

Experiments in Social Data Mining: The TopicShop System

BRIAN AMENTO, LOREN TERVEEN, and WILL HILL

AT&T Labs—Research

and

DEBORAH HIX and ROBERT SCHULMAN

Virginia Tech

Social data mining systems enable people to share opinions and benefit from each other's experience. They do this by mining and redistributing information from computational records of social activity such as Usenet messages, system usage history, citations, or hyperlinks. Some general questions for evaluating such systems are: (1) is the extracted information valuable? and (2) do interfaces based on the information improve user task performance? We report here on *TopicShop*, a system that mines information from the structure and content of Web pages and provides an exploratory information workspace interface. We carried out experiments that yielded positive answers to both evaluation questions. First, a number of automatically computable features about Web sites do a good job of predicting expert quality judgments about sites. Second, compared to popular Web search interfaces, the TopicShop interface to this information lets users select significantly more high-quality sites, in less time and with less effort, and to organize the sites they select into personally meaningful collections more quickly and easily. We conclude by discussing how our results may be applied and considering how they touch on general issues concerning quality, expertise, and consensus.

Categories and Subject Descriptors: H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—*hypertext navigation and maps*; H3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*retrieval models*

General Terms: Experimentation, Human Factors

Additional Key Words and Phrases: Cocitation analysis, collaborative filtering, computer-supported cooperative work, information visualization, social filtering, social network analysis

1. INTRODUCTION

We live in an age of information abundance. The Internet in particular confronts us with endless possibilities: Web sites to experience, music to listen to, conversations to participate in, and every conceivable consumer item to buy.

This article revises and expands material originally presented in Amento et al. [2000a,b].

Authors' addresses: B. Amento, W. Hill, AT&T Labs—Research, Bldg. 103, B282, 180 Park Avenue, Florham Park, NJ 07932; email: brian@research.att.com; D. Hix, R. Schulman, Virginia Tech; L. Terveen, University of Minnesota.

Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that the 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 the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

© 2003 ACM 1073-0616/03/0300-0054 \$5.00

Not only are there vast numbers of possibilities, but they also vary widely in quality. People thus face a difficult information management problem: they cannot hope to evaluate all available choices by themselves unless the topic of interest is severely constrained.

One information management task many Web users perform is *topic management*, gathering, evaluating, and organizing relevant information resources for a given topic. Sometimes users investigate topics of professional interest, at other times topics of personal interest. Users may create collections of Web information resources for their own use or for sharing with coworkers or friends. For example, someone might gather a collection of Web sites on wireless Internet services as part of a report she's preparing for her boss and a collection on *The X-Files* television show as a service to her fellow fans. Librarians might prepare topical collections for their clients, and teachers for their students [Abrams et al. 1998].

Current Web tools do not support this task well; specifically, they do not make it easy to evaluate collections of Web sites to select the best ones or to organize sites for future reuse and sharing. Users have to browse and view sites one after another until they are satisfied they have a good set or, more likely, they get tired and give up. Browsing a Web site is an expensive operation, both in time and cognitive effort. And bookmarks, the most common form of keeping track of Web sites, are a fairly primitive technique for organizing collections of sites.

Our approach to this problem combines *social data mining* [Terveen et al. 2001] with *information workspaces* [Card et al. 1991]. As a type of *recommender system* [Goldberg et al. 1992; Resnick et al. 1994; Resnick and Varian 1997, pp. 56–58; Shardanand and Maes 1995], a social data mining system mediates the process of sharing recommendations. In everyday life, when people have to make a choice without any personal knowledge of the alternatives, they often rely on the experience and opinions of others. They seek recommendations from people who are familiar with the choices they face, who have been helpful in the past, whose perspectives they value, or who are recognized experts. Social data mining systems extract information from computational activity records that can be used to recommend items. For our application, we need information to help people evaluate Web sites. To help with the topic management problem, social data mining can be applied to Web usage logs [Pitkow and Pirolli 1997; Viegas and Donath 1990] (answering the question, “Which sites are visited most by other people?”), online conversations [Hill and Terveen 1996; Smith and Fiore 2001; Viegas and Donath 1990] (answering the question, “Which sites are talked about most by other people?”), or the structure of the Web itself [Aggarwal et al. 1999; Bharat and Henzinger 1998; Chakrabarti et al. 1998; Goldberg et al. 1992; Pitkow and Pirolli 1997] (answering the question, “Which sites are linked to most by other people?”). TopicShop takes the final strategy.

Once the information has been extracted, it must be made available to users. TopicShop provides an information workspace in which collections of sites can be explored, ordered by any of the extracted features, added to a personal collection, grouped into subsets, and annotated.

We report here on two interrelated experiments. The first study answers the question, “Is user task performance enhanced by the TopicShop interface?” The

answer is, “Yes.” TopicShop users were able to select significantly more high-quality sites, in less time and with less effort. Furthermore, they were able to organize the sites they selected into more elaborate personal collections, again in less time. Finally, they easily integrated the two tasks of selecting sites and organizing the selected sites. The second answers the question, “Is the information that TopicShop extracts about Web sites valuable?” The answer again is, “Yes”—we found that features based on the number of incoming hyperlinks to a site, as well as simple counts of various types of content on a site, correlate well with expert quality judgments.

The remainder of this article is organized as follows. We first discuss related work and use it to situate our approach. Second, we describe the details of the TopicShop system. Third, we describe the two studies we just mentioned. We conclude by summarizing the contributions of this work, identifying areas for future work, and discussing some general issues.

2. RELATED WORK

2.1 Social Data Mining

The motivation for the social data mining approach goes back at least to Vannevar Bush’s [1945] “As We May Think” essay. Bush envisioned scholars blazing trails through repositories of information and realized that these trails subsequently could be followed by others. Everyone could walk in the footsteps of the masters. In our work, we have formulated a similar intuition using the metaphor of a path through the woods. However, this metaphor highlights the role of collective effort, rather than the individual. A path results from the decisions of many individuals, united only by where they choose to walk, yet still reflects a rough notion of what the walkers find to be a good path. The path both reflects history of use and serves as a resource for future users.

Social data mining approaches seek analogous situations in the computational world. Researchers look for situations where groups of people are producing computational records (such as documents, Usenet messages, or Web sites and links) as part of their normal activity. Potentially useful information implicit in these records is identified, computational techniques to harvest and aggregate the information are invented, and visualization techniques to present the results are designed. Thus, computation discovers and makes explicit the “paths through the woods” created by particular user communities. And, unlike ratings-based *collaborative filtering* systems [Resnick et al. 1994; Shardanand and Maes 1995], social data mining systems do not require users to engage in any new activity; rather, they seek to exploit user preference information implicit in records of existing activity.

The “history-enriched digital objects” line of work [Hill et al. 1992; Hill and Hollan 1994] was a seminal effort in this approach. It began from the observation that objects in the real world accumulate *wear* over the history of their use, and that this wear—such as the path through the woods or the dog-eared pages in a paperback book or the smudges on certain recipes in a cookbook— informs future usage. *Edit Wear* and *Read Wear* were terms used to describe computational analogues of these phenomena. Statistics such as time spent

reading various parts of a document, counts of spreadsheet cell recalculations, and menu selections were captured. These statistics were then used to modify the appearance of documents and other interface objects in accordance with prior use. For example, scrollbars were annotated with horizontal lines of differing length and color to represent amount of editing (or reading) by various users.

Other work has focused on extracting information from online conversations such as Usenet. PHOAKS [Hill and Terveen 1996] mines messages in Usenet newsgroups looking for mentions of Web pages. It categorizes and aggregates mentions to create lists of popular Web pages for each group. Donath and colleagues [Viegas and Donath 1990] have harvested information from Usenet newsgroups and chats and have used them to create visualizations of the conversation. These visualizations can be used to find conversations with desirable properties, such as equality of participation or many regular participants. Smith and Fiore [1998] also extracted information from newsgroups and designed visualizations of the conversational thread structure, contributions by individual posters, and the relationships between posters.

Still other work has focused on extracting information from Web usage logs. Footprints [Wexelblat and Maes 1999] records user browsing history, analyzes it to find commonly traversed links between Web pages, and constructs several different visualizations of these data to aid user navigation through a Web site. Pursuing the metaphor of navigation, some researchers have used the term *social navigation* to characterize work of this nature [Munro et al. 1999]. Finally, a distinct technical approach was taken by Chalmers and colleagues [1998]. They used the activity *path* (e.g., a sequence of URLs visited during a browsing session) as the basic unit. They have developed techniques to compute similarities between paths and to make recommendations on this basis, for example, to recommend pages to you that others browsed in close proximity to pages you browsed.

2.2 Mining the Web

Most relevant to the concerns of this article is work that mines the structure of the World Wide Web itself. The Web, with its rich content, link structure, and usage logs, has been a major domain for social data mining research. A basic intuition is that a link from one Web site to another may indicate both similarity of content between the sites and an endorsement of the linked-to site. An intellectual antecedent for this work is the field of bibliometrics, which studies patterns of cocitation in texts [Egghe and Rousseau 1990; Garfield 1979]. Various clustering and rating algorithms have been designed to extract information from link structure. Pirolli et al. [1996a] developed a categorization algorithm that used hyperlink structure (as well as text similarity and user access data) to categorize Web pages into various functional roles. Later Pitkow and Pirolli [1997] experimented with clustering algorithms based on cocitation analysis, in which pairs of documents were clustered based on the number of times they were both cited by a third document.

Kleinberg [1998] formalized the notion of document quality within a hyperlinked collection using the concept of *authority*. At first pass, an authoritative

document is one to which many other documents link. However, this notion can be strengthened by observing that links from all documents aren't equally valuable: some documents are better *hubs* for a given topic. Hubs and authorities stand in a mutually reinforcing relationship: a good authority is a document that is linked to by many good hubs, and a good hub is a document that links to many authorities. Kleinberg developed an iterative algorithm for computing authorities and hubs. He presented examples that suggested the algorithm could help to filter out irrelevant or poor quality documents (i.e., they would have low authority scores) and identify high-quality documents (they would have high authority scores). He also showed that his algorithm could be used to cluster pages within a collection, in effect disambiguating the query that generated the collection. For example, a query on "Jaguar" returned items concerning the animal, the car, and the NFL team, but Kleinberg's algorithm splits the pages into three sets, corresponding to the three meanings.

Several researchers have extended this basic algorithm. Chakrabarti et al. [1998] weighted links based on the similarity of the text that surrounded the hyperlink in the source document to the query that defined the topic. Bharat and Henzinger [1998] made several important extensions. First, they weighted documents based on their similarity to the query topic. Second, they counted only links between documents from different *hosts*, and averaged the contribution of links from any given host to a specific document. That is, if there are k links from documents on one host to a document D on another host, then each of the links is assigned a weight of $1/k$ when the authority score of D is computed. In experiments, they showed that their extensions led to significant improvements over the basic authority algorithm.

PageRank [Page et al. 2002] is another link-based algorithm for ranking documents. Like Kleinberg's algorithm, this is an iterative algorithm that computes a document's score based on the scores of documents that link to it. PageRank puts more emphasis on the quality of the links to a particular document. Documents linked to by other documents with high PageRank scores will themselves receive a higher PageRank score than documents linked to by low scoring documents.

In summary, much recent research has experimented with algorithms for extracting information from Web structure. A major motivation for these algorithms is that they can be used to compute measures of document quality. Yet this work has proceeded without much experimental evaluation, leaving some basic questions unanswered. First, what benefits do the more complicated link-based algorithms provide beyond simple link counts? And second, how well do the various link-based metrics (in-links, authority scores, PageRank scores) actually correlate with human quality judgments? We report on an experiment that investigates these issues.

2.3 Information Workspaces

Once information has been extracted, it must be presented in a user interface. Users must be able to evaluate collections of items, select items they find useful, and organize them into personally meaningful collections. Card et al. [1991]

introduced the concept of *information workspaces* to refer to environments in which information items can be stored and manipulated. A departure point for most such systems is the file manager popularized by the Apple Macintosh and then in Microsoft Windows. Such systems typically include a list view, which shows various properties of items, and an icon view, which lets users organize icons representing the items in a 2-D space. Mander et al. [1992] enhanced the basic metaphor with the addition of “piles.” Users could create and manipulate piles of items. Interesting interaction techniques for displaying, browsing, and searching piles were designed and tested.

Bookmarks are the most popular way to create personal information workspaces of Web resources. Bookmarks consist of lists of URLs; typically the title of the Web page is used as the label for the URL. Users may organize their bookmarks into a hierarchical category structure. Abrams et al. [1998] carried out an extensive study of how several hundred Web users used bookmarks. They observed a number of strategies for organizing bookmarks, including a flat ordered list, a single level of folders, and hierarchical folders. They also made design recommendations to help users manage their bookmarks more effectively. First, bookmarks must be easy to organize, for example, via automatic sorting techniques. Second, visualization techniques are necessary to provide comprehensive overviews of large sets of bookmarks. Third, rich representations of sites are required; many users noted that site titles are not accurate descriptors of site content. Finally, tools for managing bookmarks must be well integrated with Web browsers.

Many researchers have created experimental information workspace interfaces, often designed expressly for Web documents. Card et al. [1996] describe the WebBook, which uses a book metaphor to group a collection of related Web pages for viewing and interaction, and the WebForager, an interface that lets users view and manage multiple WebBooks. In addition to these novel interfaces, they also presented a set of automatic methods for generating collections (WebBooks) of related pages, such as recursively following all relative links from a specified Web page, following all (absolute) links from a page-one level, extracting “book-like” structures by following “next” and “previous” links, and grouping pages returned from a search query. Mackinlay et al. [1995] developed a novel user interface for accessing articles from a citation database. The central UI object is a “Butterfly,” which represents an article, its references, and its citers. The interface makes it easy for users to browse among related articles, group articles, and generate queries to retrieve articles that stand in a particular relationship to the current article. The Data Mountain of Robertson et al. [1998] represents documents as thumbnail images in a 3-D virtual space. Users can move and group the images freely, with various interesting visual and audio cues used to help users arrange the documents. In a study comparing the use of Data Mountain to Internet Explorer Favorites, Data Mountain users retrieved items more quickly, with fewer incorrect or failed retrievals.

Other researchers have created interfaces to support users in constructing, evolving, and managing collections of information resources. SenseMaker [Baldonado and Winograd 1997] focuses on supporting users in the contextual evolution of their interest in a topic. It attempts to make it easy to evolve a

collection, for example, expanding it by query-by-example operations or limiting it by applying a filter. Scatter/Gather [Pirolli et al. 1996b] supports the browsing of large collections of text, allowing users to iteratively reveal topic structure and locate desirable documents. Marshall and Shipman's VIKI system [Marshall et al. 1994] lets users organize collections of items by arranging them in 2-D space. Hierarchical collections are supported. Later extensions [Shipman et al. 1999] added automatic visual layouts, specifically nonlinear layouts such as fisheye views. Hightower et al. [1998] addressed the observation that users often return to previously visited pages. They used Pad++ [Bederson et al. 1996] to implement PadPrints, browser companion software that presents a zoomable interface to a user's browsing history.

3. TOPICSHOP

The *TopicShop* system supports users in gathering, evaluating, and organizing collections of Web sites. It consists of a Web crawler, which constructs collections of Web sites and associated information, and an interface that provides access to these data. This article focuses on the interface; however, we first give a brief overview of the Web crawler.

3.1 TopicShop Web Crawler

A Web crawler takes a set of Web pages as input (we refer to these as the "seeds"). It then fetches these pages, extracts the links to other pages, and selects some subset of the linked-to pages to fetch. The process of fetching pages, extracting links, and selecting new pages to fetch continues until some stopping condition (e.g., a maximum number of pages fetched) is reached.

Web crawlers can be distinguished by (1) the procedure they use to decide which pages to fetch, and (2) the way they analyze fetched pages. Some Web crawlers attempt to index as much of the Web as possible, so they can employ a simple control strategy such as breadth-first search. Other crawlers, however, attempt to maintain a *focus*, that is, to fetch pages that are conceptually related to the seeds. Our crawler is one of this sort. It uses heuristics based on link connectivity and text similarity to decide which pages to fetch. However, in the studies that we report in this article, we constrained the crawler simply to fetch and analyze the sites containing the seed pages, so we do not go into the details of these heuristics here.

After fetching a specified number of pages, the crawler groups the pages into sites. It does this using heuristics that look at the directory structure of URLs and the set of pages that have been fetched. For example, if the crawler is processing a URL `http://a/b/page1.html`, and `http://a/b/index.html` already has been determined to be the root page of a site, it records this URL as part of the site. A number of specific heuristics handle large hosting sites such as geocities, tripod, and so on. (The problem is that a site like geocities.com hosts many distinct, unrelated user sites; thus, to use a simplified example, two URLs `www.geocities.com/a/x.html` and `www.geocities.com/b/y.html` must be treated as belonging to different sites.) In addition, links between URLs are aggregated to their containing sites; in other words, if a URL u_1 links to a URL u_2 , and u_1

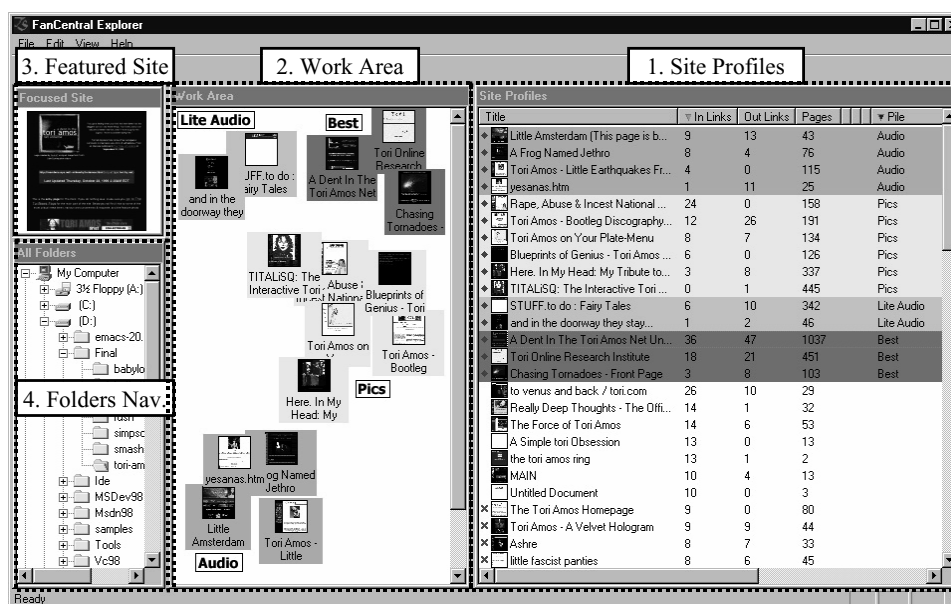


Fig. 1. TopicShop Explorer.

belongs to site s_1 , and u_2 belongs to site s_2 , then a link is recorded from site s_1 to s_2 .

As pages are fetched, their data are analyzed to produce *profiles* of the page's content. Like links, profile information also is aggregated to the site level; for example, a site records all the pages known to belong to it. As discussed below, the purpose of profile information is to help users evaluate a site. Profiles contain the data:

- title (of the site's root page),
- a thumbnail image (of the site's root page),
- links to and from other sites, and
- internal html pages, images, audio files, and movie files.

3.2 The TopicShop Explorer: Evaluating, Managing, and Sharing Collections of Resources

Topic collections can be viewed and managed using the *TopicShop Explorer* (Figure 1), an interface modeled on the normal Macintosh file manager/Windows file explorer. The TopicShop Explorer is a Windows application, written in C++ and Microsoft Foundation Classes. It interprets and processes site profile information.

The interface features two main linked views, the Site Profiles view (indicated as window 1 in Figure 1) and the Work Area (window 2). The other two windows show a thumbnail image of the currently featured site (window 3) and enable navigation through the folder space (window 4). Users can show/hide whichever views they choose to aid them in focusing on their current task. The

purpose of the Site Profiles view is to provide access to site profile information, thus helping users evaluate sites. The purpose of the Work Area is to allow users to create personally meaningful organizations of sites. We explain the interface further as we discuss the main design goals of the TopicShop Explorer.

3.2.1 *Basic Design Goals.*

Make Relevant but Invisible Information Visible. The first goal of TopicShop is to help users evaluate and identify high quality sites. We sought to achieve this goal by providing site profile data and interface mechanisms for viewing and exploring the data. Making these data visible means that users no longer had to select sites to browse based solely on titles and (sometimes) brief textual annotations. (A chief complaint of subjects in the Abrams et al. [1998] study was that titles were inadequate descriptors of site content, and that was for sites that users already had browsed and decided to bookmark.) Instead, users may visit only sites that have been endorsed (linked to) by many other sites or sites that are rich in a particular type of content (e.g., images or audio files). Users can sort sites in the Site Profiles view by any property (e.g., number of in-links, out-links, images, etc.) simply by clicking on the label at the top of the appropriate column. Users can “drill down” to investigate the profile data in detail, for example, to see a list of all the audio files on a site and all the other sites that it links to or that link to it. And users can browse the site in their default Web browser just by double-clicking it.

Make It Simple for Users to Explore and Organize Resources. The second goal is to make it easy for users to organize collections of sites for their own future use and to share with others. The Work Area lets users organize sites by arranging them spatially. This builds on the results of Nardi and Barreau [1995], who found that users of graphical file systems preferred spatial location as a technique for organizing their files. We believe spatial organization is particularly useful early in the exploration process while users are still discovering important distinctions among resources and user-defined categories have not yet explicitly emerged. Thumbnail images also serve as effective memory aids to help users identify sites they already have visited.

These two design goals were maintained throughout the iterations of TopicShop development. Informal user testing and a significant pilot study (described below) also contributed significantly to the interface design.

3.2.2 Features Resulting from User Feedback. We developed TopicShop using an iterative design process, gathering user feedback early in the design process, conducting a pilot study to gain insights into user requirements for the task of topic management [Amento et al. 1999], and incorporating our findings into further design iterations. The primary modifications to our initial design were incorporated to fulfill the following user needs.

Two Always Visible, Linked Views Support Task Integration and a Cleaner Definition of Each Task. The initial version of TopicShop contained two alternate views. In one view, users could evaluate sites using the detailed site profile

information (similar to the current Site Profiles view). In the other view, users could spatially organize larger thumbnail icons representing each site (similar to the Work Area view). Both were views onto the same collection of sites, but (exactly as in the Microsoft Windows Explorer) only one view was available at a time and users were forced to choose which view they wanted to see. However, the pilot study showed that users wanted to integrate work on the evaluation and organization tasks. First, they wanted to be able to organize items without losing sight of the detailed information contained in the site profiles. One subject commented:

I really want to organize the large icons, but don't want to lose the detailed information. Switching all the time is too painful, so I have to settle for the details view only.

Second, we realized that most items in a collection never needed to be organized, because users would not select them as worthy of further attention. Thus, most of the icons were useless and just got in the way of the user's task. Rather than supporting a single collection, a better design would support two data sets. Users can evaluate the initial, machine-generated collection based on the site profiles shown in the interface and select promising items for additional indepth analysis. With this approach, only selected items are organized, thus saving the user valuable time for further browsing.

In response, we redesigned the interface as shown in Figure 1. In this design, the site profile data and a work area for organizing sites are visible at all times. Items in the initial collection are displayed in the Site Profiles window, and the Work Area is initially empty (in Figure 1, the Work Area already is populated with the results of one subject from the user study). As users evaluate items and find good ones, they select them simply by dragging the items and dropping them in the Work Area. Because icons are created just for selected items, the Work Area is uncluttered, and provides a clear picture of the sites about which users care.

"Piling" Icons Makes It Easy to Create First-Class Groups by Spatial Arrangement. The original version of TopicShop let users make explicit groups by creating folders. Users also could arrange icons in spatial groups; however, the system did not know that these spatial groups should be treated as semantic groups. We looked for a grouping technique with the ease of spatial arrangement, but with the added benefit that the system also knows that a certain set of sites should be treated as a group.

The current design enables this by recognizing when users place one icon "next to" another. When this happens a group is formed automatically, or extended, if one icon already is part of a group. (This technique is similar in spirit to that of Mander et al. [1992], although the interface and implementation details are quite different.) How close two icons must be before they are considered to be part of a group is a system parameter, set by default to occur just when their bounding boxes touch. Each group is assigned a different color. Because the Site Profile and Work Area views are linked, both the group of icons in the Work Area and the sites' profile data in the Site Profiles window

are displayed using that color as a background. To help users better organize their groups, they can perform operations on piles (i.e., move, name/annotate, arrange, and select) as well as the normal operations on single sites.

Multilevel sorting is a useful operation that can be applied to a group; it also illustrates how the linked views support task integration. In the Site Profiles view, users can reorder the sites based on primary and secondary sort keys. In the earlier version of TopicShop, users commonly sorted first by the groups they defined in folders and then by some additional feature, such as in-links or number of pages. To support this operation in the new design, we built in a multilevel sorting technique that lets users evaluate and compare sites within a single group. Figure 1 shows just such a sort. In fact, users can sort by any two keys by simply defining the primary and secondary sort keys with mouse-clicks on the header fields. A left click on any of the header fields defines that field as a primary sort key, indicated by a blue arrow, and a right click defines the field as a secondary sort key, indicated by a green arrow. Additional clicks on the header fields reverse the direction of the sort.

Visual Indicators Make the Task State Apparent. The status of the user's task must be manifest. Most important, it has to be clear which items in the initial collection users have already evaluated and which they have not. Un-evaluated items are a kind of agenda of pending work. Subject comments made this clear:

An indication of whether or not I visited the site would be useful. I can't tell what I've already seen.

It's hard to know what you've looked at and what you haven't...

We modified the TopicShop interface to respond to this issue. Any site included in the Work Area now is marked with a green diamond in the Site Profile view and kept at the top for easy reference. Users can mark irrelevant or low-quality sites for "deletion"; this marks the sites with a red X and moves them to the bottom of the list to get them out of the way. Thus, users quickly see which sites they have already processed (selected or deleted) and which need additional evaluation.

Annotations and Large Thumbnails Support Reuse and Sharing. Subjects observed that including more graphical and textual information could improve site recall. Many subjects asked for the ability to annotate both individual sites and groups of sites. (Note that annotations also make collections more informative for others.) And other subjects asked for a larger thumbnail image to provide a better visual cue:

A larger thumbnail would be nice... It can be used to refresh your memory... and would be more effective if it looked more like the site.

The Focused Site window (upper left of Figure 1) displays a large thumbnail of the most recently clicked-on site. Users can create textual annotations for piles or individual sites in the Work Area. Annotations become visible as

“pop ups” when the user lets the cursor linger over an object (pile or individual thumbnail) for a second or two.

3.3 The Experiments

We have described the TopicShop system, its main features, and their rationale. We now turn to evaluating system performance. Specifically, we address two main research issues.

- Does the TopicShop system improve user task performance? In particular, can users *select* higher-quality sites and *organize* them more effectively? To answer this question, we require certain things. First, we need a baseline system to which TopicShop can be compared. For the task of evaluating and selecting sites, we chose Yahoo, a widely used search tool on the Web [Amento et al. 2000b]. For the task of organizing sites, we chose Netscape Communicator bookmarks, because bookmarks and the equivalents in other browsers are the primary means by which users organize Web sites. Second, we need some definition of a “quality” site; this brings us to the other main issue.
- Is the site profile information valuable? In particular, do any features that can be computed for a site tell us anything about the site’s *quality*? To answer this question, we need an independent measure of site quality. We obtain such a measure by having human topic experts rate site quality.

We had to select specific topics for the experiments. Because of its important role on the Web, we selected five topics from the popular entertainment domain, the television shows *Babylon 5*, *Buffy The Vampire Slayer*, and *The Simpsons*, and the musicians Tori Amos and Smashing Pumpkins. In addition to its popularity, we also believe that this domain is representative of other domains characterized by rich content and many links between sites, including popular scientific topics such as Mars exploration. However, we emphasize that nothing about the TopicShop system or our experiment was specific to this domain. (Indeed, we later discuss a completely different task for which the TopicShop interaction paradigm has proved effective.) We obtained a collection of sites for each topic from the appropriate category in the Yahoo hierarchy. We then applied our Web crawler to each collection to obtain the site profiles and thumbnail images that TopicShop uses.

Figure 2 gives an overview of the interface evaluation and quality assessment experiments. In the following sections, we consider each of the experiments in some detail.

4. RESEARCH QUESTION I: CAN WE IMPROVE USER TASK PERFORMANCE?

For the interface evaluation, we recruited 40 subjects from a local university, none of whom had any prior knowledge of or exposure to TopicShop. The experiment was a 2×5 , between subjects design, with user interface and topic as factors. Subjects were randomly assigned an interface and topic. To begin the experiment, subjects received 15 minutes of instruction and training in the task and user interface. The experiment combined two tasks, evaluation and

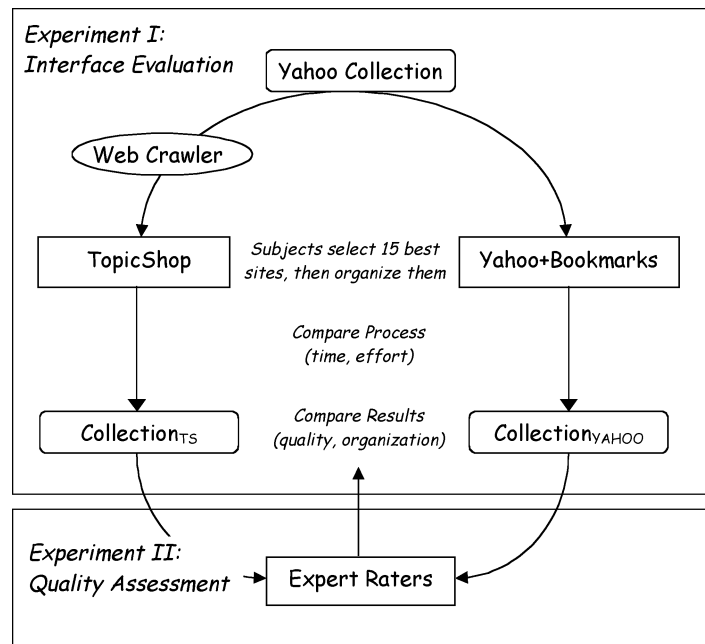


Fig. 2. Overview of the experiments.

organization. For the evaluation task, subjects investigated the sites for their assigned topic by using the interface (TopicShop or Yahoo) and browsing sites. They were asked to choose the 15 best sites (where the “best” items were defined as those that together gave a useful and comprehensive overview for someone wanting to learn about the topic). Subjects were given 45 minutes to complete the task and were kept informed of the time, although they could take more time if they chose. For the organization task, we instructed subjects to organize their selected sites into groups and annotate the groups with descriptive labels. The interfaces were instrumented to record subject actions and store them in log files.

There is a relationship between time on task and quality of results: the more time spent, the better results one can expect. By limiting the amount of time, we hoped to focus on any differences in the quality of results (i.e., the selected sites) between the two interfaces. And people do not spend unlimited amounts of time browsing, so we wanted to see whether users could find high-quality sites in a limited amount of time.

4.1 Results

Our results showed the benefits of TopicShop in supporting both the evaluation and organization tasks and in enabling task integration. We present quantitative data and subject comments to illustrate these benefits.

4.1.1 Supporting the Evaluation Task. TopicShop subjects selected significantly higher-quality sites than did Yahoo subjects. Section 5 gives the details of

Table I. Intersection Between User Selections and Top 15 Expert-Rated Sites

Topic	TopicShop	(%)	Yahoo	(%)	% Difference
Babylon 5	7.0/15	(47)	5.75/15	(38)	9
Buffy	7.25/15	(48)	3.5/15	(23)	25
Simpsons	6.5/15	(43)	5.25/15	(35)	8
Smashing Pumpkins	8.5/15	(57)	5/15	(33)	24
Tori Amos	7.75/15	(52)	3/15	(20)	32
Average	7.4/15	(49)	4.5/15	(30)	19

how the quality of sites was determined. Here, it suffices to say that, after topic experts rated sites, we defined the top sites for each topic as those with the 15 highest average expert scores. The quality of each subject's sites was measured by counting how many were among the top 15 expert sites. Table I shows the results for each topic and interface condition. On average, TopicShop subjects selected 19% more high-quality sites: 7.4 of the 15 expert sites versus 4.5 for Yahoo subjects. A 2×5 two-factor ANOVA (interface and topic) shows that the main effect of interface was significant ($F(1, 30) = 18.55, p < .0002$). Neither topic nor the interaction between topic and interface was significant (Topic: $F(4, 30) = 0.656, p < 0.627$; Interaction: $F(4, 30) = 1.097, p < 0.376$). Because topic is not significant and the interaction also has no effect, the remaining statistical results are reported using t -tests of independent means.

Second, we wanted to be sure that users didn't gain quality by putting in more effort, so we measured the amount of time subjects spent on their tasks and the total number of sites they browsed. Again, TopicShop subjects had the advantage. They took about 72% of the time of Yahoo subjects (38 vs. 53 minutes) (independent means t -test, $t(38) = -4.219, p < 0.007$), and they browsed about 67% as many sites (27 vs. 40) (independent means t -test, $t(38) = -4.788, p < 0.001$). The latter fact showed that they based many of their judgments on the data presented in the TopicShop interface, rather than browsing site content.

In summary, TopicShop subjects selected higher-quality sites, in less time and with less effort. We believe these benefits are due to TopicShop's site profile data. User comments and survey responses support this belief.

TopicShop subjects commented on the utility of the information they saw.

It presented me with lots of information very quickly. I could get a feel for what the site had to offer before visiting it, saving time to find the info that interested me. I got more than a site description, I got site facts.

The different sorting methods make it very easy for you to find what you're looking for.

And Yahoo subjects asked were near unanimous in asking for more information to judge sites.

[Show] some sort of popularity information to evaluate the sites.

[Show] something like an indication of how popular [the sites] were. Some rating of content.

Add some sort of ranking, that would be nice.

[Show] number of web pages, top 10 most visited.

List the type of audio or video offered on the multimedia pages.

I would add the approximate graphic level [i.e., number of images on a site] (so as to be able to judge the worthiness).

Subjects also were given a survey. The TopicShop survey included a question asking them to rate the utility of the site profile features. Number of in-links was first, and number of pages was second (responses were similar in the pilot study). Interestingly, in the next section we present results showing that both in-links and number of pages are good predictors of site quality, so subjects proved accurate in their utility judgments.

4.1.2 Supporting the Organization Task. The second part of the subjects' task was to organize their selected sites into groups and to name the groups. Recall that in the TopicShop condition, subjects grouped items by piling them together, whereas Yahoo subjects created folders in the Netscape Communicator Bookmarks window and placed items in the folders.

We defined a number of metrics to measure performance on the organization task. The metrics characterize the effort involved, the level of detail of the organization, and the amount of agreement between subjects on how sites should be grouped.

First, we examined the log files to compute how much time subjects spent on the organization task. TopicShop subjects spent 18% of their total time (7 out of the 38 total minutes), and Yahoo subjects spent 36% of theirs (19 out of the 53 total minutes) (independent means t -test, $t(38) = -4.893$, $p < 0.009$). Because TopicShop subjects spent less of their total time organizing sites, they were able to devote more time to evaluating and understanding the content of sites and selecting the good ones. Yet, even while taking less time, TopicShop users still created finer-grained and more informative organizations, as we now detail.

Second, we computed the number of groups subjects created. TopicShop subjects created four groups on average, and Yahoo subjects created three. Thus, TopicShop subjects articulated the structure of the topic somewhat more. In addition, TopicShop subjects grouped nearly all of their 15 selected sites (3% were left ungrouped), and Yahoo subjects left more ungrouped (15%).

Third, TopicShop subjects created more site annotations, thus making their collections more informative for their own use or for sharing with others. The experiment didn't require subjects to annotate sites (only groups). Yet 10 of 20 TopicShop subjects did so, annotating a total of 15% of their 15 selected sites. Two Yahoo subjects annotated a total of four sites.

Fourth, TopicShop subjects tended to agree more about how sites should be grouped. This is a difficult issue to investigate; in general, it requires interpreting the semantics of groups. We computed a simpler metric: for each pair of subjects within a topic and interface condition, for each pair of sites that they both selected, did they group the sites together? If both subjects grouped the pair of sites together, or both grouped them separately, we counted this as agreement; otherwise, we counted it as disagreement. Table II summarizes the results. It shows that TopicShop subjects agreed 68% of the time on average,

Table II. Agreement in Grouping Items

Topic	Average Agreement		
	TopicShop	Yahoo	Average Difference
Babylon 5	0.78	0.39	0.39
Buffy	0.59	0.44	0.15
Simpsons	0.78	0.36	0.42
Smashing Pumpkins	0.75	0.53	0.22
Tori Amos	0.48	0.41	0.07
Total	0.68	0.43	0.25

and Yahoo subjects agreed 43% of the time; thus, TopicShop subjects agreed 25% more (independent means t -test, $t(58) = -3.486$, $p < 0.005$). TopicShop subjects, on average, created more categories than Yahoo subjects, so random agreement would be less likely to occur between TopicShop subjects, yet they actually agreed more often than Yahoo subjects.

Taken cumulatively, the results show that TopicShop subjects appear to do a better job of organizing the items they select—they create more groups, they annotate more sites, and they agree in how they group items more of the time—and achieve these results in half the time Yahoo subjects devote to the task. We believe these results are due to the ease of grouping and annotation and because the rich information contained in site profiles remains visible while users organize sites. Subject comments support these beliefs.

—TopicShop subjects found it easy to group sites.

Piling Web sites and annotating them makes grouping easy. You can easily see an overview of the organization.

Easily viewing category annotations and colored groups in the Work Area helps when attempting to determine what the important areas within a topic are.

—Thumbnail images and textual annotations were effective memory aids for identifying sites and recalling their distinctive properties; TopicShop users commented on their utility, and Yahoo users expressed a desire for these types of functionality.

Treating a site as a graphical object that can be dragged and dropped like anything else in your normal windows environment was much easier to conceptualize than treating sites as text links that required cutting, pasting, editing [TopicShop subject].

A thumbnail of the site. . . would help the user who has been using several sites remember the site by looking at its thumbnail [Yahoo subject].

I used annotations to remind me about a site so I could tell the difference from the many other sites that I looked at [TopicShop subject].

Some way to take notes while surfing would be useful [Yahoo subject].

Table III. Interleaving Tasks

Quartile	TopicShop		Yahoo	
	# of Actions	% of Total	# of Actions	% of Total
Quartile 1	125	23	2	1
Quartile 2	138	26	31	18
Quartile 3	110	21	50	29
Quartile 4	160	30	89	52
Total	533		172	

4.1.3 *Relationship Between Evaluation and Organization Tasks.* We also studied the relationship between the evaluation and organization tasks. The TopicShop Explorer allows the tasks to be integrated, but doesn't force it. On the other hand, in the Yahoo/bookmarks condition, browsing sites and organizing bookmarks are separate tasks.

The log files contain data that let us quantify the relationship between tasks. Each user action is timestamped, and we know whether it was an evaluation or organization action. Evaluation actions included visiting a page in a Web browser and sorting data in the Site Profiles Window. For TopicShop, organization actions included moving or annotating icons or groups in the Work Area. In the Yahoo/bookmarks condition, organization actions included creating a bookmarks folder, naming a folder, naming a bookmarked item, and placing an item in a folder.

We computed how many actions of each type occurred in each quartile of the task; that is, how many occurred in the first 25% of the total time a subject spent on task, how many in the second 25%, and so on. Table III shows the results for organizational actions. First, it shows how much more organizational work TopicShop users did: 533 actions versus 172. (Recall they did this in half the time.) Second, as expected, TopicShop users integrated organization and evaluation to a much greater extent than did Yahoo users. They did about a quarter of their total organizational work in each of the first two quartiles, dipped slightly in the third quartile, then increased a bit in the final quartile. Yahoo users, on the other hand, did virtually no organizational work in the first quartile of their task, then ended by doing more than 50% in the last quartile. We should emphasize that TopicShop does not force task integration; rather, it enables it. And when users had the choice, they overwhelmingly preferred integration.

We also can construct detailed timelines of user activity. Figure 3 shows such timelines for two Yahoo and two TopicShop subjects. They provide vivid illustrations of the overall results. TopicShop users interleaved the two tasks throughout the course of their work and performed many more organization actions. On the other hand, Yahoo users began by focusing exclusively on evaluation; then, toward the end of the task, they shifted to focus mostly on organization. And they did much less organization.

Several comments showed that subjects appreciated the ability to integrate tasks and the fact that their task state remained visible.

—Linked views helped users integrate the evaluation and organization tasks.

In particular, they could evaluate within groups they created.

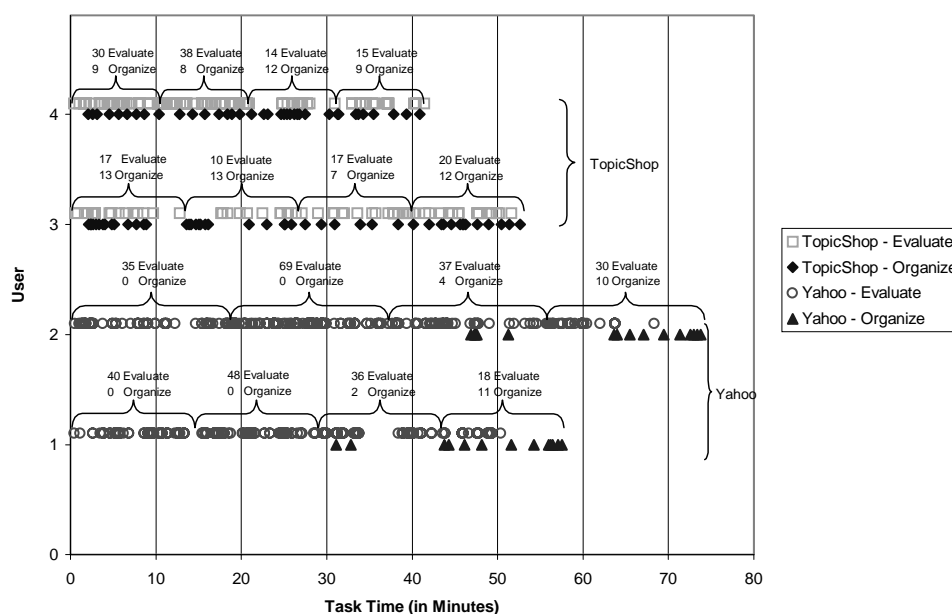


Fig. 3. Example timelines of user activity. TopicShop users did more organization actions and interleaved organization with evaluation. Yahoo/bookmarks users did less organization and did it at the end of their task.

Coloring was nice, because it gives me the ability to quickly SEE what was in what pile. Sorting within a pile was helpful for picking things out of each pile.

—TopicShop made the state of the task apparent, allowing users to treat the initial collection of sites as an agenda of items to be processed.

The graphics indicators let you quickly see what's left, because they show what you've already picked and what you didn't like.

5. RESEARCH QUESTION II: CAN WE PREDICT HUMAN QUALITY JUDGMENTS?

We now turn to a discussion of our second research issue, how to obtain quality ratings for Web sites. We needed these ratings to evaluate the sets of sites our subjects selected in the first experiment. As we said before, there has been little empirical data concerning the utility of hyperlinks as a metric of site quality or differentiating between various algorithms. In this experiment, we compare the various link-based metrics and evaluate their effectiveness at selecting quality sites.

Researchers have experimented with different metrics for computing the quality of Web sites. As we discussed in the Related Work section, link analysis algorithms, in particular, have received much attention. However, there has been little empirical evaluation of these algorithms. This leaves a fundamental issue unresolved: do link-based metrics work that is, do they correlate with

human judgments of quality? We're actually interested in a more general question, namely, whether *any* metrics we can compute for Web documents are good predictors of document quality. Accordingly, we'll investigate content-based as well as link-based metrics.

We encountered several other questions while investigating this issue. First, we wondered to what extent topic experts agree on the quality of items within a topic. If human judgments vary widely, this suggests limits on the utility of automatic methods (or perhaps that collaborative filtering, which can personalize recommendations for an individual, may be more appropriate). More fundamentally, it would call into question whether a shared notion of quality even exists. Conversely, if experts do tend to agree in their quality judgments, our confidence in the concept of quality will be bolstered, even if it is difficult to give a precise definition.

Second, we wondered whether there were any significant differences between various link analysis algorithms: for example, would one score documents D_1 , D_2 , and D_3 highly, and another score D_4 , D_5 , and D_6 more highly? If there are no such differences, then an algorithm can be chosen for other factors, such as simplicity and efficiency.

5.1 Experiment/Data Sets

For this experiment, topic experts rated the quality of sets of items obtained from the subjects. We solicited self-identified topic experts from our organization, offering each \$20 for participation. We obtained four experts for *The Simpsons*, and three for each of the other topics. Experts rated the quality of items on a scale of 1 (worst) to 7 (best). Experts rated items by filling out a Web-based form; the form presented no information about items other than the URL, so experts had to browse the sites to judge their quality. Each expert's form presented items in a random order.

The five collections we used in this study ranged from about 90 to over 250 sites. However, based on our previous experience, we knew that experts could not do a good job if they were asked to rate more than about 40 sites, so it was not possible for all sites to be rated. It wasn't even possible to rate the union of all sites that subjects selected in experiment 1; this would have resulted in sets ranging in size from about 50 to 70. Instead, we had experts rate all sites selected by more than one subject and a sample of sites selected by one or no subjects. The rationale was that high-quality sites were likely to have been selected by more than one subject. Table IV summarizes the various data sets just discussed. The major point to note is that all of the sites selected by multiple subjects were chosen for the final expert data-set along with a random sampling of singly-selected and unselected sites.

5.2 Features Computed for the Study

To compute link and content-based metrics to compare with the expert ratings, we had to analyze the Web neighborhood surrounding the items. We did this by applying our Web crawler/analyzer to each collection of items we obtained from

Table IV. Constructing Datasets for Experts to Rate^a

Topic	Number of Sites from Yahoo	Multiply Selected Sites (All Chosen)	Singly Selected Sites (Chosen/Total)	Sites Not Selected (Chosen/Total)	Total Expert Dataset
Babylon 5	173	28	8/42	4/103	40
Buffy	258	29	8/39	4/190	41
Simpsons	210	21	12/38	6/151	39
Smashing Pumpkins	95	33	6/16	3/46	42
Tori Amos	88	36	4/19	2/33	42
Total	824	147	38/154	19/523	204

^aDatasets included all sites selected by multiple subjects (e.g., there were 28 such sites for the *Babylon 5* topic), a sample of singly selected sites (for *Babylon 5*, 8 of 42 sites were selected), and a sample of unselected sites (for *Babylon 5*, 4 of 103 sites were selected).

Yahoo. As mentioned previously, for this experiment, we limited the crawler to consider only URLs on the same site as one of the seeds; we did this by accepting only URLs that contained some seed URL as a prefix.

When the crawling is complete, URLs are aggregated into sites (as described above). In addition to the basic URL graph—whose nodes are URLs, and whose edges represent hyperlinks between URLs—this results in a site graph—whose nodes are sites, and whose edges represent a hyperlink from (any URL on) one site to (any URL on) another.

From these graphs, we computed five link-based features; in- and out-degree, Kleinberg’s authority and hub scores, and the PageRank score. In all cases, we computed features for both the site and the root URL of the site. Computing these metrics at the site level was straightforward. When we computed at the URL level, we followed Bharat and Henzinger [1998] by (1) counting only links between URLs on different sites, and (2) averaging the contribution of links from all the URLs on one site to a URL on another.

The crawler also computes a set of content-based features for each URL. Page size and the number of images and audio files are recorded. This information is aggregated to the site level, and the total number of pages contained on each site also is recorded.

Finally, the crawler computes text similarity scores. Although we consider relevance and quality to be different notions (we discuss this more in Section 6.1), we wanted to test whether relevance would help predict quality. The crawler uses the Smart IR System [Buckley 1985] to generate a centroid (a weighted vector of keywords) from the content of the seed items for each topic. The relevance score of each page is based on the inner product similarity of the page’s text to the centroid. And for each site, the relevance score of the root page, the text similarity score of any contained page, and the average relevance scores of all contained pages are recorded.

Each of the features induces a ranking of the items in our data set. In subsequent analysis, we examine how well the various rankings match human quality judgments. To summarize, here is a list of all the features we used

Table V. Pairwise Expert Agreement Using Correlation^a

Topic	Correlations Between Pairs of Experts						Avg
	1-2	2-3	1-3	1-4	2-4	3-4	
Babylon 5	0.91	0.92	0.76				0.87
Buffy	0.75	0.79	0.83				0.79
Smashing Pumpkins	0.80	0.73	0.69				0.74
Tori Amos	0.61	0.63	0.50				0.58
Simpsons	0.52	0.59	0.50	0.75	0.59	0.59	0.59
Total							0.71

^aThere were four experts for *The Simpsons* and three for all other topics.

(at both the site and page level):

- In-degree: number of sites that link to this page/site,
- Kleinberg’s Authority Score,
- PageRank Score,
- Out-degree: number of pages/sites this site links to,
- Kleinberg’s Hub Score,
- Text similarity score: similarity to topic seed text,
- Size (number of bytes and number of contained pages),
- Number of images, and
- Number of audio files.

5.3 Results

5.3.1 Do Experts Agree? We first investigated how much experts agreed in their quality judgments. To the extent they do agree, we gain confidence that there is a shared notion of quality within the topic areas we investigated. We did computations to measure agreement. First, we correlated the scores assigned to items by each pair of experts for each topic. (Recall that we had four experts for *The Simpsons* and three for all other topics.) We used the Pearson product-moment correlation because the expert averages represent interval data, ranging from 1 to 7. Table V presents the results. It shows that almost all pairs of experts were highly correlated in their judgments of item quality (all correlations were significant, $p < 0.01$).

These results suggest that experts agree on the nature of quality within a topic, and that the expert judgments thus can be used to evaluate rankings obtained by algorithms. However, there is considerable variation between topics. Some variation may be due to properties of the topics. For example, we noticed that one or two *Tori Amos* sites were of quite high quality, but somewhat tangential relevance to the topic. This may have influenced some experts’ quality judgments. Second, some variation in opinions is inevitable, particularly in the area of popular entertainment, where there is no objective quality standard. Thus, results might be different in a technical domain, where there are more objective quality standards. One expert may be more interested in one type of content than another (e.g., song lyrics vs. tour schedules). Some experts may have highly idiosyncratic tastes. Where tastes do differ significantly, a collaborative

Table VI. Metric Similarity Using Correlations

Topic	In-Degree/ Authority	In-Degree/ PageRank	Authority/ PageRank
Babylon 5	0.97	0.93	0.90
Buffy	0.92	0.85	0.70
Simpsons	0.97	0.99	0.95
Smashing Pumpkins	0.95	0.98	0.92
Tori Amos	0.97	0.92	0.88
Average (Spearman)	0.96	0.93	0.87
Average (Kendall)	0.86	0.83	0.75

filtering approach ultimately will be superior. We discuss the relationship of quality to expertise and consensus in quality judgments in more detail later.

5.3.2 Are Different Link-Based Metrics Different? The second issue we investigated was whether the three link-based metrics—in-degree, authority, and PageRank—ranked items differently.

Because the different metrics use different scales that do not maintain a linear relationship, we converted raw scores into ranks and used Spearman’s rho rank correlation on the resulting ordinal data. We computed correlations between each pair of metrics.

Table VI presents the results. The correlations were extremely high (and were all significant, $p < 0.01$). We also computed the Kendall tau rank correlation. Correlations again were high, although not quite as high as Spearman’s rho; the final row in Table VI presents the average Kendall correlations.

These results (and results we present below) show no significant difference between the link-based metrics. In-degree and authority are particularly similar. This should be surprising: the primary motivation for the authority algorithm was that in-degree isn’t enough, that all links are not equal. Do our results prove this assumption false? No, but they require further consideration. The discussion below illuminates where and why more sophisticated algorithms may be needed.

By starting with items from Yahoo, we almost guaranteed that items in the neighborhood graph we constructed would be relevant to the topic. In contrast, other evaluations of Kleinberg’s algorithm [Bharat and Henzinger 1998; Chakrabarti et al. 1998; Kleinberg 1998] have begun with much noisier neighborhoods. Typically, they’ve started with a base set of items returned by a search engine, many of which are of dubious relevance, and then added items that link to or are linked to by items in the base set. This sort of neighborhood is likely to contain many pages that are not relevant to the original query. Kleinberg argued that although some of these irrelevant pages have high in-degree, the pages that point to them are not likely to have high out-degree; in other words, they don’t form a coherent topic. In such cases, the authority/hub algorithm will assign low scores to some items with high in-degree.

To follow through with this argument, we see that two processes are going on: obtaining a set of relevant items, and rating the quality of the items in this set. As commonly conceived, the authority algorithm helps with both. However,

Table VII. Number and Proportion of Good Items

Topic	Total	# Good	Proportion Good
Babylon 5	40	19	0.48
Buffy	41	15	0.37
Simpsons	39	10	0.26
Smashing Pumpkins	41	7	0.17
Tori Amos	42	13	0.31
Average			0.32

our experiment shows that if one already has a set of relevant items, in-degree alone may be just as good a quality measure. And because many manually constructed collections of topically relevant items are available from general purpose or topic-oriented directories, there are many practical situations where in-degree is a suitable metric.

A further note is that the in-degree metric we're using is *site* in-degree. By aggregating links to the site level, we avoid the problems Bharat and Henzinger identified (links between pages belonging to a common site and mutually reinforcing relationships between two sites). They showed that solving these problems resulted in significant improvements to the basic authority algorithm. The site in-degree metric accrues the same benefits.

5.3.3 Can We Predict Human Quality Judgments? We tested how well the rankings induced by each of the site profile features matched expert quality judgments. We wanted to compute the *precision* of each ranking, the number of items ranked highly by each metric that actually were (according to the experts) high quality. To do this, we needed the set of high-quality items for each topic. We defined the good items as those that a majority of experts rated as good (i.e., scored 5, 6, or 7). Table VII shows the total number of items for each topic, number of good items, and proportion of good items. The proportion of good items serves as a useful baseline; for example, it tells us that, across all topics, if you picked a set of 10 items at random, you'd expect about 3 to be high quality.

For ease of presentation, we present results for 10 metrics (all of the 9 site-based metrics as well as the size of the root page). The same 5 metrics performed best in all analyses, so we include them. We also found that all site-based metrics outperformed their URL-based counterparts in all cases (e.g., number of images on the entire site was better than number of images on the root page), so we omitted the URL-based versions. None of the text-relevance metrics performed well, but we include the best (text similarity score) for the sake of comparison.

Using the set of good items, we computed the precision at 5 and at 10 for each metric.¹ The precision at 5 (10) is the proportion of the 5 (10) highest ranked items that were among the top 5 (10) items as ranked by the experts. Table VIII presents the results, with metrics ordered by average precision at 5. The table shows that the top five metrics all perform quite well. For example, the in-degree metric has a precision at 5 of 0.76: on average, nearly 4 of the first 5 documents it returns would be rated good by the experts. This is more than

¹Since there were only seven high-quality items for Smashing Pumpkins, we could not compute precision at 10 for this topic. Accordingly, the average precision at 10 is for the other four topics.

Table VIII. Precision at 5 and 10 (Ordered by Average Precision at 5)

Metric		Babylon 5	Buffy	Simpsons	Smashing Pumpkins	Tori Amos	Average
In-degree	at 5	0.8	0.8	0.8	0.8	0.6	0.76
	at 10	0.6	0.7	0.6	N/A	0.5	0.6
# Pages on site	at 5	0.8	1	0.6	0.6	0.6	0.72
	at 10	0.8	0.8	0.5	N/A	0.4	0.63
Authority score	at 5	0.8	0.6	0.8	0.8	0.6	0.72
	at 10	0.7	0.7	0.5	N/A	0.5	0.6
PageRank score	at 5	1	0.8	0.6	0.8	0.4	0.72
	at 10	0.7	0.6	0.6	N/A	0.4	0.58
# Images on site	at 5	1	0.6	0.6	0.6	0.4	0.64
	at 10	0.8	0.7	0.5	N/A	0.5	0.63
Out-degree	at 5	0.8	0.4	0.4	0.4	0.6	0.52
	at 10	0.5	0.5	0.5	N/A	0.5	0.5
# Audio files on site	at 5	0.2	0.4	0.6	0.6	0.8	0.52
	at 10	0.2	0.2	0.5	N/A	0.6	0.38
Hub score	at 5	0.8	0.2	0.4	0.4	0.6	0.48
	at 10	0.4	0.5	0.4	N/A	0.5	0.45
Text Similarity	at 5	0.4	0.6	0.6	0.2	0.4	0.44
	at 10	0.7	0.5	0.5	N/A	0.4	0.53
Root Page Size	at 5	0.6	0	0.4	0.4	0.2	0.32
	at 10	0.5	0.2	0.3	N/A	0.2	0.3

double the number of good documents you would get by selecting 5 at random from the sites that the experts rated. And recall that most of the items that experts rated were of good quality, as they were selected by multiple subjects in Phase 1 of our experiment. Thus, we speculate that in a larger data set that contains items of more widely varying quality, the improvement obtained by using these metrics would be even greater.

Inasmuch as the link-based metrics were highly correlated, it should be no surprise that they have similar precision. However, it is surprising how well a very simple metric performs: in this data set, simply counting the number of pages on a site gives as good an estimate of quality as any of the link-based computations (and number of images isn't bad, either). We speculate that the number of pages on a site indicates how much effort the author is devoting to the site, and more effort tends to indicate higher quality.

The precision analysis abstracted away from the item scores, which could conceal significant differences. For example, suppose that two metrics have identical precision. In principle, they could return completely different sets of items; furthermore, one metric could return the best (highest ranked) of the good items, and the second return the worst of the good items. Thus, we wanted to do another analysis using item scores to check for this possibility.

We experimented with two different item scoring schemes, the average of all expert scores and a "majority score". The majority score is the ratio of the number of experts rating the item as "good" (i.e., scoring it 5, 6, or 7) and the total number of experts rating the item.

The two methods yielded similar results; here we present results using majority score.

Table IX. Majority Score at 5 and 10 (Ordered by Majority Score at 5)

Metric		Babylon 5	Buffy	Simpsons	Smashing Pumpkins	Tori Amos	Average
Majority Score	at 5	1	1	1	0.9	1	.96
	at 10	1	0.9	0.7	0.7	0.9	.84
In-degree	at 5	0.8	0.7	0.7	0.8	0.5	.71
	at 10	0.6	0.7	0.6	0.4	0.6	.57
Authority score	at 5	0.8	0.5	0.5	0.5	0.5	.69
	at 10	0.7	0.6	0.5	0.4	0.5	.57
PageRank score	at 5	1	0.7	0.5	0.8	0.4	.69
	at 10	0.7	0.6	0.6	0.4	0.4	.53
# Pages on site	at 5	0.7	1	0.6	0.6	0.4	.66
	at 10	0.8	0.8	0.5	0.4	0.3	.56
# Images on site	at 5	0.9	0.7	0.6	0.6	0.3	.62
	at 10	0.8	0.7	0.5	0.4	0.5	.56
# Audio files on site	at 5	0.3	0.5	0.4	0.6	0.8	.52
	at 10	0.2	0.3	0.4	0.4	0.6	.39
Out-degree	at 5	0.7	0.4	0.4	0.4	0.5	.49
	at 10	0.5	0.5	0.4	0.4	0.5	.45
Hub score	at 5	0.7	0.3	0.4	0.4	0.5	.47
	at 10	0.4	0.5	0.4	0.4	0.5	.44
Text Similarity	at 5	0.3	0.5	0.6	0.2	0.3	.39
	at 10	0.6	0.4	0.5	0.3	0.4	.43
Root Page Size	at 5	0.5	0.1	0.2	0.5	0.3	.31
	at 10	0.4	0.2	0.3	0.3	0.3	.28

Table X. Average Expert Scores of Top 10 sites

	Babylon	Buffy	Simpsons	Smashing Pumpkins	Tori Amos	Average
# Pages on site	5.57	5.77	4.20	4.37	3.70	4.72
In-degree	5.13	5.20	4.63	4.14	4.34	4.69
Authority score	5.40	5.07	4.42	4.17	4.37	4.69
# Images on site	5.67	5.37	4.10	4.10	3.97	4.64
PageRank score	5.30	4.93	4.45	4.14	3.74	4.51
Out-degree	4.37	4.43	4.03	3.77	4.33	4.19
Hub score	4.33	4.50	3.73	4.03	4.30	4.18
Text Similarity	4.74	4.17	3.98	3.30	4.47	4.13
# Audio files on site	3.83	3.70	4.05	4.07	4.80	4.09
Root Page Size	4.13	3.30	3.55	3.40	3.80	3.64

Table IX presents the results. For reference, we present the average scores for the top 5 and 10 items as ranked by the expert majority score itself. This is the ideal: no metric can exceed it. A score of 1 (e.g., for majority score at 10 for *Babylon 5*) means that all experts rated all items as good. A score of .8 (e.g., in-degree at 5 for *Smashing Pumpkins*) means that 80% of experts rated all 5 items as good. The best metric is in-degree. It performs about 74% of the ideal at 5 (i.e., .71/.96), and 68% at 10 (i.e., .57/.84).

We also computed a simple average expert score of the top 10 sites for each metric. Table X shows the results of this analysis. Once again, the in-degree, authority score, and PageRank score are in the top 5 metrics along with the

number of images and pages on a site. An 18×5 two-factor ANOVA of metric and topic verified that the results are statistically significant ($F(17, 810) = 3.775$, $p < .0001$) and the interaction between topic and metric was not significant ($F(68, 810) = .683$, $p < .976$). All metrics from Table X were included in the ANOVA at the site and URL level except the text similarity score and root page size which were only calculated for entire sites, resulting in 18 metrics for comparison.

The same metrics—in-degree, authority, PageRank, number of pages, and number of images—are in the top five slots in each of the past five analyses (precision/majority score at 5/10, and expert average), although their order varies a little. We wondered whether there were any significant differences among the metrics, so we performed a post hoc Least Significant Difference test after the above ANOVA. The analysis shows that the results can be broken into three distinct subsets in the data. Tables VIII through X have a triple line indicating where significant differences occur. It turns out that the top five means are not significantly different from each other, but are significantly distinct from the rest of the chart (at the .05 level). This is also true of the next four metrics. (out-degree, hub score, text similarity score, and number of audio files). Root page size was the only metric that was significantly different from every other metric in the table. Notice that once again two very simple content metrics, number of pages and number of images on a site, correlated with expert judgments.

5.4 Summary

We have investigated the utility of various computable metrics in estimating the quality of Web sites. We showed that topic experts exhibit a high amount of agreement in their quality judgments; however, enough difference of opinion exists to warrant further study. We also showed that three link-based metrics and a simple content metric number of pages on a site do a good job of identifying high quality items.

Our results contained two main surprises: that in-degree performed at least as well as the more sophisticated authority and PageRank algorithms, and that a simple count of the pages on a site was about as good as any of the link analysis methods.

6. DISCUSSION

There are a number of important issues that deserve further investigation. One direction is to seek new sources for mining information about user preferences. As we have discussed, researchers have investigated hyperlink structure, electronic conversations, navigation histories and other usage logs, and purchasing history. One area with great potential is electronic media usage, in particular, listening to digital music. By observing what music someone is listening to, a system can infer the songs, artists, and genres that person prefers, and use this information to recommend additional songs and artists, and to put the person in touch with other people with similar interests. We took a step in this direction with a system that lets users visualize individual and group listening

histories and define new playlists relative to listening history [Terveen et al. 2002]. Crossen et al. [2002] report on a system that learns user preferences from the music they listen to, then selects songs to play in a shared physical environment, based in part on the preferences of all people present.

As user preferences are extracted from more and more sources, the issue of combining different types of preferences becomes important. For example, PHOAKS [Hill and Terveen 1996] extracted preferences about Web pages from Usenet messages and presented them to users. As users browsed through this information, PHOAKS tracked which pages users clicked on (another type of implicit preference), and users also could rate Web pages (explicit preferences). Developing general techniques for combining different types of preferences is a challenge. Billsus and Pazzani [1998] presented a method for weighting different types of contributions; however, whether this is the best combination method and how to determine appropriate weights are still open issues.

It is worth pointing out that the task that TopicShop supports—selecting a subset of items from a large set and then organizing the subset that arises—is quite general and occurs in other contexts. For example, of the many people I exchange email with, a small subset are “contacts” whom I wish to keep track of, and organize into groups which I can use to manage my communication. We have applied this intuition in a project with Steve Whittaker, developing a new interface for the ContactMap contact management system [Whittaker et al. 2002a,b]. Features about potential contacts including their organization and frequency and recency of communication are extracted from email archives and presented in a table; as in TopicShop, the table can be sorted by any of the columns. And, when users find important contacts, they add them to their “map” (equivalent to the TopicShop Work Area) by dragging and dropping. Contacts on the map are organized by spatial arrangement and color coding. This experience illustrates that the general interaction paradigm of TopicShop can be applied in an altogether different domain.

Although our experiment compared TopicShop to the state-of-the-art Web directory Yahoo, it may have occurred to the reader that our techniques are suitable for integration with such a system. This is absolutely correct. Both directory systems, which contain categories of Web sites typically built by a person, and search engines, which retrieve documents based on their similarity to a query, could benefit. An effective way to apply the results of our research would be to enhance (say) Yahoo by (1) using a Web crawler/analyzer to augment each manually constructed collection of pages with the sorts of profiles our experiments showed effective, and (2) providing a TopicShop-style information workspace interface. Such a system would combine the advantages of people, applying judgment to select the initial set of collections, and computers, applying analysis techniques to provide enhanced information and to keep the collections up to date. A similar tactic could be taken by a search engine; this would be most efficient for one such as Google that already maintains a database of links between Web pages. Finally, note that this argument shows that even a very large, manually constructed set of “seed” pages can be enhanced significantly by providing additional features, grouping pages into sites, and offering a good user interface.

6.1 Focus on Quality

It is worth considering the issue of Website *quality* in more detail. We gave both subjects and experts a rough working definition, namely, that a quality site should provide generally useful information that serves as a good overview of the topic. And it turned out that experts, when keeping this definition in mind, agreed to a large extent as to what the quality sites for a topic were. However, our work to date only touches on some complex issues, including the following.

- (1) What is the difference between quality and the traditional Information Retrieval metric of relevance?
- (2) How can we be sure that, even if experts agree on the quality of a site, that the experts are right? How do we know that the “experts” really are experts?
- (3) Finally, what if there is no consensus? What if there are significant differences of opinion among the experts?

1. *Quality Versus Relevance.* In Information Retrieval, the standard metric used is *relevance*; human raters are asked to judge how relevant a document is to a specified query (in general, a query also is a document, although it may be expressed as a natural language question, a few keywords, or a short passage of text). Search engines then are evaluated in terms of how many relevant documents they return, and how many of the documents they return are relevant.

We are not arguing that relevance is not a useful concept; we simply think that quality should be treated distinctly. Perhaps an example can clarify the distinction. It seems natural to view a student paper and a collection of literary criticism as equally relevant to their topic (e.g., Shakespeare’s sonnets), while judging the latter to be of much higher quality.

Notice that in our experiment a relevance metric—inner-product text similarity of a site to the centroid text for a topic—was not a good predictor of quality. We speculate that this is because we started with a set of relevant documents; in other words, if there were more variance in relevance, perhaps higher relevance would indicate higher quality. However, in any case, in our experiment, relevance and quality appeared to be different aspects of a site.

2. *Experts, Consensus, and Community.* In our experiments, we used the judgments of self-identified topic experts as the “ground truth” concerning site quality. How firm a basis do we have for doing this? The anthropological theory of “culture as consensus” [Romney et al. 1986] suggests some interesting answers. This theory was developed to deal with situations where an anthropologist is trying to learn about an unfamiliar culture by asking members of that culture (the informants) a series of questions, such as the names of a range of plants. The problem the anthropologist faces is that neither the “correct” answers nor the cultural knowledge of the informants is known. Thus, what is the anthropologist to do when all the informants do not agree? And how many informants are necessary before there is a reasonable probability that the correct answers will emerge? These are the types of questions cultural consensus theory attempts to answer.

The intuition is that the more informants agree with each other, the more likely they are to be correct; indeed, they agree with each other because they possess the correct knowledge. A formal mathematical model was developed based on this intuition. Given a set of answers to a series of questions by a group of informants, it is possible to estimate both the correct answers and the knowledge of each informant.

For our purposes, what is interesting is the assumptions of this theory and a few of the mathematical properties of the model. It assumes that each question has a correct answer, that all questions are of the same degree of difficulty, and that informants answer independently of each other. Finally, the model as stated only works for yes–no, multiple-choice, and fill-in-the-blank questions; in particular, it does not apply to ranking data. However, the theory is fairly robust when some of its assumptions are not satisfied; for example, the model still performs well when the questions vary in difficulty.

This theory seems to speak fairly directly to the notion of quality of Web sites, and of documents in general. Experts are analogous to informants; even more, we have grown used to hearing a wide variety of groups, from scientific researchers and professional communities to dog lovers and science fiction fans described as “sub-cultures,” so the application may seem quite natural. Indeed, we agree that it is very promising. However, applying the theory would require investigating to what extent its assumptions hold, and extending the model as required. For example, it is not clear how accurate it is to say that a given site or document has a “correct” rating or a “correct” position in a rank order. It certainly seems that it is harder to rank some sites than others; to put it another way, one person who knows a lot about *The Simpsons* TV show may know more about one aspect of the show (e.g., famous people who have been “guest voices”), whereas another person knows about another aspect (e.g., Bart’s friends). Finally, one can ask: what happens if there is no shared cultural knowledge? What if the experts don’t agree? What if there aren’t even any “experts”?

3. Collaborative Filtering. Our approach has focused on extracting a single score/rating from collective opinion: do the experts agree that this is a good site? However, as we have mentioned, there is an alternative approach, ratings-based collaborative filtering [Hill et al. 1995; Konstan et al. 1997; Shardanand and Maes 1995]. In this approach, an information seeker receives a personalized recommendation based on the preferences of others with similar tastes. Rather than the “best overall” site, one can get the “best for me.”

This approach has been discussed and analyzed in great detail; we discussed it and contrasted it to social data mining in Terveen and Hill [2001]. A major difference is that collaborative filtering requires extra work from users—one must rate some number of items in order to give the system an idea of one’s preferences—and all users play the same role—everyone rates items and everyone receives recommendations. Where there is a body of collective opinion to be mined, and where there is a strong consensus of opinion on item quality, then social data mining is easier for users and computationally more efficient. Where these assumptions are not met, ratings-based collaborative filtering is a better bet.

7. SUMMARY

The popularity of the World Wide Web has made the problems of information retrieval and management more acute. More people than ever before face the problems of identifying relevant and high-quality information and organizing information for their own use and for sharing with others.

The TopicShop system improves people's ability to solve these problems. First, the features collected in TopicShop can be used to predict which Web sites are of the highest quality. Second, the TopicShop interface provides information and interaction techniques that help people select the best sites from large collections. A user study demonstrated that users can select better sites, more quickly and with less effort. TopicShop also offers 2-D spatial arrangement techniques for creating groups of sites, and thumbnail images and annotations that enhance site recall and make the collections more informative. A study showed that users found it easy and fast to create groups and annotate their work. Finally, TopicShop makes it possible to integrate the two major tasks of evaluating and organizing Web sites. A user study showed that users preferred to integrate these two tasks when permitted by the interface.

8. ACKNOWLEDGMENTS

We thank the many people who participated in our experiments. We also thank Candace Kamm, Julia Hirschberg, and Steve Whittaker for much useful discussion. Finally, we thank the reviewers of the first draft of this article for their extraordinarily detailed and helpful comments.

REFERENCES

- ABRAMS, D., BAECKER, R., AND CHIGNELL, M. 1998. Information archiving with bookmarks: Personal Web space construction and organization. In *Proceedings of CHI'98* (Los Angeles, April), ACM, New York, 41–48.
- AGGARWAL, C. A., WOLF, J. L., WU, K.-L., AND YU, P. S. 1999. Horting hatches an egg: A new graph-theoretic approach to collaborative filtering. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.
- AMENTO, B., TERVEEN, L., AND HILL, W. 2000a. Does “authority” mean quality? Predicting expert quality ratings of Web documents. In *Proceedings of SIGIR'2000* (Athens Greece, July), ACM, New York.
- AMENTO, B., TERVEEN, L., AND HILL, W. 2000b. TopicShop: Enhanced support for evaluating and organizing collections of Web sites. In *Proceedings of UIST 2000* (San Diego, November), ACM, New York, 201–209.
- AMENTO, B., HILL, W., TERVEEN, L., HIX, D., AND JU, P. 1999. An empirical evaluation of user interfaces for topic management of Web sites. In *Proceedings of CHI'99* (Pittsburgh, May), ACM, New York, 552–559.
- BALDONADO, M. Q. W. AND WINOGRAD, T. 1997. An information-exploration interface supporting the contextual evolution of a user's interests. In *Proceedings of CHI'97* (Atlanta, March), ACM, New York, 11–18.
- BEDERSON, B. B., HOLLAN, J. D., PERLIN, K., MEYER, J., BACON, D., AND FURNAS, G. 1996. Pad++: A zoomable graphical sketchpad for exploring alternate interface physics. *J. Visual Lang. Comput.* 7, 3–31.
- BHARAT, K. AND HENZINGER, M. R. 1998. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*.

- BILLSUS, D. AND PAZZANI, M. 1998. Learning collaborative information filters. In *Proceedings of the International Conference on Machine Learning* (Madison WI, July), Morgan Kaufmann, San Francisco.
- BUCKLEY, C. 1985. Implementation of the SMART information retrieval system. Tech. Rep. TR85-686, Department of Computer Science, Cornell University.
- BUSH, V. 1945. As we may think. *The Atlantic Monthly* (July).
- CARD, S. K., ROBERTSON, G. C., AND MACKINLAY, J. D. 1991. The information visualizer, an information workspace. In *Proceedings of CHI'91* (New Orleans, April), ACM, New York, 181–188.
- CARD, S. K., ROBERTSON, G. C., AND YORK, W. 1996. The WebBook and the Web Forager: An information workspace for the World-Wide Web. In *Proceedings of CHI'96* (Vancouver BC, April), ACM, New York, 111–117.
- CHAKRABARTI, S., DOM, B., GIBSON, D., KLEINBERG, J., RAGHAVAN, P., AND RAJAGOPALAN, S. 1998. Automatic resource compilation by analyzing hyperlink structure and associated text. *Comput. Netw. and ISDN Syst.* 30, 65–74.
- CHALMERS, M., RODDEN, K., AND BRODBECK, D. 1998. The order of things: Activity-centred information access. In *Proceedings of the Seventh International Conference on the World Wide Web* (Brisbane, Australia, April), 359–367.
- CROSSEN, A., BUDZIK, J., AND HAMMOND, K. J. 2002. Flytrap: Intelligent group music recommendation. In *Proceedings of IUI'2002* (San Francisco, January), ACM, New York.
- EGGHE, L. AND ROUSSEAU, R. 1990. *Introduction to Informetrics: Quantitative Methods in Library, Documentation, and Information Science*. Elsevier, New York.
- GARFIELD, E. 1979. *Citation Indexing*. ISI, Philadelphia.
- GOLDBERG, D., NICHOLS, D., OKI, B. M., AND TERRY, D. 1992. Using collaborative filtering to weave an information tapestry. *Commun. ACM* 35, 12 (Dec.), 51–60.
- HIGHTOWER, R. R., RING, L. T., HELFMAN, J. I., BEDERSON, B. B., AND HOLLAN, J. D. 1998. Graphical multiscale Web histories: A study of PadPrints. In *Proceedings of Hypertext '98* (Pittsburgh, June). ACM, New York.
- HILL, W. C. AND HOLLAN, J. D. 1994. History-enriched digital objects: Prototypes and policy issues. *Inf. Soc.* 10, 2, 139–145.
- HILL, W. C. AND TERVEEN, L. G. 1996. Using frequency-of-mention in public conversations for social filtering. In *Proceedings of CSCW'96* (Boston, November), ACM, New York, 106–112.
- HILL, W. C., HOLLAN, J. D., WROBLEWSKI, D., AND MCCANDLESS, T. 1992. Edit wear and read wear. In *Proceedings of CHI'92* (Monterey, CA, May), ACM New York, 3–9.
- HILL, W. C., STEAD, L., ROSENSTEIN, M. AND FURNAS, G. 1995. Recommending and evaluating choices in a virtual community of use. In *Proceedings of CHI'95* (Denver, May), ACM, New York, 194–201.
- KLEINBERG, J. M. 1998. Authoritative sources in a hyperlinked environment. In *Proceedings of 1998 ACM-SIAM Symposium on Discrete Algorithms* (San Francisco, January), ACM, New York.
- KONSTAN, J. A., MILLER, B. N., MALTZ, D., HERLOCKER, J. L., GORDON, L. R., AND RIEDL, J. 1997. GroupLens: Applying collaborative filtering to Usenet news (March), 77–87.
- MACKINLAY, J. D., RAO, R., AND CARD, S. K. 1995. An organic user interface for searching citation links. In *Proceedings of CHI'95* (Denver, May), ACM, New York, 67–73.
- MANDER, R., SALOMON, G., AND WONG, Y. Y. 1992. A “pile” metaphor for supporting casual organization of information. In *Proceedings of CHI'92* (Monterey, CA, May), ACM New York, 627–634.
- MARSHALL, C., SHIPMAN, F., AND COOMBS, J. 1994. VIKI: Spatial hypertext supporting emergent structure. In *Proceedings of ACM ECHI '94* (Edinburgh, September). ACM, New York, 13–23.
- MUNRO, A. J., HÖÖK, K., AND BENYON, D., Eds. 1999. *Social Navigation of Information Space*. Springer-Verlag, New York.
- NARDI, B. AND BARREAU, D. 1995. Finding and reminding: File organization from the desktop. *ACM SIGCHI Bull.* 27, 3 (July).
- PAGE L., BRIN, S., MOTWANI, R., AND WINOGRAD, T. 2002. The PageRank citation ranking: bringing order to the Web. Stanford Digital Libraries Working Paper.
- PIROLI, P., PITKOW, J., AND RAO, R. 1996a. Silk from a sow's ear: Extracting usable structures from the Web. In *Proceedings of CHI'96* (Vancouver BC, April), ACM, New York, 118–125.
- PIROLI, P., SCHANK, P., HEARST, M., AND DIEHL. 1996b. Scatter/gather browsing communicates the topic structure of a very large text collection. In *Proceedings of CHI'96* (Vancouver BC, April), ACM, New York, 213–220.

- PITKOW, J. AND PIROLI, P. 1997. Life, death, and lawfulness on the electronic frontier. In *Proceedings of CHI97* (Atlanta, March), ACM, New York, 383–390.
- RESNICK, P. AND VARIAN, H. R., Eds. 1997. *Commun. ACM, Special issue on Recommender Systems* 40, 3 (March).
- RESNICK, P., IACOVOU, N., SUCHAK, M., BERGSTROM, P., RIEDL, J. 1994. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of CSCW94* (Chapel Hill, NC, October), ACM, New York, 175–186.
- ROBERTSON, G., CZERWINSKI, M., LARSON, K., ROBBINS, D. C., THIEL, C., VAN DANTZICH, M. 1998. Data Mountain: Using spatial memory for document management. In *Proceedings of UIST'98* (San Francisco, November), ACM, New York, 153–162.
- ROMNEY, A. K., WELLER, S. C., AND BATCHELDER, W. H. 1986. Culture as consensus: A theory of culture and informant accuracy. *Amer. Anthropol.* 88, 2 (June), 313–338.
- SHARDANAND, U. AND MAES, P. 1995. Social Information Filtering: Algorithms for Automating “word of mouth.” In *Proceedings of CHI95* (Denver, May 1995), ACM, New York, 210–217.
- SHIPMAN, F., MARSHALL, C., AND LEMERE, M. 1999. Beyond location: Hypertext workspaces and non-linear views. In *Proceedings of ACM Hypertext'99*, ACM, New York, 121–130.
- SMITH, M. A. AND FIORE, A. T. 2001. Visualization components for persistent conversations. In *Proceedings of CHI2001* (Seattle, April), ACM, New York, 136–143.
- TERVEEN, L. G. AND HILL, W. C. 2001. Beyond recommender systems: Helping people help each other. In *HCI In The New Millennium*, J. Carroll, Eds. Addison-Wesley, Reading, Mass.
- TERVEEN, L. G., McMACKIN, J., AMENTO, B., AND HILL, W. 2002. Specifying preferences based on user history. In *Proceedings of CHI2002* (Minneapolis, April), ACM, New York.
- VIEGAS, F. B. AND DONATH, J. S. 1999. Chat circles. In *Proceedings of CHI99* (Pittsburgh, May), ACM, New York, 9–16.
- WEXELBLAT, A. AND MAES, P. 1999. Footprints: History-rich tools for information foraging. In *Proceedings of CHI99* (Pittsburgh, May), ACM, New York, 270–277.
- WHITTAKER, S., JONES, Q., AND TERVEEN, L. 2002a. Persistence and conversation stream management: Conversation and contact management. In *Proceedings of HICSS'02*.
- WHITTAKER, S., JONES, Q., NARDI, B., CREECH, M., TERVEEN, L., ISAACS, E., AND HAINSWORTH, J. 2002. ContactMap: Using personal social networks to organize communication in a social desktop. Submitted.

Received June 2001; revised November 2002; accepted November 2002