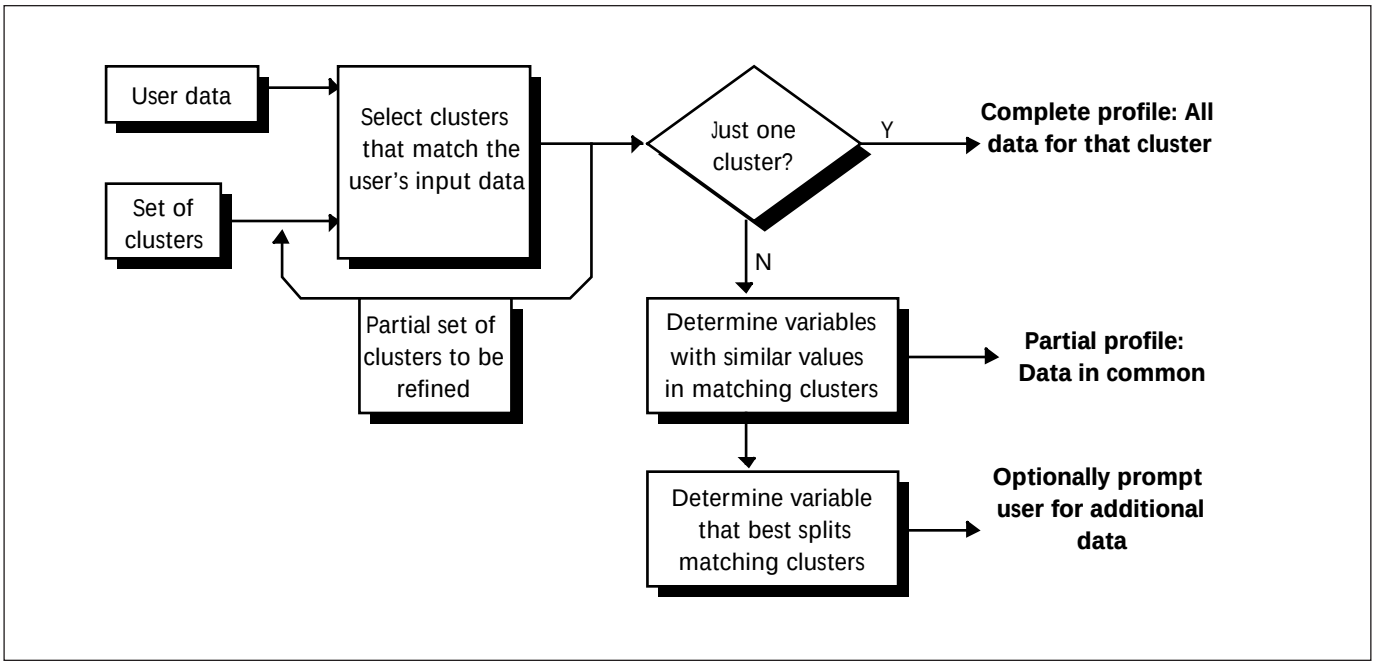


Figure 1. The Basic Process of Demographic Generalization.

benefits and drawbacks and is therefore appropriate for use in different applications or domain areas. This article presents a new method for user profiling that combines the benefits of the two previous approaches. I demonstrate that this method is highly efficient, profiles users using a small amount of information, and results in profiles with a wide scope. Although these benefits come at the expense of a slight decrease in accuracy, I show that the approach is accurate enough to be effective.

The Demographic Generalization Method

My colleagues and I have developed a novel approach to the task of user profiling called *demographic generalization*. Our method uses a commercially available database of demographic data that encompasses the interests of people nationwide.[2] Input data are used to classify users in terms of these demographic data, and these classifications are used as general characterizations of the users and their interests. The resulting profiles span the range of information contained in the demographic database.

We have been using a demographic system called PRIZM from Claritas Corporation. The PRIZM system divides the population of the United States into 62 *demographic clusters* (hereafter clusters) according to their purchasing history, lifestyle characteristics, and survey responses (Weiss 1988). The system is based on surveys of more than 40,000 people as well as U.S. census data, magazine subscriptions, catalog purchases, and the like. The demographic database contains information on more than 600 variables, each of which refers to a specific lifestyle characteristic, purchase, or activity. The variables include such items as owning a dog, purchasing Canadian whisky on a monthly basis, watching the Home Box Office (HBO) cable television station, playing or watching golf, and owning a motorcycle. Each demographic cluster has an associated mean and standard deviation for each variable, indicating the likelihood of people in the cluster to have this characteristic. The PRIZM system is one of the most commonly used demographic systems for consumer marketing, and the broad scope of its information makes it useful for a wide variety of tasks. The data are typically indexed by zip code, age, and gender for use in consumer marketing. However, these indexes are typically not available in the context of online user profiling.

The demographic generalization approach to user profiling is illustrated in figure 1. First, given a set of input data, compute the set of demographic clusters to which the user is most likely to belong. If only one cluster matches, all the data available for the cluster are used as a broad profile of the user, and the process ends. If more than one cluster match-

es the user data, the demographic variables whose values are similar in all the matching clusters form a partial profile of the user. In this way, the method can always provide a profile that is as broad as is supported by the input data to this point. The demographic variable that best differentiates the matching clusters can then be used to prompt the user for further information, and the set of matching clusters can be fed back into subsequent iterations of the algorithm to be refined. In this way, the method can converge on a single matching cluster with a close to minimal number of interactions. Alternatively, the method can use whatever data are subsequently provided to refine the profile or can halt with only the partial profile.

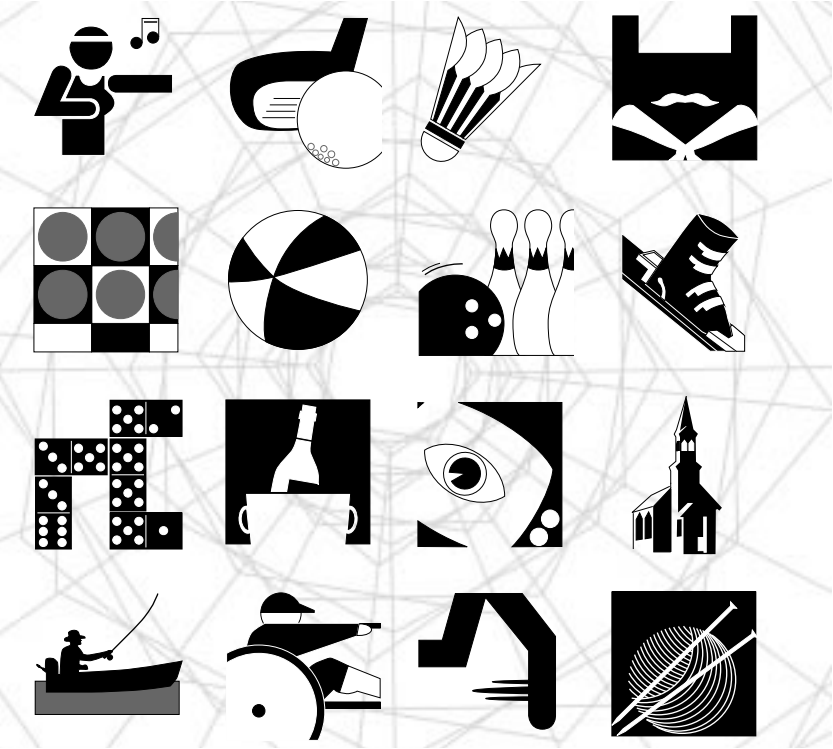
The key step in the process is the selection of the demographic clusters that match the user data. This selection is accomplished by treating each user data item as a constraint on the values of a set of demographic variables. Given the set of constraints imposed by the input data, the system computes the probability that a person in each demographic cluster would fit the user constraints. Thresholding is then used to select a list of demographic clusters most likely to describe the user. If one cluster is substantially more likely than the others, its demographic variable values are used as a profile of the user. If no single cluster is more probable than the others, the set of likely clusters is used to construct a partial profile of the user, as discussed previously.

Several characteristics of the demographic generalization method are noteworthy: First, it is designed to operate incrementally and interactively. A partial user profile is always available, the user can be prompted for new information that will be valuable in profiling, and incoming data items are used for iterative refinement. Second, it is designed to operate with a minimal amount of information from the user. The approach can typically profile users with the values for only a half-dozen demographic variables, significantly less than the amount of information necessary for other approaches. Third, and most important, the method can often profile the user in areas not addressed by the input data as long as these new areas correlate with values in the input data. Even in partial profiles, where the method cannot infer all aspects of the user's profile, it can often infer the user's interest in areas that are very different from those covered by the input data but that are nonetheless consistent with what is known.

As an example, consider the case of a user with the following input data: (1) plays tennis, (2) watches HBO cable television station, and (3) reads the business section of the newspaper.

Eight demographic clusters do these three things to a higher-than-average degree and are thus the most likely clusters to contain the user. These clusters contain individuals that are similar in a number of ways but are also dissimilar in many ways. For example, all the clusters consist of people who earn a higher-than-average income, but they vary in the type of neighborhood in which they live. Five of the clusters refer to suburb dwellers, two to urban residents, and one to those in small cities or towns. When the algorithm described earlier is run on this example, it develops a partial profile of the user indicating such things as higher-than-average income and vacations at the beach. The algorithm then determines that the demographic variable that best distinguishes between the possible matching clusters is the type of neighborhood in which the user lives.

This method, as described to this point, assumes that each user data item can be mapped to a set of required values, or constraints, for demographic variables. Computation then proceeds based on the set of demographic variable constraints. If the user data items are more abstract than the demographic variables, the set of mapping rules can be fairly complex, and the resulting computation can



11%	*Good Morning America*	3–5 times/week	
18%	*ABC World News Tonight*	3–5 times/week	Tends to like news shows
10%	*CBS Evening News*	3–5 times/week	
11%	*NBC Nightly News*	3–5 times/week	
5%	*All My Children*	3–5 times/week	
1%	*Another World*	3–5 times/week	
3%	*As the World Turns*	3–5 times/week	Tends not to like soap operas
4%	*General Hospital*	3–5 times/week	
2%	*Guiding Light*	3–5 times/week	
18%	*Coach*	2–4 times/month	
25%	*Murphy Brown*	2–4 times/month	Tends to like situation comedies
23%	*Roseanne*	2–4 times/month	

Figure 2. Demographic Data for a Particular Population Cluster.

involve a large number of constraints. An alternative approach in these cases is to compile the demographic data into a more abstract form that corresponds directly to the descriptive vocabulary of the input data. This approach enables the input data to map directly to constraints and the computation to take place on a more abstract level with a smaller number of constraints.

Suppose, for example, that the input data specify the user's interest in various genres of television shows, such as soap operas, rather than in specific programs. Rather than treat every user who likes soap operas as having an interest in all the particular soaps in the demographic database and finding the demographic clusters that tend to watch these shows, we can instead compile the demographic data into a form that directly addresses the more abstract features, such as television show genres. Then, given the information that a user likes soap operas, we find those demographic clusters that tend to like soap operas without reasoning directly in terms of particular shows. Thus, the reasoning takes place at the same level of abstraction as the input data.

Consider the demographic data shown in figure 2.[3] These data describe a particular cluster of the population in terms of their interest in the listed television shows. If we are told that a user likes *ABC World News Tonight,* we can infer that the user more likely belongs to this cluster than to clusters with a lower percentage of interest in this show. We can, however, abstract these data and record that people in this cluster tend to like news shows and situation comedies but not soap operas. Then, if we're told that the user likes soap operas in general, we can infer that the user more likely does not belong to this cluster, without having to reason about the particular shows involved.

In this example, the user could be characterized as watching movies on television (abstracted from watching HBO), playing sports (abstracted from tennis), and reading the business section of the newspaper. Abstractions of this type have been used in the LIFESTYLE FINDER experiment.

The LIFESTYLE FINDER Intelligent Agent: Waldo the Web Wizard

My colleagues and I have implemented our method as an intelligent agent that interacts with users on the internet's World Wide Web and uses their profiles to recommend web pages or sites. To make the experiment appealing for web users, we have embodied our agent as Waldo the Web Wizard, a fortune teller who asks questions using cartoon pictures and amusing prose.[4] The engaging nature of the application has enabled us to gather experimental data from over 20,000 users without resorting to test groups operating in unnatural circumstances.

To facilitate simple interactions with users and enable an efficient implementation, we compiled the demographic data into a core set of questions and answers. One such question is, "what leisure activities do you enjoy," and one possible answer is "playing sports." This

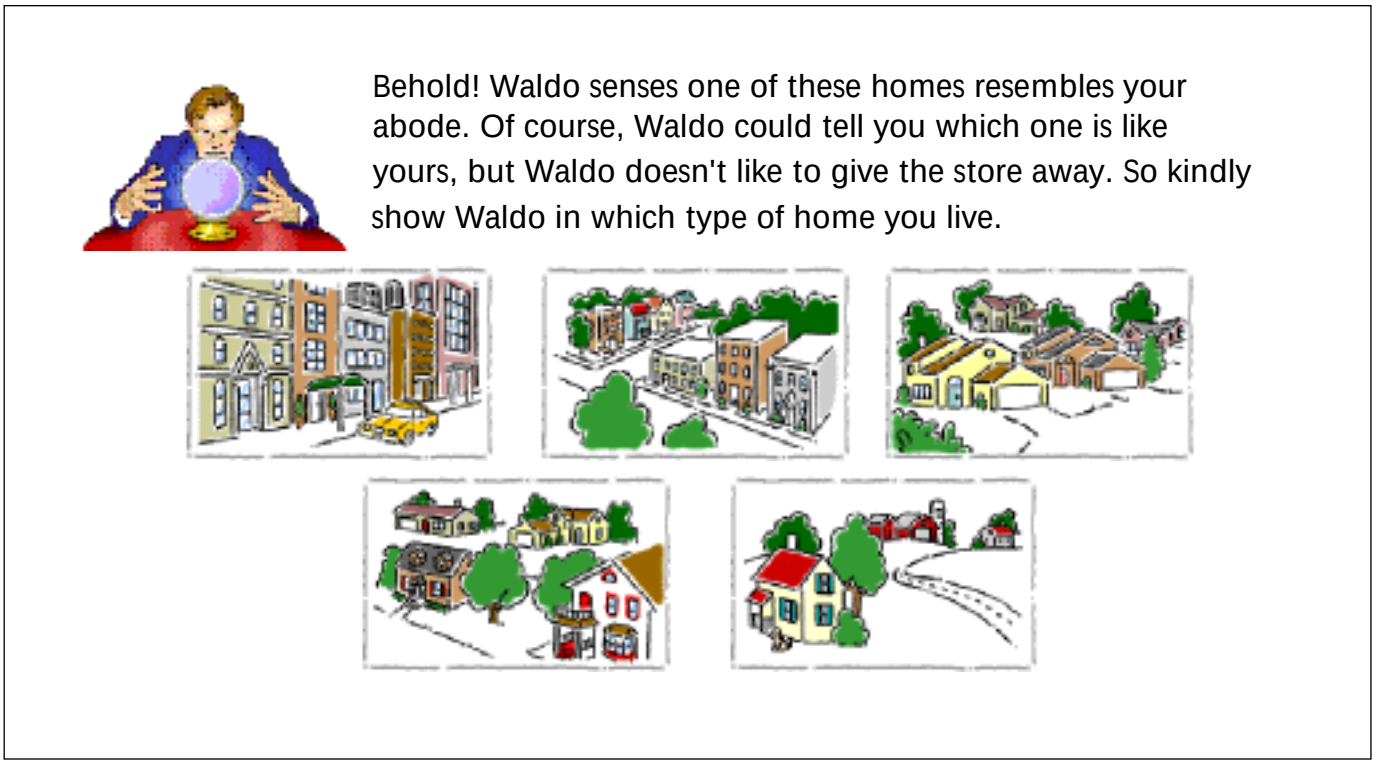


Figure 3. LIFESTYLE FINDER *Asking the User a Question.*

question-answer pair corresponds to an abstraction of the demographic data, as discussed previously, to the level of general activities rather than particular sports. We similarly abstracted other data areas to develop eight questions, with five or six possible answers each. Each question is asked of the user using graphic buttons for response, as shown in figure 3.

Consider the user in the earlier example who plays tennis, watches HBO, and reads the newspaper's business section. If this first question asked of the user is the type of leisure activity that the user most enjoys, the user would select sports, reflecting an interest in playing tennis. The next question asked by LIFESTYLE FINDER, designed to optimally reduce the number of matching clusters, is the newspaper section or magazine type that the user most often reads. The user in our example would select the business section. The next question asked is what the user likes to watch on television, to which the user responds movies. Last, LIFESTYLE FINDER asks about the user's neighborhood. If the user answers with small cities, LIFESTYLE FINDER concludes that the user is most likely a member of a particular demographic cluster, nicknamed Upward Bound, consisting of young white-collar families. Individuals in this cluster tend to use financial planning services, listen to progressive rock music, and read science and technology magazines.

Figure 4 shows the top portion of LIFESTYLE FINDER's output for this user. The top contains a picture and a short lifestyle description for people in the demographic cluster to which the user is believed to belong. The user has the chance to rate the accuracy of this "fortune cookie" description as a means of giving feedback. Below this LIFESTYLE FINDER are 15 web sites or web pages, consisting of 5 sites in each of 3 different categories: (1) things you can buy, (2) places you can go, and (3) stores you can shop at. The user can click on any of these web site descriptions to visit the site and can give LIFESTYLE FINDER feedback on the appropriateness of each suggested web site. This feedback is used to evaluate LIFESTYLE FINDER's performance.

Experimental Results

The LIFESTYLE FINDER agent, in the form of Waldo the Web Wizard, was launched on the web on 26 August 1996, and to date, it has been used over 20,000 times. Out of a base of the first 16,082 users, 6,950 users submitted their opinions of the agent's output, 2,926 users proceeded to browse some of the suggested uniform resource locators (URLs), and 4,067

From your answers, Waldo sees you in younger days listening to "In-A-Gadda-Da-Vida" in a haze of incense. Or perhaps it was "Stayin' Alive" in a frenzy of platform shoes and strobe lights. Now it's the music of Barney and Disney's mouse that fill your home. You love to exercise and to travel. You're health conscious. You feel the best you've ever felt. And yet Waldo sees more . . .

○ Sounds right ○ Somewhat ○ A bit ○ Nope

Things you can buy

○ Y ○ N CD Link - this Web site offers truespeech audio for most new albums, and some .WAV samples of every track on a CD.

○ Y ○ N Wall Street Journal Online: provides information regarding the Wall Street Journal, including how to subscribe to the online edition.

○ Y ○ N Mover Sportswear - live by aspiration, not by expectation. This Web site contains information on Mover products, a catalog and ski tips.

○ Y ○ N Webfoot Car Lot - Acura information: everything you ever wanted to know about Acuras.

○ Y ○ N Homelite Chainsaws - detailed information about Homelite chainsaws, with pricing info.

Places you can go

○ Y ○ N Sara's City Workout: give the Internet a workout - find out about the latest aerobics seminars and conventions from Sara.

Figure 4. A Portion of LIFESTYLE FINDER*'s Output.*

users filled out and submitted a follow-up survey. The data recorded from these three sets of users form the basis of the analysis of LIFESTYLE FINDER's performance and the effectiveness of the demographic generalization method.

The measurements of LIFESTYLE FINDER's effectiveness use a control group that is mixed into the agent's suggestions for every user of the system. Each user receives 15 suggested URLs in 3 categories, of which 3 (1 in each category) are randomly selected from the entire corpus of URLs rather than from the URLs relevant to the user's demographic profile. These randomly selected URLs are placed in random positions in the URL lists. In this way, one can compare how each user responded to the suggestions from the agent versus random URLs, providing an objective baseline for evaluation.

One critical question that needs to be analyzed is how dependent the method is on the source of the demographic data. As discussed previously, LIFESTYLE FINDER is based on demographic data covering only the United States but is used by the web's international audience. To measure this effect, my colleagues and I have distinguished subsets of the base of users that we are reasonably certain are browsing the web from within the United States. Although it is impossible to precisely determine a web user's country of origin, the system is designed to conservatively select U.S. users and err on the side of selecting too few users as U.S. based rather than too many. We found that roughly 36 percent of the users in our measurements were known to be in the United States, the rest living outside the United States or accessing the system from a site who's national origin was uncertain. Each measurement in tables 1 through 3 is broken down into users worldwide and U.S. based. In general, there is much less a difference between LIFESTYLE FINDER's performance on worldwide and U.S.-based users than expected, supporting LIFESTYLE FINDER's method as generally applicable.

The first measurement of accuracy is based on users answering yes or no to whether they found that suggested URLs matched their interests. In most cases, these responses were

Measurement	Worldwide	U.S. Based
Number of users submitting feedback on suggestions	5680	2039
Percentage of agent-generated URLs that the users liked	44%	43%
Percentage of random URLs that the user liked	31%	28%
Percentage of users liking four times as many agent-generated URLs as random URLs (to account for ratio of suggestions)	60%	62%

Table 1. Responses to LIFESTYLE FINDER*'s Suggested Uniform Resource Locators (URLs).*

Measurement	Worldwide	U.S. Based
Number of users clicking on URLs	2926	995
Percentage of users clicking on agent-generated URLs	89%	89%
Percentage of users clicking on random URLs	27%	31%
Percentage of users clicking on four times as many agent-generated URLs as random URLs	72%	69%

Table 2. Analysis of the Uniform Resource Locators (URLs) Selected (Clicked on) by the Users.

made before the users saw the suggested web sites and were based on the one-line description. Table 1 shows the data for users worldwide and within the United States. For the number of users given on the first line of the figure (for users worldwide and in the United States), the next two lines give the percentages of agent-generated URLs and randomly generated URLs that the users indicated were appropriate. The final line is the percentage of users that liked the agent-generated URLs more than four times as often as the randomly generated URLs, which accounts for the difference in the numbers of suggestions of each type.

It is clear that the percentages of suggested web sites that the user liked are lower than for other intelligent user-profiling systems. LIFESTYLE FINDER's goal is to develop broad profiles of users, spanning a wide variety of areas, and to do so with a stark minimum of input information. These two criteria will necessarily result in suggestions that have a lower rate of accuracy than most other systems, that develop more narrow profiles of users given a larger amount of information.

Notably, the agent's performance was virtually identical for U.S. and worldwide users, even though the agent's profiles are based on U.S. demographics. The reason appears to be that enough of the same patterns in purchasing and other lifestyle interests are similar across cultures, even if they are found in altogether different geographic regions or cultural settings.

Although the first measure of LIFESTYLE FINDER's effectiveness looked at the users explicit assessment of the URL descriptions provided by the system, the second measure looks at which URLs the users decided to actually select to browse. These selections are difficult to analyze given the tendency of users to select one and then skip the rest, for reasons having to do more with user mood and current interests than overall interests. Table 2 shows the data, including the percentage of users that selected more than four times as many agent-generated URLs as random URLs. The data show that agent-generated URLs were selected by the users notably more often than random URLs, even given the fact that four times as many agent-generated URLs were presented. It is notable, however, that a small

Assessment	Worldwide	U.S. Based
Sounds like me	24%	23%
Somewhat like me	36%	35%
Not really like me	18%	19%
Not at all like me	22%	23%

Table 3. Users' Subjective Assessments of LIFESTYLE FINDER*'s Description of Their Lifestyles.*

The demographic generalization approach to user profiling appears to make effective use of a small amount of information about a user by leveraging a large-scale demographic system.

number of users chose to select any URLs at all, compared to the number of users who provided feedback on URL interest in table 1.

The third measure of the agent's performance is the subjective assessment by the users of the agent's accuracy in describing their lifestyles. Although this measure is entirely subjective and is difficult to trust as a rigorous evaluation metric, it nevertheless provides a sense of the agent's accuracy beyond the particular URLs that were suggested. Users had the option of rating LIFESTYLE FINDER's "fortune cookie" lifestyle description as one of four grades. Table 3 shows the breakdown of user response in worldwide and U.S. populations. Again, we see that the results were virtually identical for the worldwide and U.S.-based populations.

Overall, we see that LIFESTYLE FINDER has performed substantially better than randomly at profiling users in terms of the 62 demographic clusters in the PRIZM system and was able to recommend URLs to users effectively enough to be valuable in a number of contexts. The agent performed equally well for users worldwide and in the United States. This result might be because the demographic system was clustered along general purchasing and activity patterns, not along geographic or social patterns.

Demographic Generalization and User Profiling

The demographic generalization approach to user profiling appears to make effective use of a small amount of information about a user by leveraging a large-scale demographic system. The ability to operate on a small amount of innocuous information comes at the expense of the accuracy that the system is able to achieve.

One application for this method is to improve the accuracy of online information targeting such as web advertising. Most online ads are now targeted in a random fashion, and the use of a method such as we described here would allow them to be much more accurate without requiring personal information about the users. In a survey, 80 percent of users worldwide would like to see online advertisements targeted in this fashion, and 89 percent think that LIFESTYLE FINDER's approach is a good one for this purpose. One benefit of this approach for this application is that the agent does not need any personal or private information about the user. Given the attention paid to online privacy, this benefit can be significant. Ninety-three percent of users surveyed agreed that LIFESTYLE FINDER's questions did not invade their privacy.

Another application that my colleagues and I are beginning to explore is the use of this approach to bootstrap user profiles that can then be refined and improved with other, more information-intensive methods. For example, LIFESTYLE FINDER's approach could determine general areas of interest for an individual, and other methods can be used to learn specific likes and dislikes within these areas.

As research in the area of intelligent user profiling continues, methods will be developed that use a wide range of information about users and generate profiles of various accuracy and applicability. Demographic generalization and other approaches are only the start of what will be an important area for intelligent agents and network-based systems in general.

Notes

1. The research discussed here was carried out at Andersen Consulting's Center for Strategic Technology Research, 3773 Willow Road, Northbrook, IL 60062, USA. The URL is www.ac.com/cstar. The LIFESTYLE FINDER agent is currently available at lifestyle.cstar.ac.com/lifestyle. Thanks to Anatole

Gershman for his many ideas on the project, and to Mark Jacobson for doing the programming.

2. The data are all U.S. based, but as I discuss later, the approach appears effective internationally.

3. The demographic data shown here are modified slightly because of their proprietary nature.

4. LIFESTYLE FINDER's appeal is reflected by the fact that 73 percent of the users that see the opening page run the system, 43 percent of the users that run the system give feedback, and 43 percent of the users filled out a follow-up survey.

References

Krulwich, B. 1995. Learning User Interests across Heterogeneous Document Databases. Presented at the 1995 Spring Symposium on Information Gathering from Heterogeneous Distributed Environments, 27–29 March, Stanford, California.

Krulwich, B., and Burkey, C. 1996. Learning User Information Interests through the Extraction of Semantically Significant Phrases. Presented at the 1996 AAAI Spring Symposium on Machine Learning in Information Access, 27–29 March, Stanford, California.

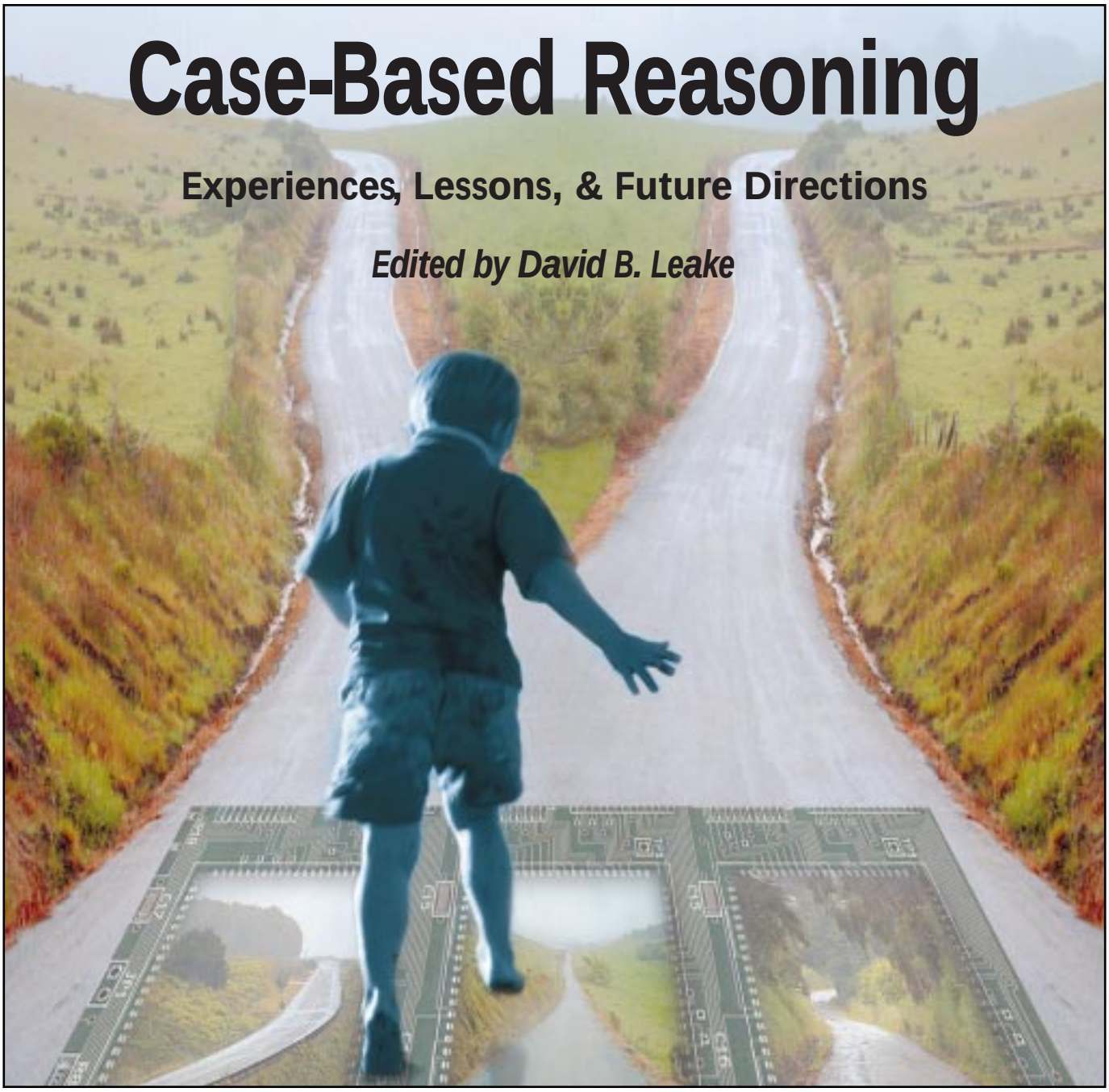
Lashkari, Y.; Metral, M.; and Maes, P. 1994. Collaborative Interface Agents. In Proceedings of the Twelfth National Conference on Artificial Intelligence, 444–449. Menlo Park, Calif.: American Association for Artificial Intelligence.

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.

Sheth, B., 1994. A Learning Approach to Personalized Information Filtering. Master's thesis, Electrical Engineering and Computer Science Dept., Massachusetts Institute of Technology.

Weiss, M., 1988. *The Clustering of America*. New York: Harper and Row.

Bruce Krulwich is a senior research scientist at AgentSoft Ltd., where he leads the Future Technologies and Applications Group. Prior to joining AgentSoft, he was a research scientist at Andersen Consulting's Center for Strategic Technology Research, where he served as the principal investigator for the Intelligent Agents Project. His research interests include intelligent agents, machine learning, information management, and planning. He received his Ph.D. from The Institute for the Learning Sciences at Northwestern University and has served on the program committees for the national AI conference and the Autonomous Agents Conference. His e-mail address is brucek@agentsoft.com.



Case-based reasoning is a flourishing paradigm for reasoning and learning in artificial intelligence, with major research efforts and burgeoning applications extending the frontiers of the field. this book provides an introduction for students as well as an up-to-date overview for experiences researchers and practitioners. It examines the field in a "case-based" way, through concrete examples of how key issues — including indexing and retrieval, case adaptation, evaluation, and application of CBR methods — are being addressed in the context of a range of tasks and domains. Complementing these case studies are commentaries by leading researchers on the lessons learned from experiences with CBR and visions for the roles in which case-based reasoning can have the greatest impact.

ISBN 0-262-62110-X. 420 pp., index. $40.00 softcover

Published by The AAAI Press. Distributed by The MIT Press,
Massachusetts Institute of Technology, Cambridge, Massachusetts 02142.
To order, call toll free: (800) 356-0343 or (617) 625-8569.
www.aaai.org