SITeseer is a Web-page recommendation system that uses an individual's bookmarks and the organization of bookmarks within folders for predicting and recommending relevant pages. Siteseer utilizes each user's bookmarks as an implicit declaration of interest in the underlying content, and the user's grouping behavior (such as the placement of subjects in folders) as an indication of semantic coherency or relevant groupings between subjects. In addition, Siteseer treats folders as a personal classification system which enables it to contextualize recommendations in classes defined by the user. Over time, Siteseer learns each user's preferences and the categories through which they view the world, and at the same time it learns for each Web page how different communities or affinity-based clusters of users regard it. Siteseer then delivers personalized recommendations of online content, Web pages, organized according to each user's folders.

Bookmarks (including Hotlists and Favorites) are a desirable mechanism for gathering preference information as they are already maintained by the user, and thus require no additional behavior for the purpose of informing the recommendation system. In contrast to a click, which can be inadvertently done and rarely takes much effort or investment, bookmarks are the result of a very intentional act, something which (especially if the bookmark is placed in a folder) takes some degree of thought and effort, making them a less "noisy" input for inference.

Bookmarks also have specific limitations. A voluntary survey of free response and multiple choice questions, posted to various Usenet news groups and drawing 40 respondents, indicated that users typically bookmark fewer than half of the sites/pages they find interesting, often because a site is easily accessible.
ble through other means such as another Web page or a search engine. In addition, users tend to bookmark for wildly different reasons, ranging from genuine interest to a transient need to return to a page. Finally, bookmarks either exist, or they don’t. There is no partial bookmark that would indicate marginal interest, and there is no way that bookmarks can be used to indicate a lack of preference, which an explicit feedback system can request.

Siteseer uses the findings of one user as implicit recommendations for another based on the bookmarked discoveries of a pool of reviewers qualified as trusted recommenders. The criteria for a reviewer being a recommender for another is straightforward. Fundamentally, Siteseer looks at each user’s folders and bookmarks, and measures the degree of overlap (such as common URLs) of each folder with other people’s folders, giving additional weight to URLs that are more obscure, (that is, less prevalent in user folders). The system does not derive any semantic value from the contents of the URLs, nor the title of the folders; it uses the URL as a unique identifier and completely ignores the title. By using overlap of contents to determine folder similarity, Siteseer establishes concept similarity without relying on the titles given to the folders.

Using this method, Siteseer forms dynamically defined virtual communities of interest, particular to each user and specific to each of the user’s categories of interest. In our example, John has a “Vacation Spots” folder which, like any other folder, is the basis for the formation of a virtual community. In this case, Mary’s “Tropical Getaways” folder has the highest overlap with John’s “Vacation Spots,” making her the most qualified recommender.

---

1This explanation is significantly simplified but conveys the main point; the mechanism for determining folder similarity is more complex and involves additional factors not mentioned.
However, the converse is not true; for the virtual community surrounding Mary’s “Tropical Getaways,” John’s “Vacation Spots” folder would not be the most qualified. The most qualified recommender would either be Christine, on the basis of her “Caribbean Paradise” folder, or someone else, based on a more highly correlated folder not shown in the diagram. By computing community membership relative to each folder (and not attempting to form a finite set of clusters), Siteseer imposes no category rigidity and is thus able to serve users even with relatively obscure or highly specific interests.

Siteseer provides as recommendations those pages which have been bookmarked by the user’s virtual neighbors, giving preference to pages drawn from folders with the highest overlap as well as those held within multiple folders in the neighborhood. Siteseer contextualizes its recommendations by delivering them in the folder served as the basis for the discovery. Thus, in our example, recommendations coming from users in John’s “Vacation Spots” neighborhood will be delivered within the context of his “Vacation Spots” folder.

While a novel recommendation and categorization system, Siteseer has intrinsic limitations because of its almost purely collaborative approach. Most notable is its inability to help the first-time user or one creating a new category for bookmarks, due to the simple fact that until a community has been discovered, there is no collective experience to leverage. Therefore, such users must first arrive at sites or pages by some other means, such as a search service or editorially built Web directory.

At press time, Siteseer had a fairly limited base of just over 1,000 users and performed best on folders having between 15 and 20 bookmarks, making recommendations 88% of the time, with an average confidence factor of 18 out of 100 (which is partially indicative of folder overlap). For smaller and larger sets, it performed fewer and lower-confidence recommendations. Overall, it was able to provide recommendations for 60% of the folders.

James Rucker (james@imana.com) is president and co-founder of San Francisco-based Imana, Inc., a pioneering firm in the field of collaborative filtering technology.

Marcos J. Polanco (marcos@imana.com) is chair and co-founder of Imana, Inc., San Francisco, Calif.

Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication and its date appear, and notice is given that copying is by permission of ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or a fee.

© ACM 0002-0782/97/0300 $3.50