

# **User Interfaces for Topic Management of Web Sites**

Brian Amento

Dissertation submitted to the Faculty of the  
Virginia Polytechnic Institute and State University  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy  
in  
Computer Science and Applications

Deborah Hix, Chair

Roger Ehrich

Rex Hartson

Will Hill

Robert Schulman

Loren Terveen

September 26, 2001

Blacksburg, Virginia

Keywords: Information Access, Information Retrieval, Information Visualization, Human  
Computer Interaction, Social Filtering, Collaborative Filtering,

Copyright 2001, Brian Amento

# User Interfaces for Topic Management of Web Sites

Brian Amento

## (ABSTRACT)

*Topic management* is the task of gathering, evaluating, organizing, and sharing a set of web sites for a specific topic. Current web tools do not provide adequate support for this task. We created and continue to develop the *TopicShop* system to address this need. TopicShop includes (1) a web crawler/analyzer that discovers relevant web sites and builds site profiles, and (2) user interfaces for information workspaces. We conducted an empirical pilot study comparing user performance with TopicShop vs. Yahoo™. Results from this study were used to improve the design of TopicShop. A number of key design changes were incorporated into a second version of TopicShop based on results and user comments of the pilot study including (1) the tasks of evaluation and organization are treated as integral instead of separable, (2) spatial organization is important to users and must be well supported in the interface, and (3) distinct user and global datasets help users deal with the large quantity of information available on the web. A full empirical study using the second iteration of TopicShop covered more areas of the World Wide Web and validated results from the pilot study. Across the two studies, TopicShop subjects found over 80% more high-quality sites (where quality was determined by independent expert judgements) while browsing only 81% as many sites and completing their task in 89% of the time. The site profile data that TopicShop provide – in particular, the number of pages on a site and the number of other sites that link to it – were the key to these results, as users exploited them to identify the most promising sites quickly and easily. We also evaluated a number of link- and content-based algorithms using a dataset of web documents rated for quality by human topic experts. Link-based metrics did a good job of picking out high-quality items. Precision at 5 (the common information retrieval metric indicating the percentage of high quality items selected that are actually high quality) is about 0.75, and precision at 10 is about 0.55; this is in a dataset where 32% of all documents were of high quality. Surprisingly, a simple content-based metric, which ranked documents by the total number of pages on their containing site, performed nearly as well. These studies give insight into users' needs for the task of topic management, and provide empirical evidence of the effectiveness of task-specific interfaces (such as TopicShop) for managing topical collections.

# User Interfaces for Topic Management of Web Sites

## Table of Contents

<b><u>CHAPTER 1: INTRODUCTION</u></b>	<b>1</b>
<b><u>1.1 INTRODUCTION</u></b>	<b>1</b>
<b><u>1.2 MOTIVATION OF RESEARCH</u></b>	<b>1</b>
<b><u>1.3 OBJECTIVES OF RESEARCH</u></b>	<b>3</b>
<b><u>1.4 APPROACH TO RESEARCH</u></b>	<b>3</b>
<b><u>1.5 CONTRIBUTIONS OF RESEARCH</u></b>	<b>4</b>
<b><u>CHAPTER 2: RELATED WORK</u></b>	<b>6</b>
<b><u>2.1 FILTERING</u></b>	<b>6</b>
<b><u>2.1.1 COLLABORATIVE/SOCIAL FILTERING</u></b>	<b>6</b>
<b><u>2.2 STRUCTURE IN THE WEB</u></b>	<b>10</b>
<b><u>2.2.1 HYPERTEXT STRUCTURE</u></b>	<b>10</b>
<b><u>2.2.2 USING STRUCTURE IN TOOLS</u></b>	<b>12</b>
<b><u>2.2.3 WEB CRAWLING</u></b>	<b>15</b>
<b><u>2.3 WEB PAGE ARCHIVING</u></b>	<b>16</b>
<b><u>2.4 INFORMATION WORKSPACES</u></b>	<b>18</b>
<b><u>CHAPTER 3: PHOAKS SYSTEMS</u></b>	<b>20</b>
<b><u>3.1 INTRODUCTION</u></b>	<b>20</b>
<b><u>3.1.1 USENET NEWS</u></b>	<b>21</b>
<b><u>3.1.2 FREQUENCY OF MENTION IN PUBLIC CONVERSATION</u></b>	<b>22</b>
<b><u>3.1.3 CLASSIFICATION RULES: DEVELOPMENT &amp; ITERATIVE REFINEMENT</u></b>	<b>23</b>
<b><u>3.2 PHOAKS ARCHITECTURE</u></b>	<b>25</b>
<b><u>3.2.1 PHOAKS NEWS AGENT</u></b>	<b>25</b>
<b><u>3.2.1.1 Filtering</u></b>	<b>25</b>
<b><u>3.2.1.2 Categorization</u></b>	<b>26</b>
<b><u>3.2.1.3 Disposition</u></b>	<b>26</b>
<b><u>3.2.2 WEB INTERFACE</u></b>	<b>28</b>
<b><u>3.3 LESSONS LEARNED</u></b>	<b>28</b>
<b><u>CHAPTER 4: TOPICSHOP SYSTEMS</u></b>	<b>31</b>
<b><u>4.1 WEB CRAWLING</u></b>	<b>31</b>
<b><u>4.2 WEBCITE</u></b>	<b>36</b>
<b><u>4.2.1 LESSONS LEARNED</u></b>	<b>37</b>
<b><u>4.3 TOPICSHOP</u></b>	<b>38</b>
<b><u>4.4 CURRENT INTERNET RESOURCE DISCOVERY TECHNIQUES</u></b>	<b>40</b>
<b><u>4.4.1 COMPREHENSIVE INDICES (WEB DIRECTORIES)</u></b>	<b>41</b>
<b><u>4.4.2 KEYWORD SEARCHES</u></b>	<b>41</b>
<b><u>4.4.3 HYBRID DIRECTORY/KEYWORD SEARCHES</u></b>	<b>41</b>

4.4.4	<a href="#">SPECIALIZED INDICES</a>	42
4.4.5	<a href="#">SOCIALY FILTERED</a>	42
4.4.6	<a href="#">TOPICSHOP</a>	42
<b><a href="#">CHAPTER 5: OVERVIEW OF USER STUDIES</a></b>		<b>45</b>
<b>5.1</b>	<b><a href="#">HYPOTHESIS</a></b>	<b>45</b>
<b>5.2</b>	<b><a href="#">EXPERIMENTS</a></b>	<b>45</b>
5.2.1	<a href="#">SELECTING A DOMAIN</a>	46
5.2.2	<a href="#">INTRODUCTION TO PILOT STUDY</a>	47
5.2.3	<a href="#">INTRODUCTION TO INTERFACE EVALUATION</a>	48
<b><a href="#">CHAPTER 6: PILOT STUDY</a></b>		<b>49</b>
<b>6.1</b>	<b><a href="#">INTRODUCTION</a></b>	<b>49</b>
<b>6.2</b>	<b><a href="#">EXPERIMENTAL DESIGN</a></b>	<b>50</b>
<b>6.3</b>	<b><a href="#">PARTICIPANTS</a></b>	<b>52</b>
<b>6.4</b>	<b><a href="#">METHODOLOGY</a></b>	<b>52</b>
<b>6.5</b>	<b><a href="#">DATA COLLECTION AND ANALYSIS</a></b>	<b>53</b>
<b>6.6</b>	<b><a href="#">QUANTITATIVE RESULTS</a></b>	<b>53</b>
<b>6.7</b>	<b><a href="#">USER EXPLORATION STRATEGIES</a></b>	<b>59</b>
<b>6.8</b>	<b><a href="#">DESIGN IMPLICATIONS</a></b>	<b>61</b>
<b><a href="#">CHAPTER 7: USER INTERFACE EVALUATION</a></b>		<b>64</b>
<b>7.1</b>	<b><a href="#">LESSONS LEARNED</a></b>	<b>65</b>
<b>7.2</b>	<b><a href="#">TOPICSHOP DESIGN ITERATION</a></b>	<b>67</b>
<b>7.3</b>	<b><a href="#">EXPERIMENTAL DESIGN</a></b>	<b>69</b>
<b>7.4</b>	<b><a href="#">PARTICIPANTS</a></b>	<b>70</b>
<b>7.5</b>	<b><a href="#">METHODOLOGY</a></b>	<b>71</b>
<b>7.6</b>	<b><a href="#">DATA COLLECTION AND ANALYSIS</a></b>	<b>73</b>
7.6.1	<a href="#">PHASE ONE: USER STUDY</a>	74
7.6.2	<a href="#">PHASE TWO: EXPERT RATINGS</a>	74
<b>7.7</b>	<b><a href="#">QUANTITATIVE RESULTS</a></b>	<b>74</b>
7.7.1	<a href="#">EXPERT METRICS</a>	74
7.7.2	<a href="#">FINDING QUALITY SITES</a>	75
7.7.3	<a href="#">USER SEARCH EFFICIENCY</a>	78
7.7.4	<a href="#">REQUIRED EFFORT</a>	80
7.7.5	<a href="#">USER CATEGORIZATION</a>	81
7.7.6	<a href="#">RELATIONSHIP BETWEEN EVALUATION AND ORGANIZATION SUB-TASKS</a>	85
7.7.7	<a href="#">EXPERT RATINGS FOR SITE BREAKDOWNS</a>	87
7.7.8	<a href="#">COMPARING HUMAN PERFORMANCE TO AUTOMATIC METRICS</a>	88
7.7.9	<a href="#">QUESTIONNAIRE RESULTS</a>	89
7.7.10	<a href="#">QUALITATIVE OBSERVATIONS</a>	91
<b>7.8</b>	<b><a href="#">DESIGN SUMMARY</a></b>	<b>93</b>
7.8.1	<a href="#">SPATIAL ORGANIZATION IN WORK AREA</a>	93
<b><a href="#">CHAPTER 8: COMPARISON OF STUDIES</a></b>		<b>95</b>
<b>8.1</b>	<b><a href="#">RESULTS</a></b>	<b>95</b>
8.1.1	<a href="#">FINDING QUALITY SITES</a>	95

<u>8.1.2</u>	<u>USER EFFORT</u>	96
<u>8.1.3</u>	<u>QUESTIONNAIRE</u>	97
<b><u>CHAPTER 9: PREDICTING QUALITY SITES</u></b>		<b><u>100</u></b>
<b><u>9.1</u></b>	<b><u>EXPERIMENT</u></b>	<b>101</b>
<u>9.1.1</u>	<u>DATA</u>	102
<b><u>9.2</u></b>	<b><u>RESULTS</u></b>	<b>103</b>
<u>9.2.1</u>	<u>EXPERT AGREEMENT</u>	103
<u>9.2.2</u>	<u>LINK-BASED METRIC COMPARISON</u>	105
<u>9.2.3</u>	<u>PREDICTING QUALITY</u>	107
<u>9.2.4</u>	<u>DISCUSSION</u>	112
<b><u>CHAPTER 10: SUMMARY AND CONCLUSIONS</u></b>		<b><u>113</u></b>
<b><u>CHAPTER 11: REFERENCES</u></b>		<b><u>115</u></b>
<b><u>CHAPTER 12: CURRICULUM VITAE</u></b>		<b><u>122</u></b>

## Table of Figures

<a href="#">FIGURE 1.1: RESEARCH ROAD MAP</a> .....	4
<a href="#">FIGURE 3.1: PHOAKS WEB INTERFACE</a> .....	27
<a href="#">FIGURE 4.1: WEBCITE USER INTERFACE</a> .....	36
<a href="#">FIGURE 4.2: FIRST VERSION OF TOPICSHOP (DETAILS VIEW)</a> .....	38
<a href="#">FIGURE 4.3: FIRST VERSION OF TOPICSHOP (ICONS VIEW)</a> .....	40
<a href="#">FIGURE 6.1: SEARCH ENGINE USAGE</a> .....	50
<a href="#">FIGURE 6.2: WEB BROWSE HISTORY FROM USER PILOT STUDY</a> .....	58
<a href="#">FIGURE 7.1: REVISED VERSION OF TOPICSHOP, BASED ON RESULTS OF PILOT STUDY</a> .....	65
<a href="#">FIGURE 7.2: A SAMPLE SUBJECT'S CATEGORIZATION OF TORI AMOS SITES. (SUBJECT 3)</a> .....	84
<a href="#">FIGURE 7.3: A SECOND SUBJECT'S CATEGORIZATION OF TORI AMOS SITES. (SUBJECT 4)</a> .....	84
<a href="#">FIGURE 7.4: GROUPS FOR TORI AMOS AS CREATED BY SUBJECTS 3&amp;4</a> .....	85
<a href="#">FIGURE 7.5: TIMELINES OF USER ACTIVITY</a> .....	87
<a href="#">FIGURE 7.6: AUTOMATED METRICS COMPARED TO SUBJECTS' JUDGMENTS</a> .....	89
<a href="#">FIGURE 9.1: DATA FOR QUALITY EXPERIMENTS</a> .....	103

## Table of Tables

<a href="#"><u>TABLE 4.1: COMPARISON OF SEARCH INTERFACES</u></a> .....	43
<a href="#"><u>TABLE 6.1: PILOT STUDY EXPERIMENTAL DESIGN</u></a> .....	51
<a href="#"><u>TABLE 6.2: EXPERT INTERSECTION ANALYSIS</u></a> .....	54
<a href="#"><u>TABLE 6.3: EXPERT WEIGHTED UNION ANALYSIS</u></a> .....	55
<a href="#"><u>TABLE 6.4: AMOUNT OF WORK</u></a> .....	59
<a href="#"><u>TABLE 7.1: MAIN STUDY EXPERIMENTAL DESIGN</u></a> .....	70
<a href="#"><u>TABLE 7.2: NUMBER OF SITES IN EXPERT SETS</u></a> .....	73
<a href="#"><u>TABLE 7.3: AVERAGE EXPERT MAJORITY SCORES FOR TOPICSHOP AND YAHOO USERS</u></a> .....	76
<a href="#"><u>TABLE 7.4: MAJORITY SCORE FOR TOP 5/TOP 10 USER SITES</u></a> .....	77
<a href="#"><u>TABLE 7.5: INTERSECTION BETWEEN USERS SELECTIONS AND TOP 15 EXPERT-RATED SITES</u></a> .....	77
<a href="#"><u>TABLE 7.6: TASK TIME (IN MINUTES)</u></a> .....	78
<a href="#"><u>TABLE 7.7: TIME TO VISIT TOP 5 SITES</u></a> .....	79
<a href="#"><u>TABLE 7.8: PERCENTAGE OF TIME SPENT BROWSING/ORGANIZING</u></a> .....	79
<a href="#"><u>TABLE 7.9: AVERAGE NUMBER OF SITES BROWSED</u></a> .....	80
<a href="#"><u>TABLE 7.10: AVERAGE SITE INTERSECTION AMONG USERS</u></a> .....	81
<a href="#"><u>TABLE 7.11: PAIRWISE CATEGORY AGREEMENT BETWEEN USERS (1-4)</u></a> .....	83
<a href="#"><u>TABLE 7.12: DISTRIBUTION OF ORGANIZATIONAL ACTIONS ACROSS TIME QUARTILES</u></a> .....	86
<a href="#"><u>TABLE 7.13: EXPERT SCORES OF SITE CATEGORIES</u></a> .....	88
<a href="#"><u>TABLE 8.1: EXPERT INTERSECTION COMPARISON ACROSS STUDIES</u></a> .....	96
<a href="#"><u>TABLE 8.2: TASK TIME COMPARISON ACROSS STUDIES</u></a> .....	97
<a href="#"><u>TABLE 8.3: COMPARISON OF NUMBER OF SITES BROWSED</u></a> .....	97
<a href="#"><u>TABLE 8.4: USER CONFIDENCE FROM QUESTIONNAIRE</u></a> .....	98
<a href="#"><u>TABLE 8.5: SITE PARAMETER RANKINGS</u></a> .....	98
<a href="#"><u>TABLE 9.1: EXPERT AGREEMENT USING CORRELATIONS</u></a> .....	104
<a href="#"><u>TABLE 9.2: EXPERT AGREEMENT, USING CATEGORIES</u></a> .....	104
<a href="#"><u>TABLE 9.3: METRIC SIMILARITY</u></a> .....	105
<a href="#"><u>TABLE 9.4: METRIC SIMILARITY, INTERSECTION OF TOP 5 AND 10</u></a> .....	106
<a href="#"><u>TABLE 9.5: LINEAR MODEL FOR PREDICTING EXPERT AVERAGE</u></a> .....	107
<a href="#"><u>TABLE 9.6: NUMBER AND PROPORTION OF GOOD</u></a> .....	108
<a href="#"><u>TABLE 9.7: PRECISION AT 5 AND 10</u></a> .....	109
<a href="#"><u>TABLE 9.8: MAJORITY SCORE AT 5 AND 10</u></a> .....	110
<a href="#"><u>TABLE 9.9: AVERAGE EXPERT SCORES OF TOP 10 SITES</u></a> .....	111

# CHAPTER 1: INTRODUCTION

## 1.1 INTRODUCTION

Web search and navigation are difficult problems that have received much attention, with search engines like AltaVista and directories like Yahoo being the most widespread solution attempts. However, users have information needs and interests that are larger in scope and longer in duration than can be satisfied by AltaVista and Yahoo. In particular users want to manage their persistent interests in broad topics and to comprehend collections of web documents relating to topics.

## 1.2 MOTIVATION OF RESEARCH

Typical search solutions are content-based, where a user query is filled by matching keywords to the text of web pages. While this approach works in many situations, it fails when users want to find quality information on a topic and manage the resulting information over a period of time. By utilizing the inherent structure found on the World Wide Web, we may gain more insight into the perceived quality of a web site. By viewing links to web pages as endorsements (a site linking to a page might validate that it contains quality content), we can use the concepts of social filtering (utilizing user preference for prediction) to create better collections of topically coherent web sites. Social filtering is a method of filtering objects (documents, videos, web pages, etc.) that concentrates on the characteristics of people and their preferences in addition to the objects' content. The focus of social filtering is shifted from strictly

assessing the content of objects to evaluating the personal and organizational relationships of the community of users accessing those objects.

An important task that many web users perform is gathering, evaluating, and organizing relevant information resources for a given topic; we call this *topic management*. 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, one might gather a collection of web sites on wireless telephony as part of a report for work or a collection on the X-Files as a service for fellow fans. Librarians might prepare topical collections for their clients, and teachers for their students.

Topic management is a difficult task that is not supported well by current web tools. A common way to find an initial set of (potentially) relevant resources is to use a search engine like AltaVista or an index like Yahoo. At this point, however, a user's work has just begun: the initial set usually is quite large, consisting of dozens to hundreds of sites of varying quality and relevance, covering assorted aspects of the topic. Users typically want to select a manageable number – say 10 to 20 – of high-quality sites that cover the topic. With existing tools, users simply have to browse and view resources 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, probably the most common form of keeping track of web sites, are a fairly primitive organizational technique.

While many web search utilities provide answers to specific queries, they do not provide convenient, efficient methods for exploring the body of knowledge available about a topic. Some search resources allow users to find a category that closely matches the topic they are interested in, but the end result is simply an alphabetical list of web sites that contain information on the given topic. New techniques that provide additional functionality need to be available on the web to support broader types of information gathering.

Most research done on search engines (See Related Work, Section 2), has concentrated on tweaking search algorithms to give very small gains in relevance ranking of results with respect to the user's query. While improving the result relevancy is still important, the small gains attained are out of proportion with the amount of work that must be done by the user. Even after these gains are realized,

there still remains the problem of what to do with the ranked list of information. With better user interfaces and visualization methods for presenting results, we may help users find information more efficiently and effectively. We created and continue to develop the *TopicShop* system (discussed in Section 3.2.2) to address this need. TopicShop includes (1) a web crawler that discovers relevant web sites and builds site profiles, and (2) information workspaces for exploring and organizing sites.

### **1.3 OBJECTIVES OF RESEARCH**

This research consisted of multiple initial goals. First, we wanted to gain a better understanding of the task of topic management and the methods that people use to complete this task, while showing that the task has limited existing support on the web. Also, we wanted to evolve the interface designs in TopicShop to be more efficient and provide users better access to the data necessary for topic management. Finally, using two controlled empirical studies, we have validated that these interfaces enable users to perform the management task effectively and demonstrate their usefulness for people maintaining persistent collections of web sites (such as links page maintainers) by enabling them to easily use the TopicShop system for their own web sites.

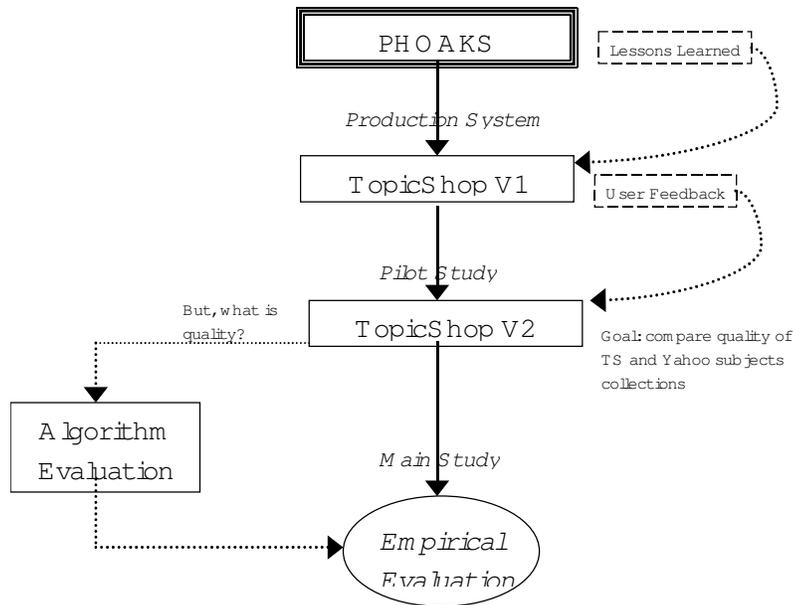
### **1.4 APPROACH TO RESEARCH**

We conducted two between-subjects empirical comparisons of TopicShop and Yahoo with users performing the task of topic management. We have also investigated the effectiveness of these two user interfaces in helping to support users' needs of managing persistent topic collections of web sites. Yahoo is a popular Internet tool that is currently used for the task of topic management and combined with bookmarks serves as a good comparison for TopicShop.

The users in both of our studies were presented data on a topic in either the Yahoo interface in its original form or a topic crawl using Yahoo sites as seed sites in our TopicShop interface. They were asked to utilize the interface provided to them and evaluate sites in the collection from the topic crawl, selecting sites they thought were the highest quality (they provide a good overview of the topic). By measuring their performance and soliciting subjective feedback, we have been able to gain an insight into the benefits of each interface concept and incorporate changes into later iterations of TopicShop that improve its

usefulness. In addition, topic experts have evaluated these same sites and we have used these expert quality judgements of sites to compare to each user's collection of sites in order to rate the quality of their collections. We have also performed numerous analyses of the notion of quality and how it can be predicted through automated measures. These studies and analyses are detailed in Chapters 4 through 10.

## Research Map



**Figure 1.1: Research Road Map**

When designing TopicShop, we kept a number of goals in mind, including: making relevant but invisible information visible, including a rich representation of the desired web sites, making it simple for users to explore and organize resources, and integrating topic management into a user's normal computing and communications environment. By following these guidelines and iterating the design based on user feedback from empirical studies, we have improved the TopicShop user interface to meet our goals and the needs of users.

### 1.5 CONTRIBUTIONS OF RESEARCH

Contributions of this research include:

- Java applet-based web crawler to efficiently gather relevant web sites about a topic using the hypertext structure of the web and return results with detailed site profiles in response to a user's topical query.

- TopicShop visualization and management user interface, which this work showed to be more effective for displaying results and managing topic collections, developed by thoroughly analyzing how users perform the task of topic management.
- Empirical evidence regarding the effectiveness of a task-specific user interface for topic management over current search engine technology through a controlled empirical study comparing topic management interfaces. We have shown that TopicShop subjects find more high quality sites, while doing less work, in less time. In addition we have shown that simple features of web sites can be used to predict their quality.

## **CHAPTER 2: RELATED WORK**

### **2.1 FILTERING**

Information filtering is a technique that uses past user data to make recommendations about something a user will want in the future. A text retrieval system can log what a user has searched for in the past and make suggestions of other documents they might be interested in, based on their past search queries. Most information filtering to date has been content-based, but there may be better methods of filtering.

#### **2.1.1 Collaborative/Social Filtering**

Social filtering is a type of information filtering where, instead of filtering on document content, systems filter on similarity of user preferences. By matching a user to other similar users, a system can suggest potential documents or items that similar users have commented on in the past. One of the earliest investigations into personal ratings for HCI-type user-modeling by Allen [4] had unencouraging results. But later attempts have built on this early work and show some very successful results. Malone et al [61] describe three types of information filtering: cognitive filtering, economic filtering, and social filtering. In an email filtering system, cognitive filtering, often referred to as content-based filtering, involves matching messages to receivers based on the actual content of the message, and economic filtering considers the estimated search cost and benefit of use to the user before suggesting a document.

An example of a content-based filtering system is Rhodes and Starner's Remembrance Agent [83], an automated information retrieval system that watches what a user is currently typing and then scans old email messages, notes, and online documents for something relevant to the user's current interest. This way, the user is not required to do anything within the system; it simply watches in the background and interrupts when relevant information is found. There are of course settings that allow users to set the frequency with which document suggestions are made. The suggestions were limited to one "nugget" of information so that they would not interfere too much with the user's current work. Because this system relied solely on text matching, it created many false positives but since they could be easily ignored, this was not a major problem.

The Information Lens [60] is an information sharing system that uses social filtering to filter email. It automatically filters by matching messages using user defined rules and performing a specified action on the messages. In Information Lens, users fill out a template of specified fields like time, topic, and meeting place for each message they send and then can write rules to automatically filter their incoming messages based on these fields.

INFOSCOPE [30][91] is a system that filters Usenet (explained in Section 3) news articles and can be thought of as an extension to Information Lens. This system works at recategorizing newsgroup messages into virtual newsgroups by matching user profiles or following user-defined rules. A virtual newsgroup is a logical entity containing articles from multiple newsgroups that match some series of patterns specified by the user. Agents monitor user behavior behind the scenes and automatically suggest new virtual newsgroups that might be useful. This system is strictly based on individual users and the information they are dealing with. There is no collaboration between the users of the system.

Collaborative filtering recommendation systems match user profiles and suggest items that similar users recommend. The first of these systems was the Video Recommender System by Hill, Stead, Rosenstein, and Furnas [46]. The interface to this system was through email and allowed users to send in their ratings of movies they have seen and then receive back a list of potential additional movies they might like to see. This was based on the idea of having a virtual community of user preference to match against a user's likes and dislikes to find similar users in order to make recommendations. By giving the system an

idea of what types of things a user likes, the system can compare to other users and locate additional movies that the user will probably also like.

Another system that supports this same type of recommendation through user profiles was developed by Shardanand [87]. The system, called Ringo [88], a precursor of Firefly, is a social information filtering system that makes personalized music recommendations. Users could indicate their listening preferences by assigning specific ratings to music. This profile they generated could then be compared to other users' profiles to determine which users are similar in their musical tastes. Then, recommendations could be made from the combined list of albums that similar users liked.

Resnick et al [82] designed and implemented another social filtering architecture based on personal ratings and demonstrated its application to filtering net news. In this system, called GroupLens, users rate articles on a numerical scale and the system correlated the existing user profiles to predict which article a user will be interested in. This system was successfully field tested with about 200 users. Filtering the net news articles could result in one of four cases: hit, miss, false positive, or correct rejection [54]. A hit and correct rejection are both desirable outcomes in a filtering system. A false positive will simply add noise to the results. Most times, human users can easily detect false positives that the system could not. Missing a relevant document, though, can be dangerous if there was necessary information contained in the document and most filtering systems attempt to limit the number of missed documents.

Another filtering system that works on Usenet news is called URN [14][15]. This collaborative Usenet interface allows users to vote on articles and provide keywords associated with those articles. These votes and weights are updated in the system and collated across all users. Now the system can determine which articles a user might want to see based on past voting and display them by popularity.

Tapestry [39][93] is a site-oriented email system that allows the entry of text annotations that can be used later to filter messages for other users. Annotations are typically rich in high quality information so adding annotations to messages provides useful data and insights about the message content for future reference. Users can search through this email repository by developing SQL like queries of the text and annotations of the messages. Both querying the system and annotating messages requires significant user effort. There is a tradeoff between the quality of data that are collected and the user effort required. This is

true of many systems. If a user must put forth a large amount of work for little benefit they are usually less inclined to use the system [40].

We have seen systems that require different types of work from the users. Some of these systems simply require a user vote, while others require users to write and attach full textual annotations to the messages in the system. The effort required to annotate documents well is too high. Instead of using high quality recommendations from a few people, it can be much more useful to have a large number of lower quality recommendations [28]. By gathering information from many diverse users, we can better predict quality documents.

The system developed by Maltz and Ehrlich [62] supported both active and passive filtering. Active filtering enables users to create explicit recommendations and then send them to specific colleagues. This way, users receive recommendations from someone they know personally. The passive filtering aspect of the system is where users can annotate documents they feel someone might be interested in and then not send them to anyone in particular, but leave them in the system for others to discover. Then, when a user happens to be reading a particular document they can see any comments made by other users.

Answer Garden 2 [2], a slightly different type of social filtering system, is an organizational memory system that provides collaborative help. By capturing the questions and answers between employees and support staff in an organization, a huge repository of information can be built to assist other employees. When an employee has a question, they can ask the Answer Garden system and if an answer is not contained in the repository it is sent to an escalation agent. This agent goes through a series of steps, each more intrusive, to attempt to get an answer to the question. First a chat room is consulted, then a newsgroup, and finally, if no answer is found, a specific expert is contacted to answer the question. After this, the answer is added to the repository and anybody asking a similar question can be given an answer immediately.

We have already seen systems that filter net news and email archives. Another potential source of rich information is bookmark lists. Siteminer [81] is a system that mines personal bookmark lists. Since bookmarks are an implicit declaration of interest in the bookmarked page's content a count of the number of times a site appears in users' bookmark lists can be seen as a quality ranking for that site. Since

individual users tend to group their bookmarks into folders, the system also attempts to gain more information about a site by observing the site groupings between multiple subjects.

## **2.2 STRUCTURE IN THE WEB**

The World Wide Web is a collection of linked hypertext documents. The underlying structure between web pages can be thought of as a directed graph. Pages are represented by nodes and links between the pages are the edges. A basic intuition derived from this structure is that links often represent an endorsement of the quality and relevance of the linked-to site. Thus, this can be considered a form of social filtering. There are some useful graph properties that can now be considered. The out degree of a node represents the number of links going from a particular node to any other node in the graph. The in degree of a node corresponds to the number of nodes pointing in at a particular node.

### **2.2.1 Hypertext Structure**

Many researchers have already done work in analyzing network structure. These same ideas can be applied to the web [48] in most cases. Network analysis has significant potential to generate insight into the communicative nature of web structures. The structure of the web can be an important component to investigate when dealing with web sites. In fact, links to another web page can be considered an endorsement of that web page.

Citation links are another area where the structure has been investigated thoroughly. Butterfly [59] is a system that accesses DIALOG's science citation databases and performs a purely structural analysis to support users in managing collections of information resources. The central UI object is a butterfly, which represents one article, its references, and its citers. By utilizing the citation structure, the interface makes it easy for users to browse from one article to a related one, group articles, and generate queries to retrieve articles that stand in particular relationship to the current article.

There are two important issues to consider when building a hypertext structure: navigation and viewing. When building the structure it is important to keep in mind how easy it will be for users to navigate and view the information contained within the structure. Furnas [31][33] studied the requirements necessary for building effectively view-navigable structures. He found that the out degree should be small with respect to the overall size of the structure. This means that nodes in the hypertext graph should not

point to every other node but rather a small subset of nodes. In addition the distance between nodes should be kept small with respect to the structure. For related nodes there should be a short path to get from one to the other, rather than having to traverse the entire graph. In the web, we have no control over this structure, but luckily, so far, it meets both of these criteria.

Botafogo et al [12] developed a number of algorithms for analyzing arbitrary networks, splitting them into structures (pre-trees, hierarchies) that are easier to visualize and navigate. These aggregate structures are inferred based on identifying articulation points in the undirected graph and removing them to create a set of subgraphs. An articulation point is a point such that removing it and its edges from the graph would disconnect the graph into two or more components. This algorithm removes indices (nodes with high out-links) and references (nodes with high in-links) because these nodes cause over-connection in the graph and in order to have good articulation points, the graph must not be highly connected. We will see later that these two types of nodes (indices and references) are an important part of the web structure.

There has been some work done to categorize pages in a hyperlink structure by Pirolli et al [75]. Their categorization algorithm uses hyperlink structure, text similarity, and user access data to categorize web pages into various functional roles such as head, index, and content. These functional roles are used to extract structures from the web determined using the spreading activation based on a set of user provided seed pages. Their system follows the links from the seed pages and allows the structure to continually grow as more pages are evaluated. A head page is the front page of a site. Nodes that have a high in degree and thus many links pointing at the page are considered content pages. Finally index pages contain a large number of links to other pages, most times content pages. These algorithms were tested on the Xerox PARC web site and were shown to categorize the pages with very good accuracy.

The web is very large and it is not possible to ensure that everything is structured well. Order can be imposed at a local level but global organization is unplanned [37]. The high level structure of the web emerges only after later analysis. Another analysis of web structure that shows the breakdown of two types of web pages was developed by Kleinberg et al [52][53] in an effort to gain information about web sites to aid user comprehension of sites. This algorithm is called HITS – Hyperlink Induced Topic Search. The two main types of pages are hub and authoritative pages. These two terms are mutually dependent: a good hub is one that links to many authorities and a good authority is linked to by many hubs. Authorities and

hubs, when isolated in a graph, should form dense bipartite communities. That is one set of pages (hubs) will point to the other set of pages (authorities). This analysis uses co-citation to categorize the pages in the structure by clustering pairs of documents based on the number of times they were both cited by a third document.

Several researchers have extended this basic algorithm. Chakrabarti et al [23][24] weight links based on the similarity of the text that surrounded the hyperlink in the source document to the query that defined the topic. Bharat & Henzinger [10] made several important extensions. First, they weighted documents based on their similarity to the query topic. Second, they count only links between documents from different *hosts*, and average the contribution of links from any given host to a specific document. That is, if there are  $k$  link 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 [74] 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.

Another project has concentrated on new techniques for inducing clusters of related documents on the web. Pitkow and Pirolli [78] describe algorithms that find lawful properties of document behavior and use. These methods again start with co-citation analysis but then use a desirability ranking of pages to improve clusters.

Two concepts that can be seen in much of the above work on hypertext structure are pages that are heavily linked to and pages that point to many other pages. The first of these have a high number of in-links and have been named content, reference, and authoritative pages. The second type have a high number of out-links and are called index and hub pages.

### **2.2.2 Using Structure in Tools**

After a thorough analysis of the structure, the next step is to use it to help users find the information they are seeking. By incorporating structure analysis into tools aimed at finding web pages and collections, we can improve the efficiency of searching on the web.

A number of researchers have created interfaces to support users in managing collections of information resources. SenseMaker [8] focuses on supporting users in the contextual evolution of their interest in a topic. They attempt to make it easy to evolve a collection, e.g., expanding it by query-by-example operations or limiting it by applying a filter. In addition Mukherjea et al. [71] designed algorithms for analyzing arbitrary networks, splitting them into substructures that are easier for users to visualize and navigate.

Pirolli and Card [77] define the term information foraging to cover activities associated with assessing, seeking, and handling information sources. The main idea gained from this metaphor is that systems need to adapt their designs in the context of the information they are seeking and the tasks that will be performed with the information. Depending on the task at hand, systems need to adapt to different information foraging strategies. In today's information rich world, the design problem is no longer how to collect more information, but how to optimize a user's time and increase the relevant information gained. Again, this goes back to the tradeoff in the value of information obtained against the cost of performing the search activity [40].

Scatter/Gather [76] browsing is a cluster based browsing technique for large text collections based on information foraging theory. The interface presents summaries of clusters of similar documents, allowing the user to navigate through the topic structure. The concept of gathering is to select individual clusters that are of interest. Scattering is the process of reclustering the selected clusters to reveal more fine-grained clusters of documents. This type of interface supports browsing of a collection of documents rather than searching the collection and is aimed at satisfying the user need to learn about the collection in general before looking for specific documents. This method gives users a chance to iteratively reveal the topic structure of the collection and eventually locate desired documents.

Two additional systems developed in support of information foraging theory are the WebBook and WebForager [20]. The WebBook uses a book metaphor to group a collection of related web pages into a compact unit for viewing, storing, and additional interaction. The WebForager lets users view and manage multiple WebBooks on their desktop. The collections of web pages required to make up a WebBook can be generated using a set of automated methods provided by the system. Typical methods of building collections include: following all links from a page one level, following relative links from a web page,

extracting book like structures by following previous and next links, and grouping pages returned from a search query.

Another browsing method for viewing pages on the web was developed using multiple hierarchical windows. Kandogan et al [50] have shown that through the extensive use of single user operations on multiple windows, their elastic windows browser provides an efficient overview and sense of current location in information structures. Their interface facilitates the organization and filtering of information and aids users in accessing previously visited pages without high cognitive demands. As users goals change, they can quickly organize, filter, and restructure the pages on the screen using this browser.

The Navigational View Builder [70] is a tool to effectively build overview diagrams of the hypertext structure behind the web. It uses binding, clustering, filtering, and hierarchization to accomplish this task. Binding is done first to bind the information attributes to the visual attributes of the nodes and links in the structure. Clustering is done to provide abstracted views to show the overall information space on a single screen by analyzing the structure and the content of pages. Filtering reduces the amount of information on the screen by specifying relationships in the links or specific content to filter out. Hierarchization is done on the resulting set of pages by inferring the hierarchy from the content and underlying structure. While this system attempts to do all this work automatically, the authors admit that they had to manually enter many of the useful semantic attributes that were not able to be extracted automatically. One of these attributes was the page topic which we will show later can be semi-automatically generated.

The web is unlike traditional hypertext systems in that it is both redundant and incomplete. In the web, when there is no link between two pages, that does not mean that they are unrelated; it simply means they have yet to be linked. The web also contains many pages containing the same information. In traditional hypertext systems, this would not be true. Spertus [88] states that content search alone is lacking and because of the untraditional nature of the web, new techniques are necessary. ParaSite is a system that analyzes the links between web pages to find additional pages that are related to a given set of pages and to infer the topic and function of the pages seen along the way.

Google [16] is another system that crawls and indexes the web making use of the structure to provide more satisfying search results. Each page encountered in a crawl in this system is assigned a page

rank that consists of the inlinks and outlinks, and the similarity of anchor text and page text. These results are given back in response to a search query.

Improving search results is the goal of WebQuery [22] which builds a graph of links and nodes from an initial search engine result set and extends it, assigning the highest rank to the most highly connected nodes. This system is unique because it allows users to visualize the results in a number of ways: cone trees, 2D graphs, 3D graphs, lists, and bullseyes. For large sets of web pages, cone trees provide the best view because they make excellent use of screen real estate. A 3D graph is the best view when there are less nodes with similar connectivity. The bullseye view helps to draw attention to the most highly ranked node and allows nodes to be selected bringing them to the front and displaying their relationships. These layout possibilities all serve different information seeking needs.

The structure within a single web site can also improve navigation for users trying to view large web sites. MAPA [29] is a system for inducing and visualizing the hierarchical structure within a web site. It extracts the structure and builds an interactive map of the site to use for navigation. A walker gathers information about individual pages within the site and then organizes the total link topology to make the interactive map visualizations.

Lamping et al. [55] explored hyperbolic tree visualization of information structures. Another system that aids in mapping a single web site is WebCutter [58]. This system has tightly integrated search and browse oriented information discovery tools that interactively crawl through a site to generate visualizations. Like other systems described above, WebCutter allows the user to explore the information using a few different visualizations: a tree control useful, for abstraction; ellipsis, or star like layout, for pursuing incremental exploration; and a fisheye view for focusing on different regions of the graph.

### **2.2.3 Web Crawling**

A more specific application of hypertext structure is web crawlers. By crawling through the structure of web pages, we can collect information about pages and build a graph of page links.

One way to improve the efficiency of methodically crawling through a set of web pages is to dynamically vary the order in which pages are visited. There is an optimal order a crawler should visit URLs (Uniform Resource Locator) in order to obtain more important pages first. Also, since the web is

very large and not all URLs can be visited in a reasonable amount of time, we want to visit pages that add the most to a crawl first. Cho et al [26] investigated metrics to use when updating the URL ordering for a crawl. Some metrics they used are: query similarity, backlink count (same as inlinks), page rank, and a location metric. The page rank is similar to an inlink count except it is weighted to consider the inlink count of each page pointing in to the site. The location metric attempts to categorize sites by looking at the URL itself to determine if a site is a homepage, a commercial site, or a number of other page types. Of these three metrics, page rank works best.

A user study was conducted on the ARC system (Automatic Resource Compiler) [23][24]. A list of authoritative web sites on a topic was compiled using this automated system by performing three tasks: search and growth; weighting; and iteration and reporting. The search and growth phase followed links one level from each node to grow the set of seed sites into an extended set. Then each page is weighted and the process is repeated. Finally the results are reported in a sorted list. The study compared the results of the ARC system with Yahoo and Infoseek. The lists were presented to the users in their original form including the title of the search engine that generated them. Users gave subjective rankings of the three lists. The results showed that ARC was able to produce a list that was almost competitive with Yahoo and Infoseek's lists and occasionally produced a slightly better list. We will show later that an enhanced crawling method similar to this, along with an efficient interface tailored to this task can actually produce significantly better results.

Miller and Bharat [65] developed a framework for site-specific web crawlers. SPHINX is a Java toolkit and interactive development environment to support users in creating maps of a single site. Users can customize the crawls by using classifiers that analyze the content of the site's pages and categorize them specific to the particular topic.

### **2.3 WEB PAGE ARCHIVING**

As users browse information on the web, they need to keep track of quality sites they have seen that may be useful to them in the future. Most browsers have some type of archiving capabilities, called bookmarks and favorites in some current popular browsers. These are usually very primitive, often consisting simply of a method to mark pages of interest for later retrieval and the ability to group the pages

in folders. As this list of marked pages gets large, it becomes difficult to handle. Users who browse the web often need better methods of archiving the best information found in their web sessions.

Abrams, Baecker, and Chignell [1] carried out a study of several hundred web users who used bookmarks. Bookmarks are a very popular way to create personal information spaces of web resources. 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 four design recommendations to help users manage their bookmarks more effectively. First, bookmarks must be easy to organize, e.g., 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. These four design goals are important to consider when creating user interfaces such as TopicShop.

One of the first systems to concentrate on bookmarks was the Group Asynchronous Browsing system [102]. It is a collaborative system that merges web sites from multiple personal bookmark lists and even different web-based topic directories. Based on the concept of a multitree, general bookmark lists that may be further categorized into folders are combined the server. Users can then query the server to specify a subset of trees from the large multitree database. The system generates an HTML document listing the web sites matching the query and any cross-reference linking to other related sites and bookmark files.

WebTagger [51] is a personal bookmarking service that provides individuals and groups with a customizable means of organizing and accessing web-based information resources. This system is a collaborative bookmarking system based on some of the ideas in the Group Asynchronous Browsing system. By sending bookmarks to the system during a browsing session, users can later retrieve bookmarks from the large repository by querying the system. The returned results list shows categories of the web page and allows the user to rate the results so the system can retrieve better quality sites corresponding to the feedback the user has given.

Automating bookmarking by keeping a history is another method that can be employed to help users with this task. Takano's Dynamic bookmark tool [92] is used to support revisiting past web pages. The system automatically watches and archives a user's navigation behavior and shows the analyzed results

as clues for which URLs to revisit. Not only will this system allow users to find past sites they are interested in, it will also support users in finding URLs that they have visited before but did not realize were important enough to explicitly add to their bookmark list.

Bookmarks are typically gathered opportunistically, as users happen to encounter interesting sites, and bookmark files usually span many different topics. We are more interested in situations where users are explicitly engaged in gathering and organizing a collection of related resources for a specific topic. Our systems will attempt to support users in performing this activity.

The Data Mountain of Robertson et al [84] represents documents as thumbnail images in a 3D virtual space. Users can move and group the images freely, with various interesting visual and audio 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.

Hightower et al [42] based their work on the observation that users often return to previously visited pages. They used Pad++ [9] to implement PadPrints, browser companion software that presents a zoomable interface to a user's browsing history. Interfaces to browsing history reduce the need for users to create collections of items explicitly, although the problems of organizing a collection are the same, however it is obtained.

## **2.4 INFORMATION WORKSPACES**

After evaluating items and selecting the interesting ones, users must organize the items for future use. Card, Robertson, and Mackinlay [19] 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 2D space. Mander, Salomon, and Wong [64] 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 on an experiment that investigates this issue.

Marshall & Shipman's VIKI system [66] lets user organize collections of items by arranging them in 2D space. Hierarchical collections are supported. Later extensions [89] added automatic visual layouts, specifically non-linear layouts such as fisheye views [34].

## CHAPTER 3: PHOAKS SYSTEMS

### 3.1 INTRODUCTION

Usenet (User's Network) news is full of pointers to useful resources but because of its immense size it is not always easy to find the best and most reliable ones, without manually sifting through many non-relevant messages. We have developed the PHOAKS (People Helping One Another Know Stuff) system at AT&T Labs to address the problem of constructing collections of web pages by scouring Usenet news and keeping a database of all web pages that have been mentioned in its everyday conversations. The basic premise of PHOAKS is that an effective way to find good information resources (web sites) about a given topic is to ask experts in that topic. Since users of Usenet newsgroups are already carrying on discussions about thousands of topics, there is a large body of information available to find recommendations for quality resources (web sites, downloadable files, etc.) available on the Internet, without requiring any additional work from the users.

The typical user searching for information could, as one of many search methods, read through a newsgroup and look for relevant resources. But newsgroups have enormous amounts of traffic and this could be a time-consuming task. Some of the more active newsgroups have thousands of posted messages per day. Most people do not want to sift through that many messages to find what they are looking for. An agent such as PHOAKS eliminates much of this work by automatically sifting out resources from all the messages posted and presenting them to the user.

### 3.1.1 Usenet News

Usenet is a large distributed depository for message exchange among interested users on the Internet. It can be thought of as a global Internet bulletin board. It is subdivided into many topic areas and users posting messages decide where their message fits in best. The network topography of Usenet is distributed over many servers around the world. A local user posts a message to their local server and from there the message propagates out and eventually reaches all other news servers in the world. Reading messages is also done from a local server that has received messages from other servers. Due to the distributed nature of Usenet, there can be a lag in when a message is posted and when different servers around the world receive the messages. This may even lead to some machines receiving a reply to a message before the original message itself because of the paths the messages followed within the distributed network.

There are over 23,000 newsgroups (taken from logs of innd, a common Usenet news server) in Usenet news. These are the topics into which all messages are divided. The structure of the subdivided topics is a very large hierarchy. At the top level there are about 600 broad categories and each level deeper into the hierarchy leads to more specific topics. There are eight major top-level topic hierarchies comprising approximately 6900 groups:

- alt (alternative) [4641 groups, ~21%]: Almost any topic can appear in this hierarchy.
- comp (computers) [903 groups, ~4%]: Related to computer hardware and software.
- misc (miscellaneous) [135 groups, ~.61%]: Themes not easily classified into the other hierarchies.
- news [30 groups, ~.14%]: Concerned with the news software, network, and administration
- rec (recreation) [708 groups, ~3%]: Consists of discussions oriented toward hobbies and recreational activities.
- sci (science) [205 groups, ~.93%]: Discussions in research or applications of science.
- soc (society) [264 groups, ~1.2%]: Topics relating to social issues and world cultures.
- talk [29 groups, ~.13%]: Geared toward debates. Topics are usually very open ended.

Some example groups are rec.boats, alt.sports.hockey.nhl.ny-rangers, and comp.lang.java.

There are other top-level topics; many of them are foreign topic hierarchies for other countries. The alt tree is different from the other trees in the way groups are added to it. In most newsgroup hierarchies, users can make requests for new groups; they are voted on and eventually approved by an administrator and added. But in the alt tree, anyone can simply add a group. This leads to a very wide variety of topics that might not make it to one of the other hierarchies. Due to the ease of group additions, the alt tree typically will have discussions of big news stories, sometimes minutes after the events have occurred.

Social filtering can help determine which web sites mentioned in the messages are most important in the topic for which they are posted. By systematically counting the number of times a web site is mentioned within a newsgroup, we can gather a list of the most talked-about sites for each newsgroup. This list can be used to rank the sites and show a user which sites were most highly recommended by the community of users participating in the newsgroup discussion. PHOAKS was developed to implement this idea by constantly monitoring newsgroups and storing in a database all web resources mentioned in the discussions.

### **3.1.2 Frequency of Mention in Public Conversation**

The metric that PHOAKS used to determine which web sites are the most popular within a newsgroup is *frequency of mention*. The social data provided in Usenet news in the form of messages posted by users can be used to determine what URL mentions the users of each newsgroup have referenced most often. Counting one vote per distinct person posting a message with the URL mention provides a frequency count for URL mentions. This prevents users from posting multiple times about a site to try to manipulate the system and cause their favorite page to move higher on the recommended resource list for a newsgroup. Currently, PHOAKS requires a threshold of only one vote to make it into the frequency page, which is a list of the top forty resources for a newsgroup ranked by frequency of mention. A better way to do this might be to accept a resource only when at least two distinct people have recommended it. This will help eliminate the spam and automatic posts that are found throughout Usenet. The main presentation page of PHOAKS shows frequency counts for web resources gathered from newsgroups.

Another order of presentation is *recency of mention*. Using this ordering PHOAKS presents web resources that users are currently talking about in a newsgroup. The recency view of PHOAKS lists all resources recommended in the most recent posts that PHOAKS has come across, in descending order by date and time of the post.

A combination of recency and frequency is also available in PHOAKS. This allows a moving time window that causes only resources that were recommended somewhere within a specified time period to be included in presentation of the top recommended resources list. Since PHOAKS started running in October 1997, many pages have built up a large number of recommendations. This makes it more difficult for newer resources to reach the top of the frequency ordered list of pages in high traffic newsgroups. A moving time window allows a user to specify that they would rather see more current recommendations and information. Since web sites for many topics are rapidly changing, the best pages may be ones that have a fair number of recent recommendations instead of a large number of old recommendations. Of course, some pages that remain around for a long time are still the best source of information available, but that means that they will probably be recommended continually.

### **3.1.3 Classification Rules: Development & Iterative Refinement**

Each URL mention in a net news post is classified by PHOAKS into one of a number of categories, by applying a set of classification rules. By manually reading through a few thousand posts, we generated an initial representative set of categories:

- Private – mentions in messages with the private header field set to true
- PHOAKS URL – any mention of a PHOAKS web page
- Spam – URLs mentioned in more than 40 newsgroups in the same message
- Kill – mentions of an URL on a list of system definable undesirable sites
- Quoted – mention was inside a quoted area of text
- Code – mention was part of a source code sample
- Signature – mention was part of a user's signature
- Organization Signature – mention was part of a user's signature and was of the posting organization

- URL in Signature – mention was part of a user’s signature and contained the user’s email name
- Approved FAQ – mention within an approved FAQ
- Unapproved FAQ – mention within an unapproved FAQ
- URL in FAQ – mention in message that appears to be a FAQ but is not specifically tagged as one
- Self Recommendation – mention is a recommendation of the user’s own site
- Recommendation List – mention is a recommendation within a list of two or more
- Recommendation – mention is a recommendation
- Other

PHOAKS uses categorization to determine which URLs will be counted as a recommendation. Currently the categories used in frequency calculations are recommendation, Approved/Unapproved FAQ, and URL in FAQ. Each of these categories consists of times when the message poster is recommending an URL. For example, in a list of frequently asked questions, the URLs found there are usually answers to specific questions and can be considered recommendations.

The rule set that PHOAKS uses to determine an URL’s category has gone through three iterations. An initial set of rules was created using rule learning software called RIPPER [27]. This software takes as input a set of text samples along with a list of features and a result to conclude about each sample (in our case, categories). RIPPER then analyzes the features and develops a boolean combination of features that best predicts the desired result for each sample. Two independent raters manually classified a set of 200 URL mentions. Then, ten percent of these URLs were used to initialize RIPPER, and the rest used by RIPPER to learn the rule set that produced similar results to the examples given. Once these initial rules were created, they were applied to another independent set of URLs randomly sampled from PHOAKS. These URLs were also manually classified and results were compared to the automated classification, leading to a more refined set of rules. Finally, one more iteration was performed and after the rules were able to predict the URL mentions accurately within about 85%, they were used in the PHOAKS system.

There are two aspects of rule accuracy: *precision* (percentage of URLs that rules classify into a certain category that actually belong in that category) and *recall* (percentage of URLs that belong to a category that rules classify into that category). For the current set of rules in PHOAKS, precision is 88% and recall is 87%, with an inter-rater reliability of 88% as determined by applying the rules to a sample set

of URLs and comparing to human categorization performed by two independent raters. It is much more important that the rules filter out false positives. A few false negatives are ok, because there are enough data coming through Usenet daily that these will tend to be overcome. But false positives lead to incorrect recommendations and must be kept to a minimum.

## **3.2 PHOAKS ARCHITECTURE**

The PHOAKS system architecture was carefully designed to be general enough to support many different text filtering and collection tasks. There are three main parts of PHOAKS: filtering, categorizing, and disposition. New functions can be plugged in to perform these three tasks to create a new system with different behaviors. Examples of different tasks this architecture is capable of supporting are URL filtering from net news (described below), FAQ collection, or personal mail filtering.

### **3.2.1 PHOAKS News Agent**

#### *3.2.1.1 Filtering*

The first part of PHOAKS extracts recommendations of web resources from Usenet messages. To do this, PHOAKS searches through every message of net news looking for a pattern (`http://`) that indicates the following text is an URL. Any message that contains binary data is ignored because these messages are typically long and take too much time to filter. Messages containing binary data do not normally contain URL mentions other than in the message header, so not many mentions are missed.

To allow general searching in the filtering module, PHOAKS searches for a textual pattern or a regular expression. The pattern recognizers allow boolean operations so the pattern can consist of multiple phrases, each assigned different weights. Also, different sections of the message or text can be specified to reduce search time. Sometimes it may only be necessary to search for patterns in the subject header while other times they should be looked for throughout the entire message body, depending on the application.

### 3.2.1.2 *Categorization*

Next, PHOAKS must classify every URL mention into one of the categories described above. This is done by first performing a sort of tokenization of the message. We developed a set of features about each message that describes aspects of the message that we thought were important for determining the category of URL mentions in a message (i.e., the URL Block feature corresponds to the block of text surrounding an URL mention). While developing the rules described above, additional features were added so messages could be classified into the most appropriate categories. These features are used in conjunction with the rules to categorize URL mentions.

Combining these syntactic features into rules gives a systematic method of categorizing URL mentions from the messages. For example, if the URL is within 20 lines of the end and occurs after a double dash, the category is a signature. If there are no special characters to set off the signature, there is another rule that looks for standard signature items like email addresses, phone numbers, etc. If the URL occurs toward the end of a message and contains any of the signature items, it is also categorized as a signature.

### 3.2.1.3 *Disposition*

Finally, after all other processing is finished, the message data and its category are stored in a database for later retrieval. At this point the HTML (Hypertext Markup Language) for each URL mention in the message is fetched. A title is taken from the page text and stored for use as the display name for the resource. Also, a reduced representation of the page text is kept so that a search index can be built on this text. This process not only provides this valuable information but also allows PHOAKS to ensure that the URL is a valid resource. If not, it is marked as unfetchable and not checked until the next iteration of the system. The web pages for each URL mention are continually checked to make sure that they are current. There is also an occasional problem of network lag and unavailable servers where a resource may appear to be invalid. If a resource once existed and is no longer available, it must be confirmed five times before the resource is dropped from PHOAKS.

Since the web is constantly changing, sites tend to move frequently. If a site move is done in a standard way, such as adding a field in the http header, or using a meta-refresh tag, PHOAKS can determine that the two sites should be equated and combine the data for the two records. If, on the other

hand, no forwarding information is left or some non-standard method of guiding users to the newly moved site is used, then PHOAKS will track two different sites. Since the title of the web pages is usually the same in these cases, PHOAKS can infer that the sites are the same, and will list the sites together when presenting the information to users. However, site information is not combined and will not affect the frequency count for the site.

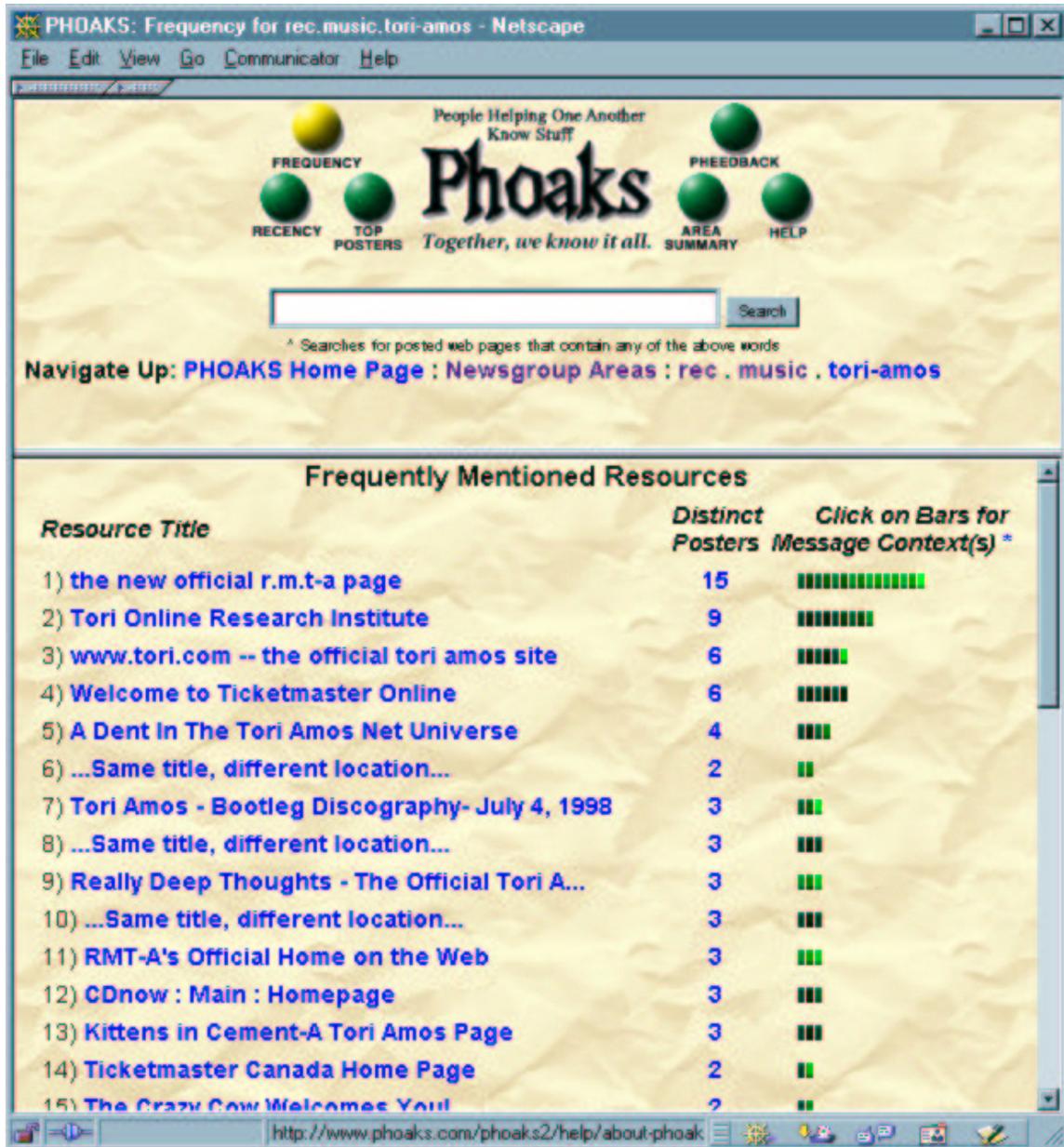


Figure 3.1: PHOAKS Web Interface

### **3.2.2 Web Interface**

The last component of PHOAKS deals with displaying information from the database to users in the form of web pages (shown in Figure 3.1). This system was developed to be easily extendable and updated. It incorporates a template language so site maintainers can build page templates to dynamically generate web pages based on the templates. This page definition language is an extension of HTML that adds iteration and conditional constructs and a set of variables specific to PHOAKS data. The language makes it easy to describe, for example, a resource summary page as an iteration over all recommended resources for a newsgroup.

There is also a software layer in PHOAKS that fills database requests from the dynamically created web pages. When a page is requested by a PHOAKS user, the template is checked and any constructs and variables are translated into database requests. Then the database layer makes queries to the database and returns all requested items and finally a page is created. Since speed of presentation of web pages was important, we developed a caching feature to pre-cache popular pages for each newsgroup and keep a cached copy of any page that a user has requested (as long as it is still valid). Now, when a page is requested, a CGI (Common Gateway Interface) script first checks to see if the page has been cached. If it has, then the page is simply displayed in the browser. If not, the page is generated, cached, and displayed. Since 75% of the pages accessed by users of PHOAKS are resource summary pages and index pages, these are the pages that are pre-cached every time the database is updated for a newsgroup. This helps keep simultaneous database accesses to a minimum and lets users get commonly accessed pages back immediately.

### **3.3 LESSONS LEARNED**

PHOAKS effectively solved the problem of automatically collecting quality web sites about a topic. In addition, we showed that Usenet messages are an abundant source of recommendations of web pages, that recommendations could be recognized automatically with high accuracy, and that there is some correlation between the number of recommenders of a web page and other metrics of web page quality. However, there were a number of aspects of PHOAKS that needed improvement.

The basic unit of the items recommended by PHOAKS was the web page. However, for many purposes, the web *page* is the wrong unit of information. The World Wide Web consists of many web *sites*, coherent, structured multimedia documents consisting of many individual web pages. Many times PHOAKS would contain recommendations of multiple web pages within a single web site. Clearly, recommendations for these pages could be aggregated to count as recommendations for the common web site that they are a part of. We want to group web pages into sites and present the consolidated structure of the web site to users. But, we also must keep the original pointers to individual parts of the web site so that we may indicate which areas of a web site might have been more popular.

A general goal of PHOAKS was to collect as many relevant web pages as possible while including few non-relevant pages. Because PHOAKS monitored newsgroup discussions and there were some off topic web pages mentioned within newsgroups, PHOAKS sometimes collected web pages that did not concern a particular newsgroup's topic directly. Another common occurrence in newsgroups is the posting of general publications across many newsgroups. These publications may contain a few resources relevant to the newsgroup and many that are not. The opposite situation also arose. Since PHOAKS had a "no self promotion" rule, web pages mentioned by the site maintainer were not included. In a few of these cases, since the page was already mentioned in the discussion, additional users felt no need to repeat the recommendation and therefore the page was not included in PHOAKS. In future designs, we want to filter out irrelevant resources and include more relevant resources.

PHOAKS was based mainly on a single ranking metric: the number of distinct individuals who recommended an item. This metric is useful for many purposes, but situations arise where users need additional metrics to evaluate sites. By including numerous ways of comparing web sites, we can meet this need, plus help to eliminate the case where a quality site is excluded because of a low ranking within a single metric. The main representation of a web page in PHOAKS is the title, which may not always be the best way to communicate what a page is useful for. It is our goal to construct representative profiles of web site content and structure that make it easy for users to evaluate sites, helping them to determine both site quality and function.

Finally, there was no information workspace included in PHOAKS. We found that users had a desire to define and organize personal collections of web resources. In future interfaces we want to

implement an information workspace that allows users to easily manage their resources and make it easy for them to share their collections with others.

## CHAPTER 4: TOPICSHOP SYSTEMS

### 4.1 WEB CRAWLING

There are many different sites on the web for any given topic. An alphabetized list of all known sites is rarely the best method for finding useful information. The inherent hierarchical structure of the web can be used to gain further information about web sites. By following all hypertext links on a web site, a topic crawl can be generated for all sites linked to by a particular site. Continuing the crawl deeper into these sites will eventually provide a large body of topically-related sites that can be analyzed and presented to a user. This is based on the assumption that quality sites point to other relevant quality sites. Since site designers have theoretically already put effort into filtering out poor quality sites and only linking to quality sites, a crawl can simply follow links to build a better representation of the scope of sites for a given topic.

The basic unit used in many search engines is the web page. While this may work for very specific topics, many times users need to be guided to appropriate sites containing information on a variety of sub-topics. In building topic crawls, the basic component we use is web *sites* rather than web pages. A site contains a coherent body of content on a given topic and is divided into pages, usually grouping related information, to ease navigation. Pages that make up a site can be roughly sorted into three basic categories: navigation pages, content pages, and links pages. *Navigation pages* provide structure to the site content by giving indices and table of contents that a user can click on to find further information. One

navigation page usually represents the top or front page of the site and provides the starting point for navigating the rest of the site. This page is also commonly called the index page and is intended to be the first page a user sees when viewing the site. *Content pages* contain information about the topic that the site is representing. *Links pages* usually do not add any additional content of their own and are simply collections of links to other sites related to the topic. Of course not all sites follow this format, but many use this or a similar structure to provide users with easier ways to maneuver through the site. Pages are grouped into sites using heuristics that look at the directory structure of URLs. For example, if the crawler encounters a link to the URL `http://a/b/page1.html`, and `http://a/b/index.html` is a site known to the crawler, it records this URL as part of the site. Further, if the link was encountered while the crawler was analyzing the site `http://x/y/`, a link is recorded from the site `http://x/y/` to the site `http://a/b/index.html`.

Users can generate topic crawls by giving the crawler a user-defined set of *seed pages*. These seeds can be obtained in various ways: a list of pages that a user already knows about, the output from a search engine, or a list of URLs from PHOAKS. The crawl starts from these seed sites and follows links found on the seed pages by fetching the HTML of the corresponding page for each link on the seed page and analyzing the content. This process of analyzing content and following links continues for all pages within two links of a seed page. If links are internal to the site (point to a page from the same site), then the pages are added to the collection of pages already found for this site and the internal site structure is slowly revealed as the crawl progresses. Links that are external to the current site (point to a page on a different site) add to known sites about the topic. As more sites are visited, the inter-site structure is recorded during the crawl. A link does not have to point to the top page of the site; it can go from any page on a site to any page of another site. The resulting structure can be thought of as a directed graph of the sites with vertices representing sites and edges representing links between sites, called a *site graph*.

Our crawl uses a *clan graph* as the primary information structure. A *clan graph* is a directed graph where nodes represent documents and edges represent a reference to the node pointed to. A local clan graph is the subgraph whose nodes are closely connected to the user-specified set of seed sites. Building on concepts from social network analysis [48][86], co-citation analysis [35], and social filtering [46] we have developed the notion of an NK local clan graph.

- The NK local clan graph for a seed set  $S$  is  $\{(v,e) \mid v \text{ is in an } N\text{-clan with at least } K \text{ members of } S\}$ .

An N-clan is a graph where every node is connected to every other node by a path of length N or less, and all of the connecting paths go through only nodes in the clan. Our crawler uses a 2-clan (the 2K local clan graph) because it represents a useful substructure extracted from the large structure of the web. By requiring that sites relate to a certain number of seeds (K), we ensure that we find not just dense graphs, but graphs in which a certain number of the seeds participate.

There are three types of inter-document relationships where a relationship between two of the documents can be inferred based on a known relationship between the other two. Co-citation analysis says that two documents B and C are related if a third document, A, cites them both. Social filtering says that if documents B and C both refer to a third document, A, then B and C may be likely to link to similar sorts of items in general. Transitivity says that if document A refers to B and B to C, then A implicitly refers to C. These three relationships are the minimal 2-clans which are, in our case, necessary because no smaller structure allows us to make inferences about document relatedness, and sufficient because no large structure enables other simple inferences [95].

During a crawl, a number of parameters describing sites are gathered. Number of images, audio files, and movie files are recorded, as well as the number of in-links and out-links. The number of links pointing to a site by other outside sites is called the *in-links*. This parameter can be used to determine if the site is a popular site by finding the number of site designers that think it is good enough to be linked to. This is a form of social filtering. By considering each in-link to a site to be an endorsement to that site we can generate a list of the most linked-to or most endorsed sites. An *out-link* is where a site links to another site. The site with the most out-links can be considered a good index site with many links about the desired topic. Combining these two parameters can provide further information. If a site is pointed to by many sites, but does not point to any other sites, it may be an official site (perhaps a corporate site) on the topic since many sites think its important, but the site itself does not point to any other sites. If, on the other hand, a site is not pointed to by many other sites but itself points to a large number of other resources, it may be a newer site that other site designers have not noticed yet. Most likely, it is a link collection site if it has a high number of out-links.

While a crawl is being performed, two metrics are used to ensure that highly relevant sites are visited in the early stages. First, a weighted sum of the number of in-links of all sites that point to a page is

used to rank the page on its potential for not only being a quality site but for recommending other quality sites. As a crawl progresses, this ranking is improved because more data about visited sites are collected. If a site is pointed to by many other sites with a high number of in-links (and hence are considered good sites because they are endorsed by others), then this site can also be considered a good site. Because of the immense size of the web, a crawl can take a very long time but by using this metric, more relevant sites are found by the crawler near the beginning of a crawl and a crawl can be stopped after some user-defined threshold number of sites is found. In addition, anchor text is searched for keywords related to the crawl. *Anchor text* is the text description, written by the site designer, that is displayed for each link and is what the user clicks on to visit the site linked to. This text is usually highly related to what the site contains. So during a crawl, all occurrences of anchor text are saved for each site and can be searched to gain relevance feedback. If a match is found, then the ranking for the site is improved; if no match is found, nothing is done, because that does not necessarily mean a site is off-topic.

We noticed differences among the structure of crawls for certain topics. In particular, sites whose purpose was sales and business-related activities tended not to link to other sites and were isolated from the rest of the site graph. For business topics, 79% of the sites were isolated as opposed to only 32% isolated sites for non-commercial topics such as entertainment. Similarly, the average density of the graph for business-related topics was 0.004, but for other topics was much higher at 0.071. Topics dominated by merchants competing for the same customers do not exhibit collaboration and are not good candidate topics for our systems. Collaborative filtering based on linking can work only for topics with a significant number of inter-site links.

An interactive Java applet is used to generate crawls on a server based Java crawler. The user enters into the applet a few seed sites on a topic that they want to crawl. The applet sends this information to the crawler running on our server. The crawl begins and the user is given feedback about the status of the crawl in the form of sorted lists of thumbnail images. Users choose what parameters they would like to watch (in-links, images, pages, etc.) and then rows of images in sorted order are displayed so the user can see what sites are being gathered in real time. When a crawl has completed and satisfies the user-specified parameter of number of sites to gather, individual files are compressed into a zip archive and downloaded

by the user to use as data for further visualization. This client/server architecture allows the system to be used by multiple web users efficiently and allows us to monitor the crawls being performed.

Once a crawl has been completed and a database of related sites has been compiled, data can be further analyzed and presented to the user. Two visualizations of collected data that we have developed are described in the following sections. The first interface, WebCite, is a graphical interface and the other, TopicShop, combines thumbnail images with a simpler text-based representation.



Figure 4.1: WebCite User Interface

## 4.2 WEBCITE

The WebCite user interface (shown in Figure 4.1) displays a graphical thumbnail image for every site visited in a crawl. The layout is a group of concentric semi-circles based on an auditorium seating

metaphor with sites most central to the topic located on the inner ring of images and fanning outward in elliptical rows where graphical thumbnail images get smaller as they get farther from the center. Centrality is determined by a metric that combines in-links and out-links of each site (the number of 2-clans the site occurs in). The title of the main page of each site is placed next to each thumbnail and the original seed sites are marked with an asterisk. The interface is interactive; when a user moves the mouse cursor over one of the thumbnail images, it is highlighted by increasing its size. In addition, links to and from the page are shown with colored arrows indicating the links. Arrows pointing toward the highlighted site from other thumbnail images indicate that other sites link to the highlighted site. Likewise, arrows pointing toward other sites indicate the highlighted site links to other sites. Any site that does not contain a link to the highlighted site and is also not pointed to by that site fades to black, so that just the relationships to and from the highlighted site can be seen. Sites can be visited by double clicking on them, which opens up a web browser and displays the top page of the site. Clicking the left mouse button while holding the shift key over any thumbnail image will display internal pages for that site.

#### **4.2.1 Lessons Learned**

The WebCite user interface is a visually pleasing way to display collections of web sites on a topic. Layout of the thumbnail images begins to show some relationships of web sites in the collection. Much of the important information, such as in-links and out-links, is hidden within the interface. Users are required to manipulate the interface by clicking and moving the mouse to see the link structure between the sites emerge. This is, of course, better than presenting all inter-site links at once, resulting in a cluttered, useless display of the structure. We will see later (section 4.3, TopicShop) that by using the number of in-links and out-links as parameters of the site and presenting ordered lists, we can allow the user to see important aspects of the inherent structure of the sites.

User evaluation of this interface revealed that users wanted to change the structure and move the thumbnail images to suit their needs. We want to support this by letting users organize collections to reflect their own understanding of the topic area by grouping and categorizing items.

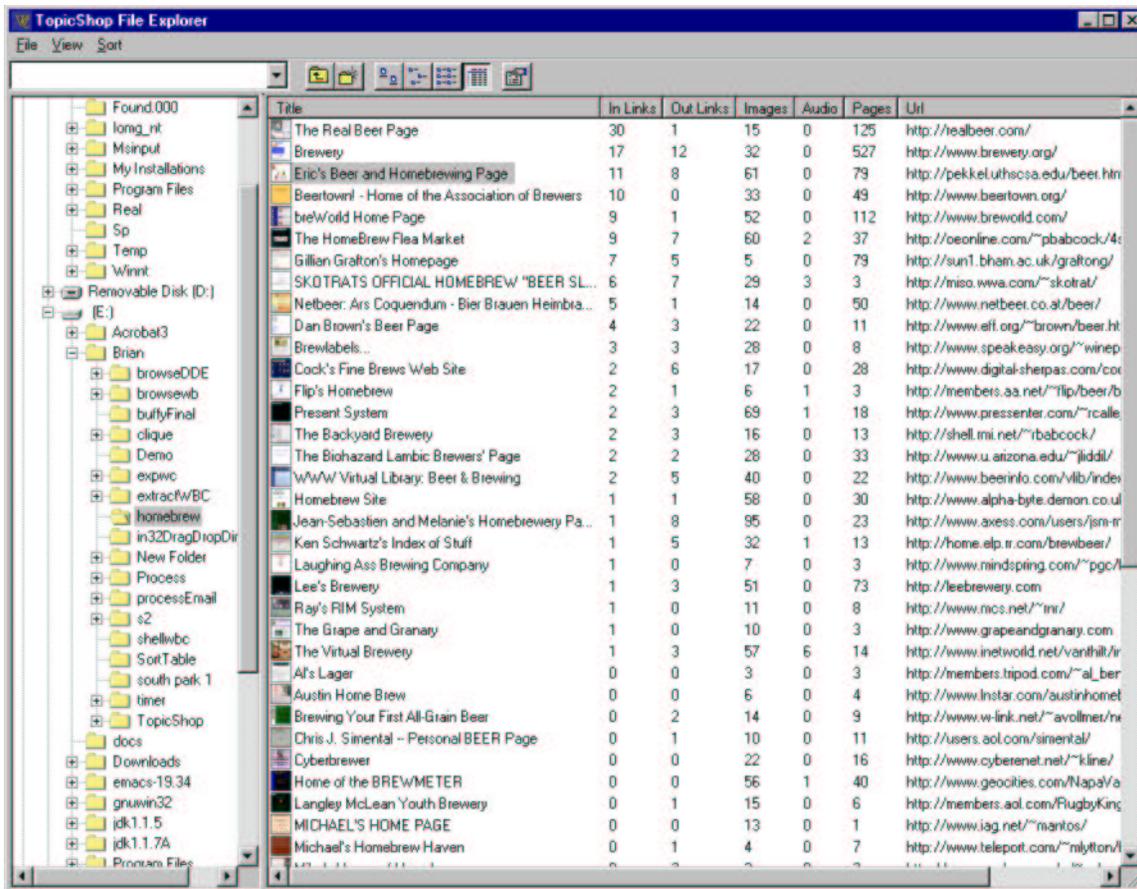


Figure 4.2: First version of TopicShop (Details View)

### 4.3 TOPICSHOP

Another visualization for viewing and managing collections is the TopicShop Explorer, a customized version of the normal Windows file Explorer. The TopicShop Explorer, shown in Figure 4.2, is a very small Windows executable that knows how to read and process site profile files.

Users can view their collections in two different ways: details or icons. The main feature of the details view is that it shows site profile information, and the main feature of the icons view is that users can arrange icons spatially (Figure 4.2 shows the details view; Figure 4.3 shows the icons view). We had three main design goals for TopicShop Explorer:

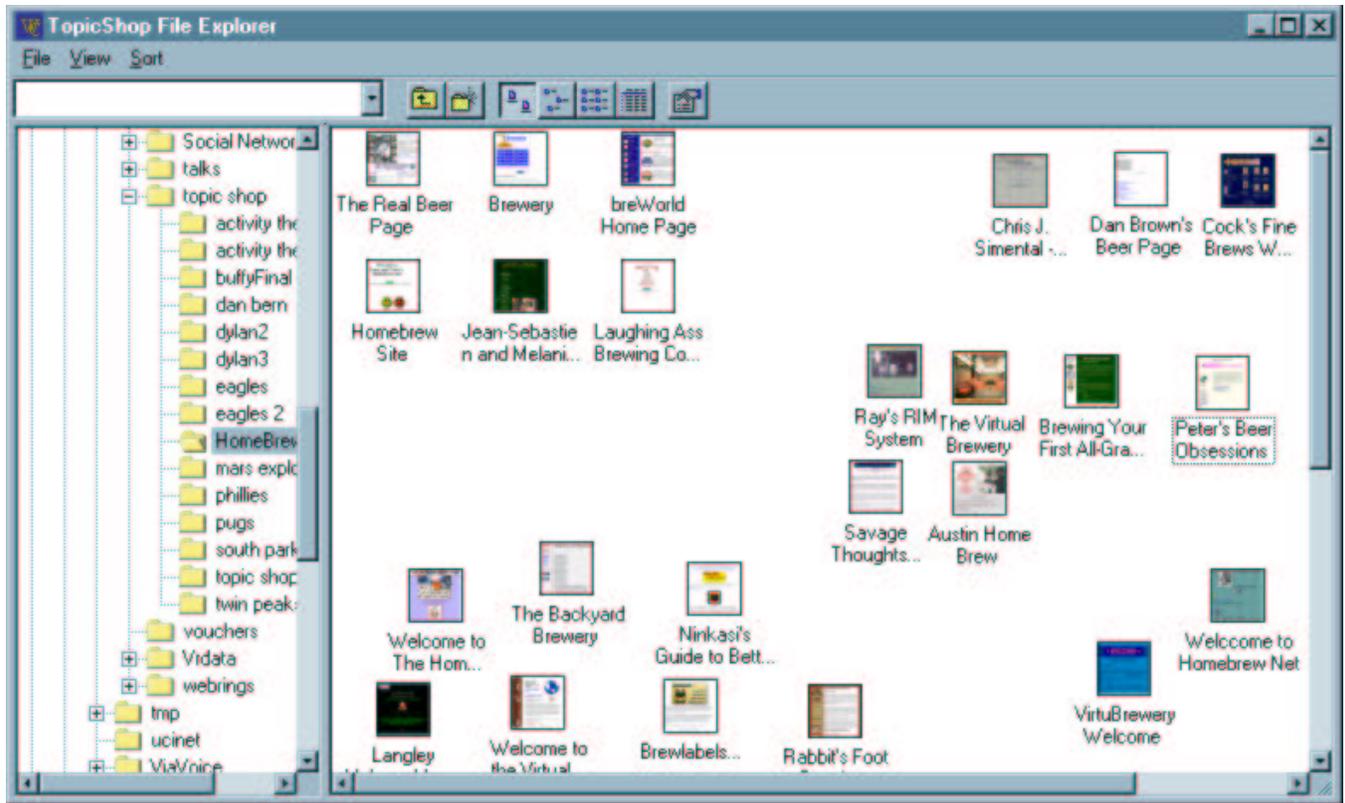
1. *Make relevant but invisible information visible.* We hypothesize that making site profile information visible will significantly inform users in evaluating a collection of sites. No longer must they decide to visit sites — a time-consuming process — based solely on titles and (sometimes) brief textual

annotations. (A chief complaint of subjects in the Abrams et al., [1] 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 can choose to 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). In addition to site profile data, the thumbnail images also are quite useful; most notably, for sites a user has visited, thumbnail images are an effective visual identifier for sites.

2. *Make it simple for users to explore and organize resources.* In the details view, users can sort resources by any of the properties (e.g., columns showing number of in-links, out-links, images, etc.) simply by clicking on the label at the top of the column. In either view, right-clicking on a site brings up a window that shows profile data from which the numbers in the columns are derived (e.g., lists of all sites that link to the selected site and all internal pages of the site). Double-clicking on a site will send the user's default web browser to that site.

Users can organize resources both spatially (in the icons view) and by creating subfolders and moving resources into the subfolders. Nardi & Barreau [73] 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. As categories do become explicit, users can create folders to contain sites in each of the categories.

3. *Integrate topic management into a user's normal computing and communications environment.* The TopicShop Explorer may not look like a novel interface at all; interestingly enough, this was an explicit goal. We wanted it to be as similar to the normal Windows Explorer as possible so Windows users could apply all their existing knowledge, meaning there would be little or no learning time and similar ease of use. Further, this decision makes it very easy for collections of resources to be shared. Since a collection is just a normal Windows folder containing files (of the special type that we designed), they can be shared in all the normal ways. As we already have explained, a collection can be compressed and downloaded. It can also be emailed. And if users share a common network, collections simply can be read directly from any machine on the network.



**Figure 4.3: First version of TopicShop (Icons View)**

The TopicShop Explorer interface allows users to organize their web site collection from any view. In the details view, users can change the order of the collection of web sites to represent their personal choice of best quality sites. This ordering becomes an additional column in the interface that can be sorted like any other column. In the icons view, spatial organization is allowed and web site icons can be arranged into groups before being moved to a new folder.

#### **4.4 CURRENT INTERNET RESOURCE DISCOVERY TECHNIQUES**

Existing web search approaches can be broken down into a number of types: comprehensive indices, keyword searches, hybrid directory/keyword searches, specialized indices, socially filtered interfaces, and task-specific interfaces (i.e., TopicShop).

#### **4.4.1 Comprehensive Indices (Web Directories)**

One popular search approach is the comprehensive index (Yahoo, Netscape, Lycos, Infoseek, etc.). This type of search engine has human web librarians that search and evaluate sites and place them in an appropriate category usually in alphabetical order. The result can be considered a web directory broken down into categories. This typically leads to highly relevant sites, but relies on human involvement to make decisions about which sites are on-topic and which sites are not. Because of the human role, directories can often provide better results than search engines. Search features are available to automate finding correct categories and/or sites for a particular query, by allowing users to specify keywords that are matched against category headings.

#### **4.4.2 Keyword Searches**

Another type of web search engine is the keyword search engine (Alta Vista, Magellan, Excite, etc.). On these sites, a user specifies a query in the form of a list of words and is given back a list of pages ordered by a textual matching metric. This is done by implementing automated crawlers that catalog the web by following links on each page it finds and building indices to search on, thus eliminating the need for human intervention. Many times there are multiple pages from the same site listed in the results. These search engines constantly visit web sites on the Internet in order to create catalogs of web pages. Because they run automatically and index so many web pages, keyword search engines may often find information not listed in directories.

#### **4.4.3 Hybrid Directory/Keyword Searches**

Many search engines that began as keyword search engines are slowly incorporating a categorical directory index in their database. Designers of these search engines apparently saw that providing some sort of structure to their list of sites would be very beneficial to their potential users. By default a user's search query is still answered with a list of pages from the search engine's catalog of the web, but a rough directory index is available as well for at least some of the sites returned.

#### **4.4.4 Specialized Indices**

A fourth type of search approach is the specialized index, which can be further broken down into two distinct categories: links pages and web ring interfaces (Links Pages, WebRing, Looplink, etc.). People interested in a topic that want to provide resources to other users create links pages. Typically links pages are just a list of resources presented on a page of a web site, sometimes categorized by sub-topics. Web rings attempt to provide some structured information to groups of pages by enabling users to form rings, linking together sites related by topic. Usually one user is the ringmaster and allows other site maintainers to join the ring. Navigation among the sites in a ring is accomplished either by going to an index page of sites or moving around the ring of sites by following links on each page. This provides a sort of topic community of interested site maintainers that support each other by bringing users to the whole ring rather than just to their site. The interface is not very efficient because users must traverse through pages in the ring to find information they are interested in.

#### **4.4.5 Socially Filtered**

There are also search approaches that attempt to utilize user behavior as a predictor for relevant web sites (Alexa, Firefly). These systems watch where users navigate and also collect ratings from users about sites they visit. Mapping current behavior to the database provides a method for matching users and making recommendations on what other sites are likely to be related to the current site. These systems are highly automated, but still require users to give some feedback in voting for sites. They lead to a collection of many related resources but still do not provide a comprehensive overview of available information.

#### **4.4.6 TopicShop**

While directories appear to contain higher quality resources, with all items likely to be relevant to the topic, they require a large amount of human effort to construct and maintain a good collection of items for a topic. Search engines are automated, requiring no human effort, offer much more data, and may include relevant items that were missed by a human librarian maintaining a topical collection. However, search results often contain irrelevant information due to the ambiguity of most queries, usually have poor organization, and almost always contain duplicate pages and dead links. TopicShop attempts to combine

the best of these two approaches with the computational means of a search engine to construct high-quality topical collections of a directory, with the addition of representative profiles that users can use to evaluate the quality and function of the resulting items.

Task-specific interfaces, like TopicShop, may make finding relevant information faster and more efficient. Collecting user feedback is a step in the right direction for automatically gathering information about sites, but requires all users to perform some amount of work before gaining the benefit of getting the information. Table 4.1 shows some features of current search interfaces.

<b>Interface</b>	<b>Graphical/ Textual</b>	<b>Structural Data</b>	<b>Sub-Topic Categorization</b>	<b>Static/Dynamic Interface</b>	<b>Computer Generated/ Human Filtered</b>
<b>Categorized (Yahoo)</b>	Textual			Static	Human
<b>Rings (Web Ring)</b>	Textual	X		Static	Human
<b>Search Engines (AltaVista)</b>	Textual			Static	Computer
<b>Links Pages</b>	Textual		X	Static	Human
<b>SortTable</b>	Textual	X		Dynamic	Computer
<b>WebCite</b>	Graphical	X	X	Dynamic	Computer
<b>TopicShop</b>	Graphical	X	X	Dynamic	Computer

**Table 4.1: Comparison of search interfaces**

There appears to be a distinct division of labor between those people who prepare content sites and topic guides and those people who utilize them. Many people are not interested in rating sites and would rather just find what they are looking for, but there is a small collection of motivated people who do want to provide information. Harnessing the knowledge and motivation from these people will benefit other information seeking users. We propose to make this easier for both classes of people by semi-automating the preparation process with our web crawling system and then improving management of information with task-specific user interfaces. In addition, many site maintainers have already put forth some effort in linking their site to other sites that they feel are adequate resources about their topic. Using this information directly eliminates the need for users to contribute personal ratings. Each link into a site can be considered an endorsement of that site, and used to rank linked-to sites. Web crawls can be performed by a

few individuals, who are highly interested in the topic, and presented for other users to view. These people can be considered *topic librarians* and are responsible for managing collections of web sites for a topic. The other class of user is the one attempting to view what resources are available about a topic. This person is a *topic novice* and can either be a novice regarding the topic at hand or a novice regarding the availability of web resources on that topic. Either way, they need guidance on the structure of available resources and an introduction to the most useful sites about their topic of interest.

This is a promising way that resource discovery and ranking techniques presented above can be used. Index pages exist on the web for many, if not most, topics. However, indices have at least two problems. It is difficult for them to be comprehensive and up-to-date, and, paradoxically, the more comprehensive they are, the harder it may be to focus in on just high-quality sites. Our systems can address both problems. A person maintaining an index can apply techniques from our systems to follow links from the current index and discover new sites that may be relevant. The discovered sites can be presented to an index maintainer who then can decide which ones to add to the index. And site connectivity information can be used as an aid in ordering sites within the index.

This process is collaborative in two ways. First, over time, a topic index becomes a product of emergent collaboration, since it contains sites because they were linked to by sites from earlier incarnations of the index. Second, this also is a human-computer collaboration process, with a web crawling algorithm continuously suggesting new sites to an index maintainer based on their relevance to sites already in the index, with the maintainer retaining the final decision over what sites to add.

## CHAPTER 5: OVERVIEW OF USER STUDIES

The need for topic management was motivated in Section 1.2. This research is an attempt to explore and produce effective and efficient mechanisms for supporting this task.

### 5.1 HYPOTHESIS

- TopicShop, with sort orders, categorization, user collections, etc., is more effective and efficient for the task of topic management than typical web search engines and indices that use simple alphabetization and site annotations (e.g., Yahoo).
- Socially filtered data over time will provide a better set of topical resources than automated keyword search engines.

### 5.2 EXPERIMENTS

We performed a series of evaluations to determine the effectiveness and efficiency of several different user interfaces to socially filtered data, and provide empirical evidence regarding the benefits of transparent data-rich interfaces. In addition to the normal parameters of usability studies, the social filtering interfaces we evaluated provide an additional variable to control: topic content. The subject area in which a web crawl is created for these studies can either be pre-determined and held constant for all

subjects, or it can be varied to fit with each user's personal interests. Each of these approaches can provide valuable feedback and have unique benefits in a usability study.

Personalized topic areas guarantee that a subject already knows a good deal about the topic and may even have a grasp for the breadth of web pages available for the topic. If they already know what is on the web, they have an idea of what they want to look for and will be able to generate better, more specific search strategies to find exactly what they want to see. A topic expert has already seen a large portion of the available web content on a topic and will be interested in finding additional sites that they have not been exposed to and any new information that may have been recently generated. With prior knowledge of topic content, these users will be able to concentrate on the interface itself and be able to compare it to other interfaces they have used in the past to investigate their topic.

The other option for providing web crawls is to assign all users an identical topic for which they are a novice. This will ensure that each user has a similar experience in using the interfaces because they will be given the same initial data. A better comparison across users will be possible when the data set is held constant. When a user is not familiar with the topic, they will rely on the interface to provide them with good content that will begin to teach them about the topic they are researching. Their needs differ from users who are selecting their own topic because topic novices will be more interested in first gaining a broad overview of a topic which is likely to be contained in the most linked-to sites.

A combination of these two approaches has been used for the two evaluations of TopicShop. The expert evaluation was performed by topic experts, who were presented with sites on the topic they were interested in. Experts were selected based on their self-perceived knowledge in one of the topic areas that we selected for the studies. Those same topic collections were then presented to novice users who had no prior knowledge of the topic.

### **5.2.1 Selecting a Domain**

The web is an immense repository of information and continues to grow at a very rapid pace. In order to do almost anything on the web, users must apply some type of search strategy. Early in the research, we speculated that analyzing current methods that people use when searching the web leads to insights into how user interfaces can better support users in finding information efficiently on the web.

To quantify this, we studied a set of approximately 770K queries issued to the Magellan search engine between March 1997 and August 1998. The Magellan search engine published on their web page a random sampling of twelve queries that users were currently performing with their search engine. Our sample was taken by collecting the twelve sites every 10 minutes and writing them to a standard text file. By breaking these searches down into their keywords and eliminating common stop words, we discovered that out of 1,473,077 keywords, only 159,725 (10.4%) of them were unique. So there is a large overlap in the topics that people are investigating on the web.

The way that users generate queries to Magellan can be either small keyword phrases or large natural language queries. Not surprisingly, small keyword queries are by far more popular with users. Queries containing three keywords or less accounted for 85% of the total queries (one word queries=35%, two word queries=30%, three word queries=20%). Queries of four and five words were an additional 11% of the total, which means that queries of six words through 66 words occurred only 29,475 times (~4%).

We also wanted to investigate the major topics that people are researching on the web. We decided to analyze a sample of the top 10% of all queries. The 515 most commonly occurring queries accounted for ~96,000 queries and represented approximately 10% of the total queries in the data sample. By categorizing the top 515 queries, we got a good idea of what topics are important on the web. We chose the top-level categories by reading through the 515 queries and coming up with seven distinct categories: business, current events, entertainment, sex, internet/technology, travel, and uncategorized. A brief description of each category and the list of queries were given to two independent raters and they were asked to categorize the 515 queries. After analyzing the entire categorization (inter-rater reliability of .85, Cohen's Kappa,  $p < .0001$ ), we determined that 42% of the queries had to do with entertainment topics, including media fandom. The next two most popular topics were sex and internet/technology, accounting for 25% and 23%, respectively. Entertainment topics made up almost half of the queries performed and as such are a representative area to study using the TopicShop interfaces.

### **5.2.2 Introduction to Pilot Study**

The initial study we performed was a pilot study comparing Yahoo with TopicShop using 16 subjects and 8 topic experts. These subjects were given sets of 60 sites on one of two topics. This small-

scale study was used to verify that our hypotheses were on target and also to ensure that the methodologies we designed would capture the types of data we were interested in analyzing. Instead of using a couple of pilot subjects in a larger scale study, we designed this smaller study so that we would be able to not only eliminate any problems in our experimental design, but also obtain meaningful results without wasting any user data. With these data, we were also able to iterate on the design of TopicShop and make improvements based on this first user study. Details of this study are presented in the following chapter.

### **5.2.3 Introduction to Interface Evaluation**

After redesigning TopicShop to reflect the things that we learned from the pilot study, we wanted to perform a more thorough evaluation of our interface. This larger evaluation used 40 subjects and 15 experts covering 5 different topic areas on the web. The number of sites each subject was given was also increased to better represent the magnitude of content available on the web. One important design change that users requested was better support for organizing the collections of sites they selected. Because of this, the task was changed to include organization and categorization of the selected sites. This allows us to look at what ways we can better support this operation and also investigate how much agreement there is about categories within a topic. Details of this study are presented in chapter 7.

## CHAPTER 6: PILOT STUDY

### 6.1 INTRODUCTION

We wanted a topic management web tool as a suitable baseline for comparison to TopicShop. Yahoo is the most widely used means of finding and browsing collections of web resources. Figure 6.1 shows an analysis of search engine usage on the web between March 1997 and September 1999. These data were provided by Media Metrix [103] and were based on a sample of 50,000 web users. The chart shows the percentage of these web users that use each different search engine on a regular basis. Clearly Yahoo has been the most popular search engine, with Netscape and Microsoft Network following close behind. These are all category-based search directories and, as shown in the graph, they are more popular than the keyword search databases like Excite and AltaVista. Yahoo is also the largest of the search directories cataloging over one million sites, while the closest competitor contains less than half that amount. Bookmark lists are probably the most common means of organizing collections of resources. According to user surveys done at Georgia Tech [104] over the past 4 years, bookmarks were cited as a browsing strategy for locating information on the web by 80% of all participants. Bookmarks are built into most web browsers and are easy to use, requiring only a mouse click or keystroke to save a site for future reference. The study by Abrams, Baecker and Chignell [1] and the Data Mountain system by Robertson et al. [84] also indicate that bookmarks are a popular method of storing personal collections of sites.

Therefore, we decided that subjects would use either TopicShop or Yahoo/bookmarks. We chose two entertainment topics for the pilot study: homebrewing and the TV program “Buffy the Vampire Slayer”. Each contained about 60 sites on their corresponding Yahoo page. Our choice of these topics was influenced by the fact that pursuing special interests, including hobbies and media fandom, is one of the main ways people use the web.

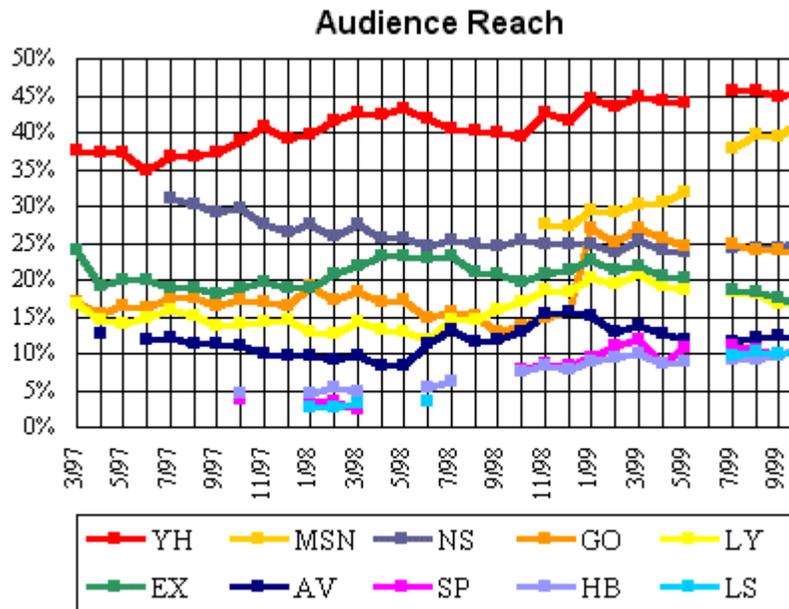


Figure 6.1: Search Engine Usage. (YH=Yahoo, MSN=Microsoft Network, NS=Netscape, GO=go.com {InfoSeek}, LY=Lycos, EX=Excite, AV=AltaVista, SP=Snap {search.com}, HB=HotBot, LS=LookSmart) (MediaMetrix [103])

## 6.2 EXPERIMENTAL DESIGN

To verify that the user interfaces support the tasks they were designed for, our pilot study compared TopicShop and Yahoo/Bookmarks. This evaluation concentrated on the initial version of the TopicShop interface along with Yahoo’s widely used interface on the web. In phase one, four experts in each topic evaluated the sites and gave their quality judgements. In phase two, a 2x2 between-subjects analysis was conducted with 16 subjects (see Table 6.1). Two topic collections (Buffy the Vampire Slayer and Homebrewing) were randomly presented in two different interfaces (TopicShop and Yahoo). The crawls were limited to sites that also existed on the Yahoo page for each topic to keep data sets consistent. Order of presentation of the sites shown to each subject was randomized. We randomly assigned each of

the 16 subjects into one of the 4 conditions resulting in 4 people per condition (topic/interface combination).

<b>2x2 Experimental Design</b>		<b>Interface</b>	
		<i>TopicShop</i>	<i>Yahoo</i>
<b>Topic</b>	<i>Buffy the Vampire Slayer</i>	4 Subjects	4 Subjects
	<i>Homebrewing</i>	4 Subjects	4 Subjects

**Table 6.1: Pilot study experimental design**

The two main metrics we wanted to measure were the quality of resources users gathered and the amount of effort (time and total number of sites browsed) required. To give a quality baseline, in phase one, four topic experts were presented a list of 60 sites (in random order) from each topic; only titles were presented, but no annotations or profile data. This meant that experts had to browse each site and evaluate it based on its content and layout. Each expert collected the 20 “best” sites. For this study, we defined “best” as a set of sites that collectively provided a useful and comprehensive overview for someone wanting to learn about the topic. During analysis, we used the “expert intersection”, the set of sites that all experts for a given topic selected, as the yardstick for measuring the quality of resources selected by subjects. It turns out that the “expert intersection” was 12 sites for both topics; we will discuss expert intersection in more detail below.

In phase two for both the TopicShop and Yahoo conditions, topic novice subjects were presented with 60 sites from the appropriate topic, whose quality they were to evaluate. Yahoo subjects saw (as usual) site titles and, for about half the sites, a brief textual annotation for all sites in the appropriate Yahoo category. For the TopicShop condition, we applied our web crawler to the Yahoo sites to produce site profiles, which TopicShop then displayed.

There were two main goals of this pilot study. First, we wanted to verify that the web crawler interfaces we had iteratively designed work for certain types of web search tasks, like maintaining a links page. Second, we wanted to develop some expert rankings of collections of web sites in an attempt to quantify what factors go into quality web sites.

### **6.3 PARTICIPANTS**

Subjects for the pilot study consisted of volunteers from AT&T. Topic experts included graduate students from Virginia Tech and employees from AT&T.

### **6.4 METHODOLOGY**

In phase one, experts in each topic were given a list of web site titles of the 60 sites for their topic in random order. The instruction sheet they were given contained information explaining the task and the definition of a quality site. They were asked to look through the links exhaustively and choose the 20 sites they thought were the best quality sites, keeping them in ranked order. Experts took approximately four hours each to complete this task.

In phase two, subjects (topic novices) were assigned randomly to one of the four conditions. To begin the experiment, subjects received 15 minutes of instruction and training in the task and user interface. TopicShop subjects were shown the basic interface features and taught how to collect sites by dragging and dropping icons into folders. Yahoo subjects were shown a sample list of sites and taught how to collect sites by bookmarking. After training, subjects performed a short task to ensure that they were comfortable with collecting and organizing sites.

For the main task, subjects investigated the sites for their assigned topic by using their assigned interface (TopicShop or Yahoo) and browsing to sites. In both interface conditions, subjects were presented with the same collection of sites for their topic. They were asked to choose the 15 “best” (as defined previously) sites and rank them by quality. Because people do not spend unlimited amounts of time browsing, we wanted to see whether users could find high-quality sites in a limited amount of time. Subjects were asked to complete the task in 45 minutes and were kept informed of the elapsed time at five-minute intervals. Clearly, there is a relationship between time on task and quality of results: the more time spent, the better the results one can expect. By limiting the amount of time, we hoped to focus on any differences in the quality of results (i.e., sites users selected) between the two interfaces.

The task ended when subjects were satisfied with their collections of sites or after 45 minutes had elapsed. Subjects then completed a short questionnaire. Finally, we conducted an informal interview to reveal strategies subjects used to perform the task, their reactions to the interface, and what could help them complete the task more effectively.

## **6.5 DATA COLLECTION AND ANALYSIS**

During the pilot study we observed a number of variables. We recorded time on task for each interface and broke it down into time spent in the interface and time spent browsing web sites. In addition, a keystroke level log captured mouse movement and interface component clicks. This resulted in data regarding where in the interface the subject was and what they were doing during the experiment. Analysis of the browsing history showed each subject's browsing behavior during the experiment. This included percentage of time a subject spent on web pages rather than in the interface, total number of sites visited, and average visit position of the subject's top five ranked sites. This last piece of data is taken from a list of web sites in the order that the subject visited them. The subject's top five sites compared with the first five sites they visited shows whether the subject was able to quickly find the sites they thought were best. By comparing rankings of the sites that the subject selected to topic experts' rankings, we computed a quality metric to rate how similar each subject's list of sites was to topic experts' list of sites. Survey results were also tallied along with some additional statistical analyses on the questionnaire data.

There were two factors to look at when comparing topic novice subjects' lists of best quality sites to topic experts' opinions: endorsement and ranking. Experts not only selected quality sites but also ranked them in order by quality. One method of comparison is the strict intersection of the four experts for each topic. This gives a set of sites that can be considered to contain quality information with a nice layout for the topic, that were endorsed independently by a total of four people. Another way we looked at the expert data was to assign a score for each of the 60 sites in a topic. The score was the number of experts (1 to-4) that recommended the site in their list of 20 quality sites. This gives a larger set of sites that were recommended by at least one expert. Finally, since the expert sets were ranked, a weighted score was computed for each site by averaging its position in each expert's ranked list. This weighted score was then compared against the site position in the subjects' lists to measure similarity to the experts.

## **6.6 QUANTITATIVE RESULTS**

We first compared the set of sites chosen by each novice subject to the expert intersection. For each topic, the expert intersection contained 12 sites. For the Buffy topic, Yahoo subjects selected an average of 5.0 sites that were in the expert intersection, while TopicShop subjects selected 7.5 expert-endorsed sites. For homebrewing, Yahoo subjects matched 4.3 sites and TopicShop subjects matched 9.3.

Overall, Yahoo subjects selected 4.6 sites from the expert intersection, while TopicShop subjects selected over 80% more, or 8.4 sites. We performed an analysis of variance to look at the interaction between topic and interface for the expert intersection results. We are only interested in the main effect of the interface factor, but we want to be sure that topic is not significant and there is no interaction. A 2x2 between-subjects two factor ANOVA (interface and topic) shows that topic is not a significant factor ( $F(1,12)=0.585$ ,  $p=0.459$ ). Also, the interaction between topic and interface was not significant ( $F(1,12)=3.659$ ,  $p=.08$ ). The interface factor we are investigating is significant ( $F(1,12)=32.927$ ,  $p<.0001$ ). Since there is no interaction and no significance of the topic factor, the rest of the results in this section will be presented based on a pooled independent means t-test.

Expert intersection results are summarized in Table 6.2. Thus, TopicShop subjects found significantly more better quality sites in the time given to complete the task. Notice that choosing sites at random would result in obtaining 3 sites in the expert intersection. (Users selected 15 out of 60 sites, or 25%; 25% of the 12 sites in the expert intersection is 3 sites.) The Yahoo score of 4.6 is not that much better than random selection (one sample t-test (test-value=3),  $t(7)=3.87$ ,  $p<0.006$ ). This probably is due to task time limit of 45 minutes. If Yahoo subjects had had unlimited time, undoubtedly they would have been able to find more high quality sites. So, we see that TopicShop users found significantly better sites in the time given to complete the task.

Topic	Interface Type	
	Yahoo	TopicShop
Buffy	5.0	7.5
Homebrewing	4.3	9.3
Average over Topic	<b>4.6</b>	<b>8.4</b>

**Table 6.2: Expert intersection analysis (average number of expert endorsed sites selected)**

If instead we compare the subjects' list of sites to the experts' weighted union, we can see a similar trend. The weighted union is a sum of the ratio of experts that selected each of a subject's sites (if 3 of the 4 experts selected a site, the ratio of experts would be 0.75). The Yahoo subjects' average expert weighted unions were 6.5 and 6.63 for Buffy and homebrewing, respectively, with a total average of 6.56. For TopicShop, the scores were higher, 9.5 for Buffy and 10.81 for homebrewing. The total average of TopicShop subjects was 10.16, or 55% higher than Yahoo subjects (pooled independent means t-test,  $t(14)=-3.97$ ,  $p<.0007$ ). Table 6.3 shows results for the expert weighted union.

Topic	Interface Type	
	Yahoo	TopicShop
Buffy	6.5	9.5
Homebrewing	6.63	10.81
Average over Topic	<b>6.56</b>	<b>10.16</b>

**Table 6.3: Expert weighted union analysis**

It also is revealing to examine the amount of work subjects performed to complete their tasks. A study of data from the search engine Excite [49] (51,473 queries; 18,113 users) showed that 86% of all users look at three or fewer pages (each search results page contained 10 sites) of the search results. This shows typical users are willing to consider no more than 30 pages when browsing the web, many of which can be rejected by examining the title only. In our study, Yahoo subjects browsed an average of 44 sites, while TopicShop subjects visited about 36 (pooled independent means t-test,  $t(14)=1.14$ ,  $p<0.14$ ), or about 19% less. Further, the task of constructing a high-quality collection of resources is more difficult than doing a simple search; the task is global, since a user is trying to develop a comprehensive overview of a topic, so more sites must be considered. By providing additional dynamic data up front, TopicShop enables users to make better decisions about which sites to immediately rule out and which to investigate further. Yahoo users can rely only on textual annotations, which are provided by site maintainers. While these annotations are sometimes helpful, they can be out-of-date or self-promotional, so are not necessarily good indications of the perceived quality of a site.

Our results for the number of sites visited where subjects looked at three or fewer pages were very similar to the Excite study. Subjects went only to the front page (first page of site) of 52% of the total visited sites and navigated to a second page on an additional 20%. Analyzing these results further reveals that sites where subjects visited two pages or less were many times selected into their final list of sites. So, many subjects judged the quality of sites after viewing only the first page or two of a site. In fact, 61% of the sites that subjects selected, matching the expert intersection, had only one or two pages browsed by all subjects.

Subjects tended to visit more sites than necessary while selecting quality sites because they wanted to be sure there were no additional quality sites they might have missed. Even though they viewed more sites than necessary, subjects found quality sites for their final collections more rapidly using TopicShop than using Yahoo. We can analyze this by looking at the visit position of sites for each subject.

A site's visit position is calculated by considering the entire temporal sequence of sites each subject has visited, and calculating the position of the site in that list. The average visit position, of the top five sites from subjects' final sets of selected sites for Yahoo subjects was 21, while TopicShop subjects visited their top five sites within 13 visits on average (pooled independent means t-test,  $t(14)=1.52$ ,  $p<0.08$ ). So, even though Yahoo subjects browsed to an average of 44 sites and TopicShop subjects an average of 36 sites, their most productive browsing took place within the initial 42% of the sites browsed (an average of 48% for Yahoo subjects, 36% for TopicShop subjects (pooled independent means t-test,  $t(14)=1.40$ ,  $p<0.09$ )).

We also analyzed time on task. We did not expect a large difference in this metric since we gave subjects a (soft) limit of 45 minutes to complete the task and kept them aware of elapsed time during the experiment. Since subjects were encouraged to finish within 45 minutes, their times were usually not much more than the limit. Some subjects would have taken more time to complete this task had it been available to them. Still, TopicShop subjects took about 11% less time to finalize their selections (41.5 minutes vs. 46.6 minutes for Yahoo; the difference was not statistically significant but was in the predicted direction. (pooled independent means t-test,  $t(14)=-0.845$   $p<0.21$ ).

In a task like topic management, one of the goals of the interface is to give users some additional information and let them make decisions without having to browse through every page and have the time cost of downloading more pages. Users probably do not want to exhaustively search every available site in order to find a few that they are interested in. Instead, if they have a way to evaluate a collection of sites without visiting every one, they can more efficiently find the information they are interested in. So, the time that they spend in the interface rather than the browser should be maximized. Because TopicShop provides more information in the interface, users can spend more time evaluating sites based on their site profiles and not have to browse to each page to evaluate its content. The percentage of time that subjects spent in the interface rather than the browser was an average of 24.6% for Yahoo subjects and 34.5% for Topic Shop subjects (pooled independent means t-test,  $t(14)=-3.11$ ,  $p<0.004$ ). TopicShop was able to shift 40% more of a subject's time from the browser to the interface. This means that they were able to make more judgments about the potential quality of a site before browsing and visiting the site.

The questionnaire administered to subjects at the conclusion of the experiment asked them how confident they were with their results on a scale of 1 to 7 (1 being very confident, 7 being not at all

confident). TopicShop subjects were slightly more confident than Yahoo subjects (4.5 vs. 4.75). This is probably explained by the fact that they were given the data derived from a web crawl. Since an in-link can be considered an endorsement of a site, TopicShop subjects felt that if they agreed that a highly linked-to site was a quality site, they were agreeing with the existing opinion of other site designers. Yahoo subjects had only their own opinions to rely on and no data to help strengthen the perceived validity of their selections.

The questionnaire gave data on what information subjects found most useful in evaluating a site. TopicShop site profiles include the title and number of in-links, out-links, images, audio files, and pages in a site. The questionnaire asked subjects to rate these properties from most to least useful on a scale of 1 to 7. Subjects rated three of these properties—in-links (2.00), title (2.75), and number of pages (3.00)—most highly. The other four properties had an average score greater than 5. Even though many subjects noted that title is not a very good indication of quality, it still was perceived as one of the most useful site properties. In interviews, subjects explained that titles were useful mainly as memory aids for sites. Thus, subjects considered the number of endorsements (in-links) and the size of a site (number of pages) to be the most useful indicators of quality [6].

The questionnaire also asked subjects what additional information would have helped them in evaluating sites. Six of the eight Yahoo subjects said that the number of links between sites would be very useful. One subject even made it a point to go to the links page of every site visited to see not only what sites were linked to, but also to read any annotations or recommendations made by the site author. Thus, link information was rated as highly useful by those subjects who had seen it and as very desirable by those subjects who had not.

Browser logs show the order that subjects visited sites during the experiment. If we take the ordered list of viewed sites and look at whether subjects selected a site or not, we can see a couple of trends. Figure 6.2 shows a representation of each subject's browser history. It shows the site visitation order, where each site is represented by a character describing whether the subject selected the site and whether it coincided with the expert intersection. Some trends worth noting are:

- In general, shorter length is better
- Long strings of periods (.) represent wasted work



by a period in the browser trails). The other two categories shown in the browser trails cannot be considered productive or wasted work because they represent a difference in opinion between the subject and the experts. TopicShop subjects were productive for an average of 23% of the sites they browsed, while Yahoo subjects were only productive with 10% of their sites (pooled independent means t-test,  $t(14)=-5.38, p<0.00005$ ).

	TopicShop		Yahoo	
	Number	% of Sites Visited	Number	% of Sites Visited
<b>Productive Work</b>	67	23%	37	10%
<b>Wasted Work</b>	162	56%	223	63%
<b>Other</b>	60	21%	93	37%

**Table 6.4: Amount of work (from Browser History)**

We also observed that most subjects made their judgment of a site by viewing only the front page of the site. It makes sense that the “front door” page of a site should be both attractive and representative of the site as a whole – after all, the site author presumably designs it to be the initial impression a visitor to the site experiences. One can usually obtain a good idea of the amount and type of content available on the site as well as the production quality.

Subjects navigated to a total of 642 web sites (the total number of symbols in Figure 6.2 above), and looked at only the front page of over half the sites. And of the 240 sites that subjects selected for their collection of the best sites, subjects browsed only the front page of 91. Among the 402 sites that subjects rejected, 285 sites were rejected after browsing the front page. Overall, subjects viewed an average of 2.39 pages per site. Thus, we see that a subject’s initial impression of a site is extremely important. The quality of the front page is very representative of the quality of the entire site.

## **6.7 USER EXPLORATION STRATEGIES**

Most Yahoo subjects, lacking any better options, simply looked through the 60 sites in alphabetical order, reverse alphabetical order, or sometimes a combination of the two. A few users tried reading all the titles and annotations to make some judgments about sites before browsing them; however, many times their initial judgment of a site proved inaccurate once it was browsed, so even these users often reverted to exhaustive alphabetical search. Of course, users still read annotations as they proceeded

methodically through the list of sites, but did not rely on annotations to decide which sites to browse. Users also often browsed a few sites at random to try to cover a good sample of available sites.

TopicShop subjects used different strategies, ones that were informed by data in TopicShop Explorer. They spent more time prior to browsing sites on exploration within the TopicShop interface, sorting columns and watching how the arrangement of sites changed. They were mainly looking for sites that appeared near the top in multiple sorts. Many also attempted to get a rough idea of how sites were distributed in each column. Eventually, subjects tended to proceed by selecting a property they thought was useful and evaluating the first few sites in that column. After they exhausted the quality sites in that column, they would move on to another column and continue. Some subjects would also visit some sites at the low end of the columns to convince themselves that the profile data could be trusted.

As evidence of the influence of the TopicShop Explorer on user strategies, we looked at overlap in sites selected by subjects. TopicShop subjects arrived at a much larger common set of sites. The intersection for the eight TopicShop subjects across both topics was 9.5 sites, while the eight Yahoo subjects averaged an intersection of only 2.5 sites. It makes sense that TopicShop users would agree with each other quite a bit, even more than they agreed with the experts, since they relied on the same data, i.e., profile features, and tended to pursue the same strategies for selecting resources.

To better evaluate the utility of TopicShop data, we created purely automated sets of the 15 best sites using the “gather from the top of the column” strategy. We defined six sets of sites mechanically: five of the sets consisted simply of the top 15 sites for each numeric site profile property, and the sixth consisted of the top three sites on each property.

Recall that the Yahoo subjects had an overall average expert intersection of 4.6 (out of 12). All the automated TopicShop strategies performed better, with an average expert intersection of 5.6. We found it surprising and noteworthy that a purely mechanical strategy using only automatically computed data could outperform human subjects who had to rely only on Yahoo’s site titles and annotations. Of course, TopicShop subjects, human subjects with the added utility of the TopicShop data, outperformed the automated strategies, with an average expert intersection of 8.4. (Again, we assume that the task time limit was a factor; with enough time to browse and evaluate site content, we expect that people would outperform these mechanical strategies. Of course, who has enough time?)

We also observed a common, but unproductive strategy: nearly all subjects initially assumed that personal home pages (as determined by title and site location) would be of low quality. They supposed that they could immediately eliminate these sites and select only from the resulting, smaller subset. However, subjects quickly realized that this was not true – after visiting a few personal pages, they found that some were of quite high quality, so subjects abandoned this strategy.

As we observed subjects, we noted that about one-third of them kept their list of high-quality sites sorted by quality as they were constructing them. The other subjects selected 15 to 20 sites and went back to sort them later. Yahoo subjects had a very difficult time with this because, after looking at so many sites, they could not recall site content from just the title. Usually they had to revisit all of the sites a second time to order them. Subjects in the Abrams et al. [1] study complained that titles were inadequate descriptors of site content. Our subjects corroborated this result and were not able to recall enough information about a site by simply seeing the title. TopicShop subjects had a much easier time because they used thumbnail images to refresh their memory of different sites. Even the small icons in the details view were useful once a site was visited; they contained enough information (color, general layout, etc.) to trigger subjects' memory and help them remember site content.

## **6.8 DESIGN IMPLICATIONS**

Observations, interviews, and questionnaires suggested three significant design improvements to the TopicShop user interface. The first design improvement we considered for incorporation into TopicShop version 2 was better methods for creating subcategories of a topic. A key need that subjects in both interface conditions discussed was support for lightweight, flexible categorization. As subjects explore sites, they create rough mental groupings, using site similarity, site type (general information sites, specific subtopic sites, personal sites, etc.), or even site layout.

While the initial version of TopicShop lets subjects create folders and group subcategories of sites within folders, our observations of subjects showed that this seems to be too much overhead for users when they are starting out. Their mental groupings remain indistinct until they have encountered a sufficient number and variety of sites to enable them to articulate the organizing principle of their categories. Further, categories may be split or combined several times in early stages of exploration. And while the icons view (Figure 4.3) of TopicShop does support this flexible, lightweight categorization (and several

subjects used and liked it), this view hides the important site profile data from immediate view. We have two potential design solutions that could be added to TopicShop version 2 to better support categorization.

Linked views are one solution to this problem. One window would show the icons view and another would show the details view, with user selections of thumbnail images mirrored in both windows. Users then could spatially arrange sites as they form opinions about types of sites within a topic, while simultaneously sorting sites based on profile data. As users develop firm categories, they could create folders to hold sites within each category.

Another potential design solution is a color-coding scheme. Users could assign a color to a small informal grouping of sites and add others to the group as they continue to browse. Then, when sites are sorted, they would be sorted first by color (informal group), then whatever other property the user specified (e.g. inlinks, images, etc.). This would let users quickly create groups and still keep all sites in a single window. Again, when users are satisfied that a group really is a category, a folder can be created to contain it. This solution can easily be combined with the previous solution to give users more flexibility.

A second improvement to the design of TopicShop is to add two levels of annotations. One of the TopicShop design goals was to make it easy to reuse and share topical collections. Subjects affirmed that this was important. In support of this desire, all 16 subjects mentioned that they wanted to record comments about sites as they visited and collected them. Comments could be recorded for individual sites as well as user defined categories. These comments would be useful both to original users when they returned to their collections in the future and to people with whom they shared the collections. Their comments would explain why sites were selected, why they were considered to be of high quality, and what they were good for.

The final design change involves sorting techniques within TopicShop. Currently, sorting in TopicShop version 1 is limited to a single column, but subjects expressed a desire for several more powerful sorting techniques. First, they wanted to combine several columns, e.g., sorting by the sum of inlinks and out-links. Second, they wanted to be able to do a multi-level sort. For example, one might want to sort sites primarily by number of pages, then break ties by using another property, such as number of inlinks.

In the next section we will discuss the redesign of TopicShop describing which of these design changes we incorporated, and how they affected use of TopicShop's interface, based on another empirical study we conducted.

## CHAPTER 7: USER INTERFACE EVALUATION

We developed a new version of the TopicShop Explorer interface (shown in Figure 7.1), incorporating the design changes described in the previous chapter. There are 4 main components to this interface. The first is the *Work Area*, an initially blank space, where users can drag selected sites for further investigation. The *Site Profiles window* displays all detailed information that our crawler has collected for each site. The *Focused Site* at the top left corner of the screen shows a large thumbnail image of the last site that a user clicked on. The final section of the interface is the *Folder Selection* area, where users can select which topic to display in the interface.

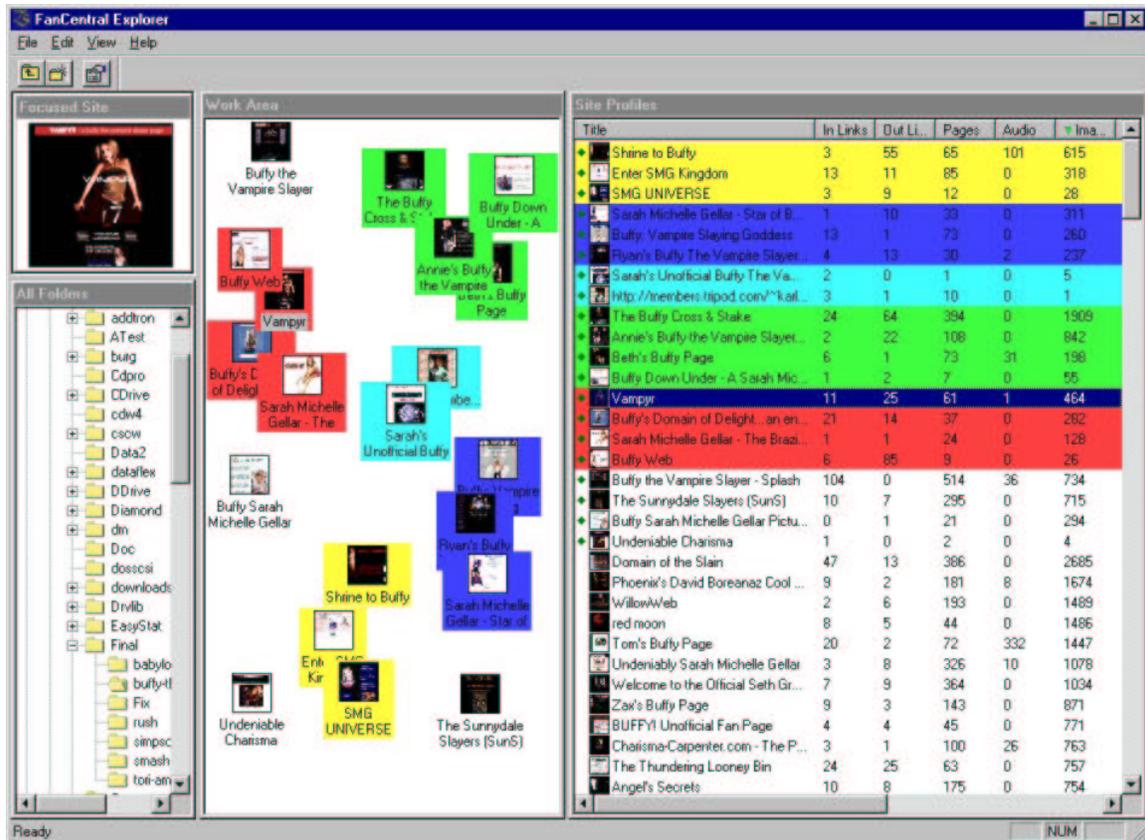


Figure 7.1: Revised version of TopicShop, based on results of pilot study

## 7.1 LESSONS LEARNED

Like all artifacts, the initial version of the TopicShop Explorer embodied claims about how users will conceive and carry out their tasks [21]. With its two separate windows for exploring site details and for organizing icons into groups, only one of which could be visible at a time, it embodied a claim that the tasks of site evaluation and organization must be carried out separately. Further, it assumed a single data set (the collection of all topic-relevant items), which could be manipulated in two ways (exploring site profiles or organizing by spatial grouping). The pilot study revealed problems with both implicit claims.

First, users wanted to organize items without losing sight of detailed information contained in 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.*

The interface must allow users to integrate the two tasks of site evaluation and organization.

Second, users preferred to group sites by spatial organization rather than by creating explicit folders. While the icons view supported this, the resulting groups were not first class objects. We wanted to explore spatial techniques to make it very easy to create and manipulate groups.

Third, we realized that most items in a collection never would need to be organized, because users would not select them as worthy of further attention. Thus, rather than supporting a single collection, a better design would support two data sets. Users can evaluate the initial, machine-generated collection and select promising items. Organization will only be done for selected items.

This also has implications for the nature of task integration. Users must be able to explore within groups they have created; for example, some users selected fairly large sets of similar sites, say ones that contained multimedia information, then wanted to keep only the best of these sites and throw the rest away. To do this, the interface should make it easy to sort within a user-defined group, e.g., to find multimedia sites with the most in-links or largest number of pages.

Fourth, site recall could be improved by including more graphical and textual information. 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.*

Fifth, the state of the user's task must be manifest. Most important, it had to be clear which items in the initial collection users had already evaluated and which they had not. Unevaluated 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...*

## 7.2 TOPICSHOP DESIGN ITERATION

Results and comments from the prior study guided us in designing an interface intended to more effectively address users' needs for topic management. Major changes in the second version of TopicShop include the following:

*Two always visible, linked views support task integration and a cleaner definition of each task.*

In an attempt to assist users in dealing with the overwhelming number of web sites available on any given topic, we provided site profile data and a work area for organizing sites, keeping both visible at all times. Items in the initial collection are displayed in the Site Profiles window, and the Work Area is initially empty (unlike Figure 7.1, which shows results of a subject from the main user study). As users discover sites that they are interested in, using the Site Profiles view, they select them simply by dragging and dropping them in the Work Area. Since icons are created just for selected items, the Work Area is uncluttered and provides a clear picture of sites users care about.

*“Piling” icons makes it easy to create first-class groups by spatial arrangement.*

Users seem to have the desire to group things spatially by making piles [73]. “Piling” is an easy way to allow a lightweight form of categorization because users are not required to create anything to contain the new category or to name it; they simply arrange thumbnail images in the Work Area by dragging icons. As users find sites that they feel are similar, they can arrange the sites to be close together in the Work Area window. When a user positions one icon “close enough” to another, a group is automatically formed. (How close two icons must be before a pile is formed is a system parameter, set by default to occur just when their bounding boxes touch.) Each group is assigned a color. As the views are linked, both the group of icons in the Work Area and the features for sites in that group in the Site Profiles window are displayed using the color as a background. Then users can add additional sites to the grouping as they visit similar sites. After the user is confident that their temporary category contains enough sites to be considered a sensible category for the topic, they can assign a meaningful name to it. One of the columns in the details view contains the user's category information so that sites can be sorted by category. 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.

Multi-level sorting is a useful operation that can be applied to a pile; it also illustrates how linked views support task integration. In the Site Profiles view, users can reorder sites based on primary and secondary sort keys. Users commonly sorted first by the groups they defined and then by some additional feature, such as in-links or number of pages. This lets users evaluate and compare sites within a single group. Figure 7.1 shows just such a sort.

*Visual indicators make the task state apparent.*

Any site included in the Work Area is marked with a green diamond in the Site Profiles 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. 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.*

The Focused Site window (upper left of Figure 7.1) displays the most recently clicked-on site. This large thumbnail image of the site is now displayed on the main screen to give users a more prominent view of the layout of the currently selected site. This is in direct response to users' claims that a large preview of sites was extremely useful but was too time-consuming to use in the initial TopicShop interface. It also serves as a memory aid to help users quickly remember additional details about the site.

Users can create textual annotations for piles or individual sites in the Work Area. Providing two levels of annotations allows users to describe groups they have formed and also give an indication of what type of content can be found on a given site. Annotations are useful as individual memory aids, but also allow users to personalize their collection of sites by adding comments to share with other users.

The interface also allows more customization by the user. Users have the option to show or hide any of the views they would like (Work Area view, Site Profiles view, Focused Site view, etc.). Subjects in the pilot experiment performed the task of topic management in many different ways. To support different approaches users take in evaluating and maintaining a collection of sites, we wanted to provide an interface that a user can tailor to their own specific needs. In the revised TopicShop, columns of displayed data can be hidden as well as moved to any position that the user desires. This way, columns that are important to a

user can be displayed first, while less relevant data for the user's crawl can be moved to the end or hidden all together. More user feedback has been integrated into the interface for common operations such as drag & drop, selection, and mouse movement. There were quite a few situations where a user had accomplished an operation and did not even realize it because of the lack of useful feedback.

### **7.3 EXPERIMENTAL DESIGN**

The second experiment was similar to the pilot study but was larger in scale and was redesigned in light of lessons from the pilot study. One major change from the pilot study was due to the fact that the topic collections were much larger, ranging from about 90 to over 250 sites. Since experts are required to comprehensively browse each site while establishing their ratings, we wanted to limit the number of sites experts rated to about 40. It would be unrealistic to expect experts to rate all the sites. It was not even possible for experts to rate all the sites that any subject selected, because this subset was also too large. However, we were able to come close. We chose sites for experts to rate by including first all the sites selected by multiple subjects and then a sample of sites selected by a single subject (A more precise explanation is provided later in section 7.5). Of course, this means that the order of the two phases was reversed from the pilot study. We first gathered user data and used those results to decide which sites to present to experts for rating in phase two.

This main user study consisted of two tasks that were performed simultaneously, a selection task and an organization task. Again, the experimental design has two levels of interface (TopicShop Explorer and Yahoo/bookmarks), but covers five different topics rather than two. As discussed before (section 5.2.1), analysis of the Magellan search data showed that entertainment was a very popular category that users searched for on the web, so we again selected topics from the domain of popular entertainment, including the television shows Babylon 5, Buffy The Vampire Slayer, and The Simpsons, and the musicians Tori Amos and Smashing Pumpkins. The experimental design was a 2x5 between-subjects design (see Table 7.1). Because results from the previous experiment were statistically significant with only four subjects per cell, we used the same number of subjects per condition for a total of 40 subjects.

<b>2x5 Experimental Design</b>		<b>Interface</b>	
		<i>TopicShop</i>	<i>Yahoo</i>
<b>Topic</b>	<i>Babylon 5</i>	4 subjects	4 subjects
	<i>Buffy the Vampire Slayer</i>	4 subjects	4 subjects
	<i>The Simpsons</i>	4 subjects	4 subjects
	<i>Smashing Pumpkins</i>	4 subjects	4 subjects
	<i>Tori Amos</i>	4 subjects	4 subjects

**Table 7.1: Main study experimental design**

We again obtained collections from Yahoo and then applied our web crawler to obtain site profiles and thumbnail images for use in TopicShop. For this experiment, we configured the crawler to start from the set of sites found on a Yahoo page for each topic, but this time we configured the crawler to crawl beyond the initial set and include any new sites found during the crawl. Because Yahoo is a human-generated index, many of the newer web sites for a topic will not be displayed in the Yahoo index. Our crawler has the ability to find the newest sites before they show up in Yahoo. Topic experts also evaluated the sites that our crawler discovered beyond the initial set of seed sites. This allowed us to evaluate the quality of recent sites that have not yet been added to Yahoo. However, in the experiment, the sets of sites were still kept the same and limited to only the sites that appear on the Yahoo page. By once again maintaining the same data sets across the two interfaces, we could evaluate the efficiency and effectiveness of each interface.

The experts' task was to rate a collection of web sites derived from the sites selected by users. We had 16 experts evaluating sites with 4 experts for the Simpsons and 3 experts for each of the other four topics. This time we decided it would be both easier for them and more informative for us if experts rated the quality of sites on a scale of 1 (worst) to 7 (best) instead of ranking them in order. Again, experts rated sites by filling out a web-based form; the form presented sites in a random order. And it gave no information other than the URL, so experts had to browse each site to judge its quality.

#### **7.4 PARTICIPANTS**

Participants for this study were students from Virginia Tech and were compensated ten dollars per hour for their time. Topic experts consisted of upper level graduate students and faculty from Virginia

Tech, as well as AT&T Employees. The topic experts received Amazon gift certificates for participating in our study.

The novice subjects for our study came from 13 different majors and were between the ages of 17 and 35. Graduate students made up 39% and the rest were undergraduates. The majority of subjects used computers and the web daily and were predominantly PC users, with a few UNIX users and one Macintosh user.

## **7.5 METHODOLOGY**

Again, we used a two-phase approach, except this time the order was reversed from the pilot study, to ease the task of the topic experts. For this study, sizes of the collections in Yahoo had grown considerably, ranging from 88 sites to 258 sites. Since it would be unrealistic to ask an expert to visit and accurately rate that many sites, we culled the list of sites evaluated by experts using results from the first phase to select which sites experts rated. A more detailed explanation of site selection method is provided below.

Subjects were assigned randomly to one of the ten conditions (2 interfaces, 5 topics). The study began with a pre-questionnaire to gather some demographic information from the users. We then had users read instructions on the web explaining how to use TopicShop or Yahoo, depending on their assigned experimental condition. TopicShop subjects were shown its basic interface features and taught how to collect and organize sites by dragging and dropping icons in the Work Area. Yahoo subjects were shown a sample list of sites and taught how to collect sites by bookmarking and how to arrange them into categories. After answering any questions they had about their assigned interface, we had users complete a short practice task to get them familiar with their interface and ensure that they were comfortable with collecting and organizing sites.

In phase one of this study, the task was to collect the 15 “best” (as defined previously in section 6.1) sites and organize them into logical groups, with descriptive group labels, as they were collected. Since ranking the best sites was sometimes difficult and many times arbitrary in the pilot study, this time subjects were simply asked to collect sites and not worry about their relative rank. To complete this task, subjects utilized any information provided in their interface (for Yahoo: title and annotation; for TopicShop: site profiles and thumbnail images) along with site content.

In the pilot experiment, subjects were given a soft time limit of 45 minutes. They were warned at five-minute intervals when they were getting close to the time limit and were encouraged to attempt to finish within that time. This forced subjects to quickly choose sites they had already seen to fill their list of 15 quality sites. As a result, most subjects finished the task in approximately 45 minutes. The difference in task time between TopicShop and Yahoo subjects was very small in the pilot study. In the final experiment, no time limit was given for the task. Subjects were free to take as long as they needed to evaluate the sites for their topic.

The task ended when subjects were satisfied with their collections of sites. Subjects then completed a short questionnaire. Finally, we conducted an informal interview to reveal strategies subjects used to perform the task, their reactions to the interface, and what would have helped them complete the task more effectively.

In phase two, to collect expert ratings, we gathered three experts for each topic and asked them to fill out a short questionnaire detailing their interest in the topic and self-rated knowledge of the topic. They were then given an instruction sheet containing a description of their task and a definition of quality for our purposes. Experts in each topic were given a list of web site titles for approximately 45 sites for their topic in random order and asked to exhaustively browse the list of sites rating them on a scale of 1 (worst) to 7 (best). When they were finished, they filled out a final questionnaire to give feedback about the task and any problems they ran into.

We asked the experts in the pilot study to look at a set of 60 sites, selecting 20 and ranking them by quality. This turned out to be a difficult and time-consuming task. As mentioned before, we simplified the expert task in this second experiment, by reducing the set of URLs that experts were asked to look at, and instead of having them rank the best sites, we simply had them assign a rating to each site they visited. The set of sites presented to experts consisted of four subsets: URLs selected by multiple subjects, URLs selected by a single subject, URLs selected by no subjects, and URLs discovered by our crawler. We included all URLs that were selected by more than one subject in the main user study because, according to our subjects, these were the best sites. This is analogous to the standard information retrieval theory that “good” items are very likely to be in the intersection. A small random sampling of URLs that were not chosen by any of our subjects was included so we could test our hypothesis that sites selected by subjects

would be given the best expert ratings. We included 5 URLs that were discovered by our crawler and then randomly added sites from the other two groups (URLs select by one subject and URLs selected by no subjects) in a ratio of 2 to 1 until the set was approximately 45 URLs. Table 7.2 shows exact sizes and breakdown of the expert sets for each topic, indicating how many were selected from each group and the total size of the original sets when subsets were selected randomly.

<b>Topic</b>	<b>Number of Sites</b>	<b>Multiply selected Sites</b>	<b>Singly selected Sites</b>	<b>Sites not chosen</b>	<b>Discovered Sites</b>	<b>Total expert dataset</b>
<b>Babylon 5</b>	173	28	8/42	4/104	5	45
<b>Buffy</b>	258	29	8/39	4/190	5	46
<b>Simpsons</b>	210	21	12/38	6/151	5	44
<b>Smashing Pumpkins</b>	95	33	6/16	3/45	5	47
<b>Tori Amos</b>	88	36	4/19	2/33	5	47
<b>Total</b>	824	147	38/154	19/523	25	229

**Table 7.2: Number of Sites in expert sets. (In cases where we randomly selected a subset of sets, we use the notation x/y to show that we selected x sites out of a possible y.)**

## 7.6 DATA COLLECTION AND ANALYSIS

The pilot experiment contained a flaw in the way the browsers were set up that might have affected user task time. When browsing sites on the web, network lag can be introduced due to Internet congestion or downed servers and routers. In the pilot experiment, this lag could have affected the task time of some subjects since our browsers simply loaded pages from the original server location on which each site resided. This did not seem to create a noticeable delay, but still might have had some minor impact on the pilot results. For the final experiment, we installed a cached server that pre-loaded all web sites for the study to a local hard drive, providing a frozen snapshot. This way, we ensured that all subjects had the same page load times and the times were consistent across all subjects.

In addition to data collected for the pilot experiment (user selections, task time, browser history, etc.), we also collected some additional interface data in this second main study. By adding interface instrumentation, we collected log files describing usage of the application and its individual interface components. In this way we could find what features of TopicShop are used most often. A user's behavior using the interface can be tracked more easily through these logs. In the pilot experiment these data were available but would require searching through hours of videotape.

### **7.6.1 Phase One: User Study**

The main user study was automated so that data collection would be easier than it was for the pilot study. Batch scripts were running that automatically timed tasks, logged data, and transferred files. For each user, we collected a list of sites that the user selected, either a bookmark page for the Yahoo condition or an icon list for TopicShop. The site categorization that users created was also derived from these two files. In addition, a snapshot of a browser history file was written for each user. Two log files were captured to time users and watch what they were doing during the course of the task. One was a system-level log that registered which windows were active at all times, and the other was an application log showing exactly what users were doing by logging where in the browser or TopicShop window users clicked. The two questionnaires that subjects filled out before and after the study were web-based forms that recorded results to files. The pre-questionnaire gathered basic demographic information along with computer and Internet search experience. Subjects provided their strategies, confidence with results, and interface feature comments in the post-questionnaire.

### **7.6.2 Phase Two: Expert Ratings**

Expert ratings were collected remotely using forms on the web. This allowed experts to rate sites at their leisure and do a more comprehensive job than they might have in a lab setting. We once again collected demographic information with a pre-questionnaire and also asked about their perceived familiarity with the topic they were evaluating. Then the site rating scores from 1 to-7 were collected followed by a post-questionnaire that gathered information including how they went about doing their evaluation and how long they spent doing it.

## **7.7 QUANTITATIVE RESULTS**

### **7.7.1 Expert Metrics**

Since we collected a numeric rating of each site viewed by experts, we have various applicable methods of using these expert data in our analysis. Below are the two main expert metrics that we used in analyzing our results: *expert average* and *majority score*. We of course looked at other metrics, but since

results were comparable for all metrics, we chose these two because they are easy to understand and are the most logical for the types of analysis presented in this section.

The first metric that we used was a straightforward *average* of the three experts. This is simply an average rating from 1 to 7 and is easy to use in calculations. The other metric, *majority score*, is a bit more complicated. Majority score can be defined as the percentage of experts that rated the site 5 or higher. Since humans tend to apply different scales when rating quality, we wanted to collapse the 1 to 7 ratings from our experts into two bins: good and bad. We decided that a rating of 5, 6, or 7 was considered to represent a good site and anything below 5 was deemed a bad site. So by counting the number of experts that rated a particular site as good and dividing by the total number of experts, we get a ratio of how many experts considered a site to be of high quality. We considered URLs with a majority score greater than one half to be high quality and less than that low quality, according to our experts' ratings. This is equivalent to saying that high quality sites (those with a majority score of one half) have been rated as "good" by more than half the experts for that topic. Note that for four of our five of our topics, the majority score must be 2 out of 3 experts for a site to be considered high quality but for the other topic, the Simpsons, the majority score must be 3 out of 4 experts.

### **7.7.2 Finding Quality Sites**

One of the main goals of our study was to help users find better quality sites. The first analysis we performed looked at the quality of sites found in each interface condition (TopicShop or Yahoo). Using the expert ratings, we computed an average expert majority score for the set of URLs selected by any subject. Recall that experts only rated a subset of the singleton URLs, which were selected by only one subject (e.g., only 8 of the 42 Babylon 5 singletons were rated). The average expert majority score includes a normalized expert score for any un-rated (singleton) URLs that were part of a subject's collection. The normalized expert score is based on the ratings of URLs that experts did judge. For each topic, we computed the normalized expert score as the average of all expert-rated singly-selected URLs. When computing the average expert majority score, we substituted in the topic-specific normalized expert score for each un-rated singleton URL, rather than using zero to indicate that the URL was not rated. This way a

subject's average expert majority score was not penalized because experts were unable to rate their selected URLs due to time constraints. Table 7.3 shows a summary of scores for each topic.

<b>Topic</b>	<b>TopicShop</b>	<b>Yahoo</b>	<b>% Increase (TS over Yahoo)</b>	<b>Maximum Possible Score</b>
<b>Babylon 5</b>	0.52	0.38	36.11%	0.91
<b>Buffy the Vampire Slayer</b>	0.50	0.27	80.65%	0.84
<b>Simpsons</b>	0.40	0.22	80.75%	0.83
<b>Smashing Pumpkins</b>	0.38	0.25	53.48%	0.55
<b>Tori Amos</b>	0.49	0.26	92.73%	0.75
<b>Average</b>	<i>0.46</i>	<i>0.28</i>	<i>65.72%</i>	<i>0.78</i>

**Table 7.3: Average expert majority scores for TopicShop and Yahoo users**

The scores presented are majority scores, which show the percentage of sites in a subject's collection that would be rated good by the experts. Overall, TopicShop subjects were able to select 66% more high quality sites than Yahoo subjects. The majority score for TopicShop subjects was 0.46 and only 0.28 for Yahoo subjects (2-way ANOVA, interface factor  $F(1,30)=36.94$ ,  $p<.00001$ ). The topic factor was significant ( $F(4,30)=2.84$ ,  $p<0.06$ ), but the interaction was not ( $F(4,30)=0.49$ ,  $p<.74$ ). There was some variability across topics, which may be due to differing amounts of quality content about each topic. Most topics contained only a small number of high quality sites, so the total expert majority scores for a subject's collection of 15 sites must include some lower quality sites that were not rated good by a majority of the experts.

The last column in Table 7.3 shows the maximum possible average majority score for the best 15 sites in each topic. Since there were so few good sites, it is worthwhile to look at the top 5 and top 10 sites in each subject's collection to see how many of the good sites subjects selected were rated high in quality by the experts. This gives an indication of how many of the limited quality sites subjects are able to find using each interface. Table 7.4 shows results of this analysis.

<b>Topic</b>	<b>Sites</b>	<b>TopicShop</b>	<b>Yahoo</b>	<b>% Increase</b>
<b>Babylon 5</b>	5	0.98	0.82	20.41%
	10	0.70	0.53	31.25%
<b>Buffy the Vampire Slayer</b>	5	0.92	0.62	48.65%
	10	0.63	0.34	85.37%
<b>Simpsons</b>	5	0.88	0.55	59.09%
	10	0.50	0.31	63.27%
<b>Smashing Pumpkins</b>	5	0.80	0.65	23.08%
	10	0.55	0.37	50.00%
<b>Tori Amos</b>	5	0.90	0.52	74.19%
	10	0.68	0.34	97.56%
<b>Average</b>	5	0.90	0.63	42.06%
	10	0.61	0.38	65.49%

**Table 7.4: Majority score for Top 5/Top 10 user sites**

Limiting the analysis in this fashion shows that 90% of TopicShop subjects' top 5 sites were rated good by experts, compared to 63% for Yahoo subjects' sites. When looking at the top 10 sites in each subject's collection, the percentage of good sites were 61% and 38% for TopicShop and Yahoo subjects, respectively. Again, TopicShop subjects found more of the better sites than Yahoo subjects. Two-way ANOVAs were run to check statistical significance. For the majority score of the top 10 sites, the interface factor was significant ( $F(1,30)=21.37$ ,  $p<0.00005$ ), but the topic factor and the interaction were not significant (topic factor:  $F(4,30)=1.94$ ,  $p<0.13$ ; interaction:  $F(4,30)=0.43$ ,  $p<0.79$ ). The analysis for the majority score for the top 5 sites was similar (interface:  $F(1,30)=28.37$ ,  $p<.00009$ ; topic:  $F(4,30)=2.08$ ,  $p<0.11$ ; interaction:  $F(4,30)=0.84$ ,  $p<0.51$ ).

<b>Topic</b>	<b>TopicShop</b>	<b>Yahoo</b>	<b>% Increase</b>
<b>Babylon 5</b>	7.00	5.75	21.74%
<b>Buffy the Vampire Slayer</b>	7.25	3.50	107.14%
<b>Simpsons</b>	6.50	5.25	23.81%
<b>Smashing Pumpkins</b>	8.50	5.00	70.00%
<b>Tori Amos</b>	7.75	3.00	158.33%
<b>Average</b>	7.40	4.50	76.20%

**Table 7.5: Intersection between users selections and top 15 expert-rated sites**

We performed another analysis to look at the number of sites from a subject's collection that intersected with the top 15 expert-rated sites. This metric is a bit more straightforward than majority scores presented above. For each topic, we can generate a 'good set' of sites by sorting all expert-rated sites and selecting the 15 best sites. We can then measure the quality of a user's collection by looking at how many

of their selected sites match with the best sites according to experts. Table 7.5 shows the average number of sites that intersected with the ‘good set’ for each interface condition. Of the 15 best sites, TopicShop subjects found 7.4 on average, while Yahoo subjects found only 4.5 sites. This is a 76.2% increase in quality for TopicShop users. Notice that the relative benefit of TopicShop over Yahoo varies from one metric to another (i.e., 76.2% better for the intersection and 65.72% better for the majority score analysis), because in each of these analyses we looked at the data in a different way. We computed a 2x5 two factor ANOVA on this metric since this metric was also computed in the pilot study. Results showed that the main effect of interface was significant ( $F(1,30)=18.55, p<.0002$ ). Topic and the interaction between topic and interface were both insignificant (topic:  $F(4,30)=0.656, p<0.627$ ; interaction:  $F(4,30)=1.097, p<0.376$ ). Once again, since topic is insignificant and the interaction also has no effect, we report, below, the remaining statistical results using pooled independent means t-tests.

### 7.7.3 User Search Efficiency

It is not only important to find quality sites, it also is important to find them quickly. The time and effort that subjects take is important when evaluating interfaces designed to help users search the web. The next two sections show analyses done on a user’s time and effort spent selecting the sites in their collections. Recall that in this experiment, subjects were given as much time as they needed. TopicShop subjects were able to complete the task in a little over a half an hour, but Yahoo subjects typically took almost an hour. The average time to complete this task for TopicShop subjects was 37.7 minutes. This is 28.2% faster than Yahoo subjects, who took over 52 minutes on average. Table 7.6 shows all the task times for each topic in both interface conditions. Statistical analysis shows that overall averages across the five topics were significantly different. (pooled independent means t-test,  $t(38)=-4.219, p<0.00007$ ).

<b>Topic</b>	<b>TopicShop</b>	<b>Yahoo</b>	<b>% Diff</b>
<b>Babylon 5</b>	41.45	51.45	19.44%
<b>Buffy the Vampire Slayer</b>	41.23	61.77	33.26%
<b>Simpsons</b>	33.36	54.05	38.38%
<b>Smashing Pumpkins</b>	35.55	43.71	18.66%
<b>Tori Amos</b>	36.87	51.68	28.67%
<b>Average</b>	37.69	52.53	28.25%

**Table 7.6: Task Time (in minutes)**

We also calculated the time it took each subject to find the 5 sites from their collection with the highest expert ranking. Since most people do not want to search through more than a few web sites to find the information they are looking for, it is important to analyze how quickly an interface allows subjects to find the best material. These results are summarized in Table 7.7.

<b>Topic</b>	<b>TopicShop</b>	<b>Yahoo</b>	<b>% Diff</b>
<b>Babylon 5</b>	13.12	15.93	17.64%
<b>Buffy the Vampire Slayer</b>	9.20	24.52	62.48%
<b>Simpsons</b>	16.64	25.78	35.46%
<b>Smashing Pumpkins</b>	7.99	9.05	11.72%
<b>Tori Amos</b>	15.46	20.59	24.91%
<b>Average</b>	<i>12.48</i>	<i>19.17</i>	<i>34.90%</i>

**Table 7.7: Time to visit Top 5 sites**

Again, TopicShop subjects were able to find the quality sites faster. TopicShop subjects selected their 5 best sites in an average of 12.48 minutes, which is 34.9% faster than Yahoo subjects who took 19.17 minutes to select their 5 best sites. These average times were statistically significant (pooled independent means t-test,  $t(38)=-2.356$ ,  $p<.01$ ). As shown before in Table 7.4, the quality of the 5 best sites found by TopicShop subjects was also much higher than the 5 best sites found by Yahoo subjects.

The task for the first phase of this experiment included selection and then organization. Most users intertwined these two sub-tasks together by browsing and selecting a small number of sites and then incorporating them into their overall site organization. Another interesting analysis is to look at the percentage of time users spent performing each sub-task.

<b>Topic</b>	<b>TopicShop</b>		<b>Yahoo</b>	
	Browsing	Organizing	Browsing	Organizing
<b>Babylon 5</b>	82.73%	17.27%	61.61%	38.39%
<b>Buffy the Vampire Slayer</b>	82.82%	17.18%	71.32%	28.68%
<b>Simpsons</b>	73.90%	26.10%	71.32%	28.68%
<b>Smashing Pumpkins</b>	83.66%	16.34%	67.08%	32.92%
<b>Tori Amos</b>	84.97%	15.03%	52.28%	47.72%
<b>Average</b>	<i>81.62%</i>	<i>18.38%</i>	<i>63.65%</i>	<i>36.35%</i>

**Table 7.8: Percentage of time spent Browsing/Organizing**

The results in Table 7.8 show that TopicShop was able to shift subjects' time from organizing to browsing and selecting. This suggests that organization is much easier in TopicShop and users can more

efficiently categorize their collections of sites, freeing up more of the time spent on this task for browsing. TopicShop subjects browsed sites on their topic for 82% of the time they participated in the experiment, while Yahoo subjects spent only 64% of their time actually viewing content about their topic (pooled independent means t-test,  $t(38)=-4.893$ ,  $p<0.00009$ ). Note that this is a slightly different metric than calculated in the pilot study. Instead of looking at the time a user spent using the interface versus the time they spent using their browser, here we compared the time they spent viewing content in the browser or interface with the time spent organizing their selected sites.

#### 7.7.4 Required Effort

One indication of the amount of effort required of users to find quality sites is the number of sites that they must visit to find a set acceptable to them. This roughly equates to the amount of work that users performed in completing this task and also gives an indication of how much of that work was wasted effort. Clearly if a user is trying to find 15 sites and must look at 50, they are wasting a lot of their time sifting through low quality sites. Instead, with TopicShop, we can reduce the number of sites users must investigate to find a representative set of quality sites. This is summarized in Table 7.9.

Topic	Total # of Sites	TopicShop	Yahoo	% Diff
<b>Babylon 5</b>	173	31.00	45.00	31.11%
<b>Buffy the Vampire Slayer</b>	258	25.00	50.00	50.00%
<b>Simpsons</b>	210	22.50	38.25	41.18%
<b>Smashing Pumpkins</b>	95	33.25	36.25	8.28%
<b>Tori Amos</b>	88	23.50	31.75	25.98%
<b>Average</b>	<i>164.80</i>	<i>27.05</i>	<i>40.25</i>	<i>32.80%</i>

**Table 7.9: Average number of sites browsed**

On average, TopicShop subjects only browsed 27 sites while Yahoo subjects visited 40. Comparing these numbers to the total number of sites, we can see that subjects in both conditions visited far less sites than the total available, but TopicShop subjects considered 32.8% less sites on average (pooled independent means t-test,  $t(38)=-4.788$ ,  $p<0.00001$ ). Since TopicShop gives users additional information about the sites, it is logical that they can rapidly find sites they think will be high quality and eliminate the need to visit a large number of low quality sites. Subjects in the Yahoo condition also

attempted to eliminate sites up front to avoid visiting them, but at best, they could only guess based on title and annotation.

### 7.7.5 User Categorization

From the organization sub-task, we had a collection of groupings that subjects placed on their collections of web sites. We analyzed categories that subjects made to assess whether or not they agreed with other subjects in their interface condition on which sites should be categorized together. In order to evaluate different categorizations that users made, we first looked at the size of the site intersections between users' selected sets. By looking at the number of sites users had in common with each other, we obtained a better idea of how many they might group similarly.

<b>Topic</b>	<b>TopicShop</b>	<b>Yahoo</b>	<b>% Diff</b>
<b>Babylon 5</b>	5.17	3.00	72.22%
<b>Buffy the Vampire Slayer</b>	5.00	3.17	57.89%
<b>Simpsons</b>	7.00	4.17	68.00%
<b>Smashing Pumpkins</b>	6.83	6.00	13.89%
<b>Tori Amos</b>	5.67	3.17	78.95%
<b>Average</b>	5.93	3.90	58.19%

**Table 7.10: Average site intersection among users**

The values presented in Table 7.10 represent the average number of pair-wise intersections within each topic and condition. These average pair-wise intersections were calculated by computing the set of sites selected by each pair of subjects within a topic and interface condition and then averaging the size of those sets across the 6 pairs of subjects (all possible pairs of the 4 subjects in each condition). On average, TopicShop users selected 6 sites that were also selected by other users and Yahoo users selected 4 such sites (pooled independent means t-test,  $t(58)=-4.256$ ,  $p<0.00004$ ). Unfortunately this set of sites to compare across subjects is fairly small. It would be nice to have more similar sites among users to begin investigating how they formed categories. But we can at least get an indication of how much agreement there was within topic categorization with a small intersecting set.

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

We first computed how much time subjects spent on the organization sub-task (by examining the log files). TopicShop subjects spent 18% of their total time, while Yahoo subjects spent 36% of theirs. Since TopicShop subjects spent less 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 discuss next.

We also computed the number of groups that subjects created. TopicShop subjects created 4 groups on average, and Yahoo subjects created 3. Thus, TopicShop subjects articulated the structure of the topic somewhat more. In addition, TopicShop subjects grouped nearly all of their selected sites (3% were left ungrouped), while Yahoo subjects left more ungrouped (15% were not grouped).

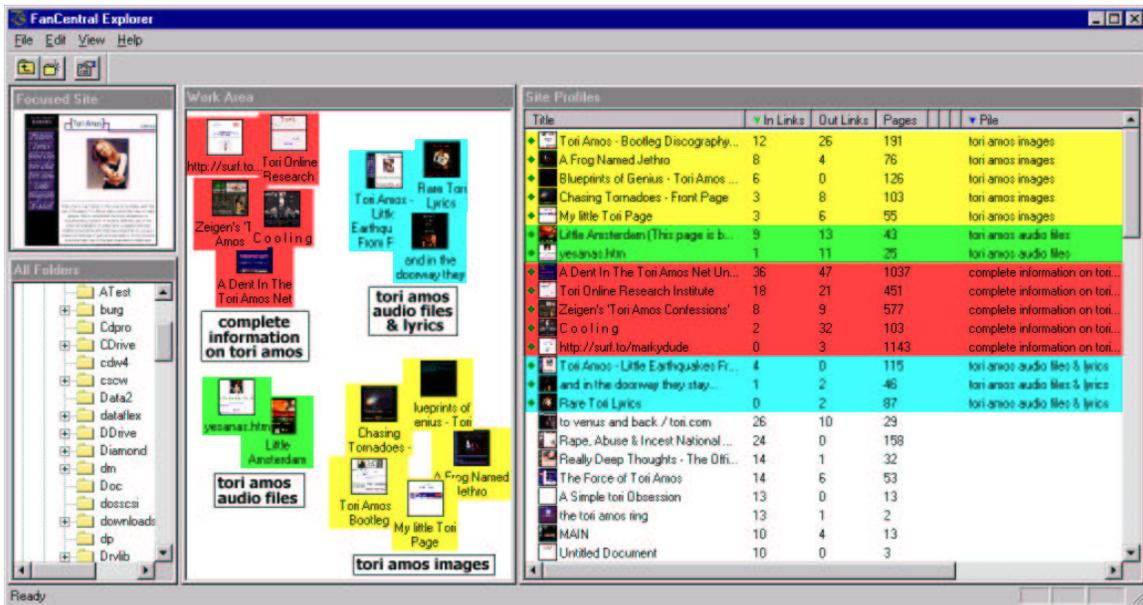
TopicShop subjects created more site annotations, thus making their collections more informative for their own use or for sharing with others. The experiment did not require subjects to annotate sites. Yet 10 of 20 TopicShop subjects did so, annotating a total of 15% of their selected sites. Two Yahoo subjects annotated a total of four sites. TopicShop subjects annotated sites using the group and site annotation features of TopicShop. Since Yahoo subjects were using the bookmarking feature of their browser, they were also able to annotate sites and groups directly in their bookmarks.

One way to gain insight into how groups are formed is to find out what percentage of sites users categorize similarly. This is a difficult issue to investigate; in general, it requires interpreting the semantics of groups. We computed a simpler metric; by looking at each pair of subjects within a topic and interface condition and then looking at each pair of sites that they have in common, we can decide whether they agree on their categorization, i.e., whether they both put the sites in the same group or in different groups. 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.

Topic	Users	1,2	1,3	1,4	2,3	2,4	3,4	Avg	%
Babylon 5	TSP	0.64	1	0.73	1	0.5	0.8	0.78	99.57%
	Yahoo	1	0	0	0.67	0	0.67	0.39	
Buffy the Vampire Slayer	TSP	0.79	0.5	0	0.73	0.83	0.7	0.59	33.46%
	Yahoo	0.67	0.33	1	0	0.33	0.33	0.44	
Simpsons	TSP	0.76	0.81	0.93	0.81	0.71	0.67	0.78	116.13%
	Yahoo	0.67	0	0	0.4	0.6	0.5	0.36	
Smashing Pumpkins	TSP	0.76	0.9	0.72	0.87	0.71	0.53	0.75	40.31%
	Yahoo	0.9	0.6	0.4	0.5	0.4	0.4	0.53	
Tori Amos	TSP	0.5	0.4	0.62	0.33	0.33	0.67	0.48	17.28%
	Yahoo	1	0.5	0	0.33	0.6	0	0.41	
Average	TSP							0.68	61.35%
	Yahoo							0.43	

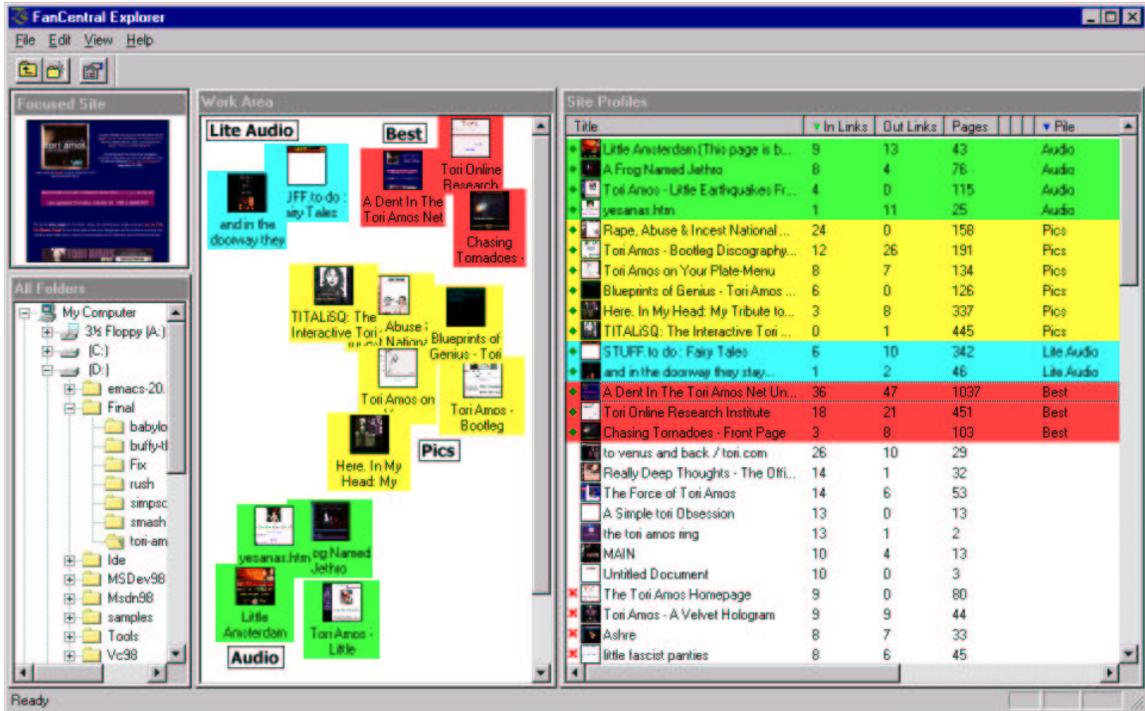
Table 7.11: Pairwise category agreement between users (1-4)

Table 7.11 shows details of this analysis. Each column represents the percentage of site agreement between any two subjects. On average, TopicShop subjects agreed on their categorization of 68% of the sites they had in common with other subjects, while Yahoo subjects agreed on only 43%. 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 (since the pairwise category agreements are not independent, a t-test was not provided for this analysis). The organizational facilities provided by TopicShop allowed users to easily group and evaluate sites in their collection. Subjects used these abilities along with the site profiles and therefore had an advantage over the Yahoo condition when forming groups.

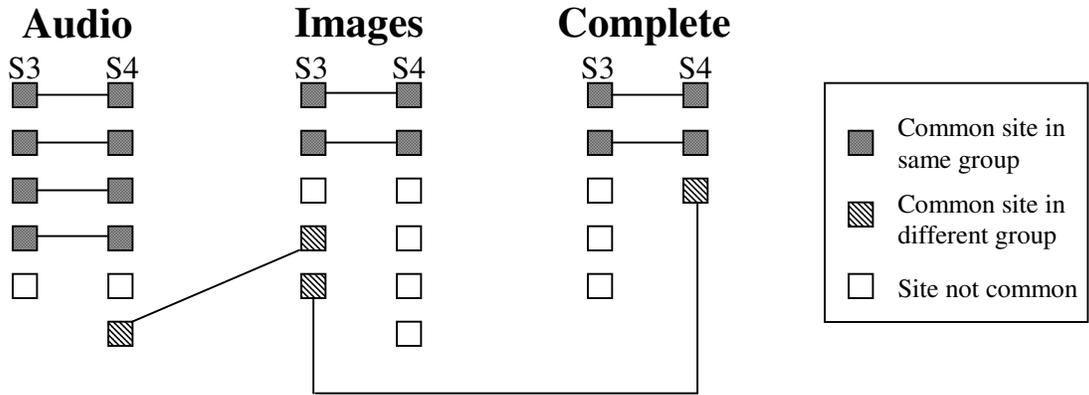


**Figure 7.2: A sample subject's categorization of Tori Amos sites. (Subject 3)**

Figure 7.2 & Figure 7.3 show the groups formed by two Tori Amos subjects (subjects 3 & 4 in Table 7.11). These two subjects had ten sites in common in their final set of selected sites and agreed on 67% of their categorization of pair-wise URLs. Each subject had two separate Links groups devoted to audio. If we combine each of these, we can see that they categorized 4 sites similarly as audio sites. Both subjects had two additional groups that formed. For subject 3 they were: 'tori amos images' and 'complete information on tori amos' , and for subject 4: 'Pics' and 'Best'. These two subjects categorized their sites into 3 semantically identical categories: audio sites, image sites, and a category representing the best available comprehensive sites on Tori Amos. Figure 7.4 is a representation of the 3 main groups, showing how these two subjects agreed on their categorization. They placed 8 of the 10 sites that the subjects had common in the same groups. But two of the sites that subject 3 classified as image sites were classified by subject 4 as audio and best. This is not surprising since most sites can be classified in a few different categories depending on a user's personal preference and interpretation.



**Figure 7.3: A second subject's categorization of Tori Amos sites. (Subject 4)**



**Figure 7.4: Groups for Tori Amos as created by subjects 3&4**

These 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. Further, they achieve these results in half the time Yahoo subjects devote to the task. We believe these results are because TopicShop makes grouping and annotation very easy, because of the rich information about sites that is available and remains visible while users organize sites.

### 7.7.6 Relationship between evaluation and organization sub-tasks

We also studied the relationship between evaluation and organization sub-tasks. The TopicShop Explorer allows these sub-tasks to be integrated, but does not force a user to perform them integrally. On the other hand, in the Yahoo/bookmarks condition, browsing sites and organizing bookmarks can only be performed as separate sub-tasks.

The log files contained data that let us quantify relationship between these sub-tasks. Each user action was timestamped, and we knew 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.

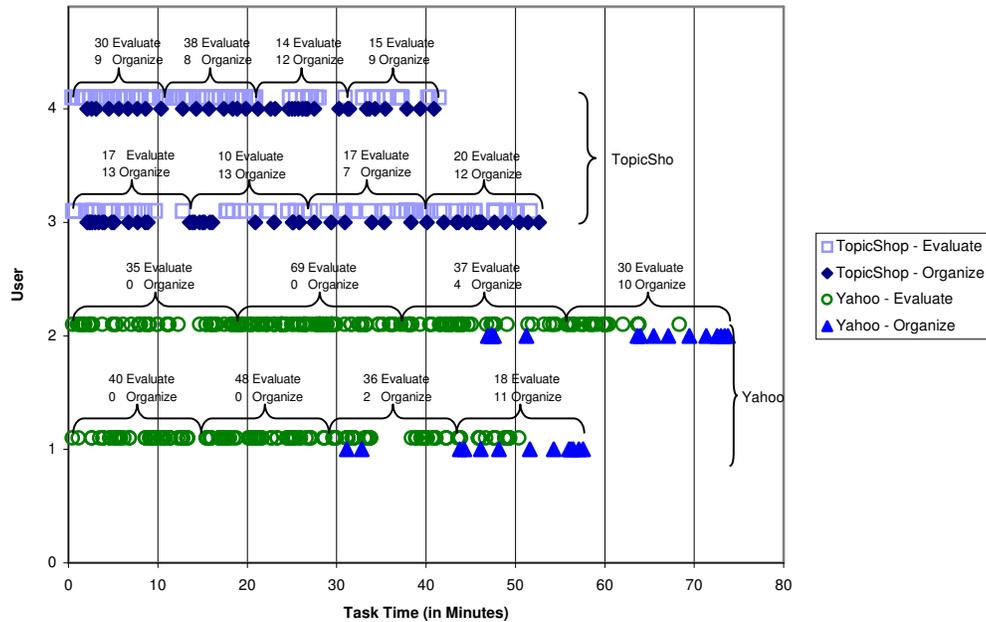
We computed how many actions of each type occurred in each quartile of the task, i.e. how many

occurred in the first 25% of the total time a subject spent on task, how many in the second 25%, etc. Table 7.12 shows results for organizational actions. First, it shows how much more organizational work TopicShop users did: 533 actions vs. 172. (And 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 sub-task integration; rather, it enables it. And when users had the choice, they overwhelmingly preferred integration of the sub-tasks of evaluation and organization.

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%
<b>Total</b>	533		172	

**Table 7.12: Distribution of organizational actions across time quartiles**

We also constructed detailed timelines of user activity. Figure 7.5 shows such timelines for two Yahoo and two TopicShop subjects. They provide vivid illustrations of the overall results. TopicShop users interleaved the two sub-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.



**Figure 7.5: 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**

### 7.7.7 Expert Ratings for Site Breakdowns

To reduce the amount of work our experts had to do, we selected a subset of sites on each topic to be presented to the experts, as explained above (section 7.5). There were four categories that sites in the expert set fell into: sites selected by multiple users, sites selected by one user, sites selected by no users, and sites discovered by our crawler (and shown to no users). Our expectation supporting this choice was that a site selected by more than one user would be better quality than a site selected by a single user and any site not chosen by any users would be the lowest quality. We validated this claim by looking at the average expert rating for each of the groups (a high expert rating is considered good).

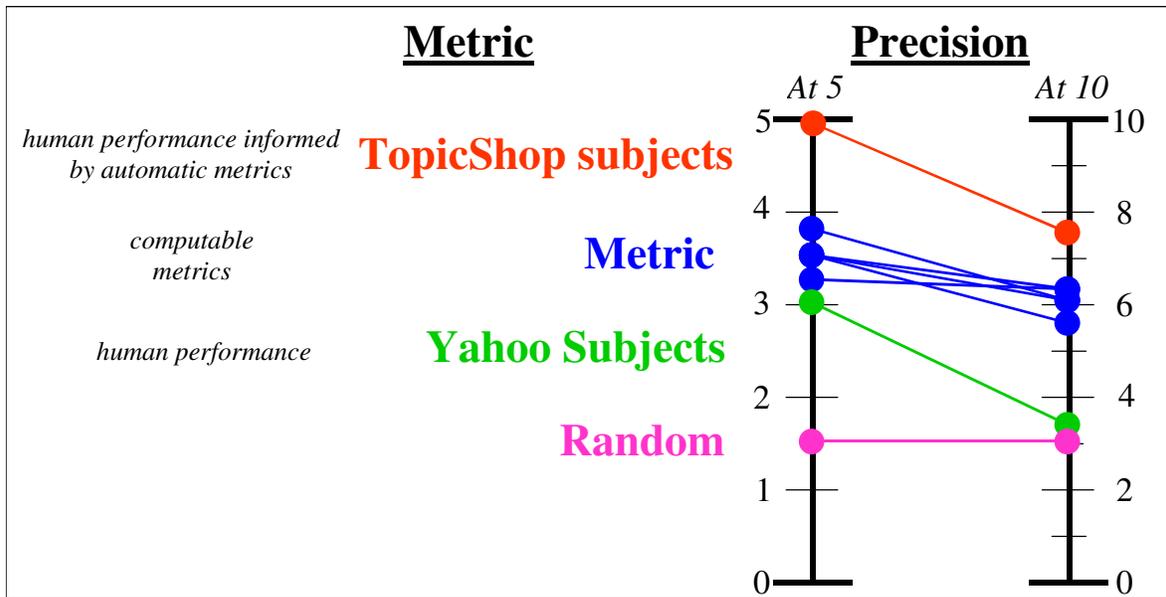
<b>Topic</b>	<b>Multiply selected Sites</b>	<b>Singly selected Sites</b>	<b>Sites not chosen</b>	<b>Discovered Sites</b>
<b>Babylon 5</b>	4.78	3.37	2.84	4.70
<b>Buffy</b>	4.28	3.58	1.75	4.20
<b>Simpsons</b>	3.91	2.33	2.33	3.80
<b>Smashing Pumpkins</b>	3.45	2.33	1.44	3.40
<b>Tori Amos</b>	3.95	3.16	2.12	3.93
<b>Total</b>	4.05	2.91	2.09	4.01

**Table 7.13: Expert scores of site categories**

As shown in Table 7.13, experts scored the multiply selected sites much higher (4.05) than the singly selected sites (2.91) and the sites not chosen (2.09). This means our expectation holds and we produced expert ratings on higher quality sites in each topic while filtering out lower quality sites. The last component of the set of sites given to experts was the discovered sites. Recall that our crawler, using the hyperlink structure of the web, found new sites that were not yet available on the Yahoo list. Each expert was given the top 5 of these sites, based on number of in-links, from their topic; resulting expert average scores are shown in the last column of Table 7.13. The experts rated these sites as high quality, almost as high as the multiply selected sites within each topic. Which shows that the best crawler discovered sites are on the same level as the best sites from Yahoo. This indicates that an automatic crawler can enhance human-generated collections significantly.

### **7.7.8 Comparing human performance to automatic metrics**

In Chapter 9 we present an analysis of the quality of the sites in our main study. For that analysis we had to compute how each metric that we collected in our site profiles ranked the collections of sites. We then looked at how many of the top 5 and 10 sites (precision) on those ranked lists were considered to be good by our experts. We compared that to how many of our subjects' first 5 and 10 sites were considered high quality. Figure 7.6 shows this comparison.



**Figure 7.6: Automated metrics compared to subjects' judgments**

The link and content metrics have a precision between .7 and .8 at 5 site and around .6 at 10 sites. Choosing sites randomly results in a precision of .3, so our automatically computed metrics performed more than twice as good as random choice. Strictly human performance (subjects in the Yahoo condition) resulted in a precision at 5, just under the automated metrics, but at 10, the precision of humans making selections from a large directory of sites, without much additional information, performed barely above random choice. On the other hand, automated metrics augmented with human judgment (subjects in the TopicShop condition using our site profiles) results in almost perfect precision at 5 and very high precision at 10. Link and content metrics are completely framed on the bottom by human performance and on the top by automated metrics plus human performance.

### 7.7.9 Questionnaire results

Using the two questionnaires given at the beginning and end of the experiment, we collected demographic information and computer background from subjects, and elicited feedback about how they performed the task and what features of their interface they found most useful in completing the task.

Several questions asked about web page bookmarking behavior. Ninety percent of our subjects primarily used bookmarks for storing their favorite links. The remainder built web pages and utilized the history features of their browser. Most users have a moderate size bookmark list, with 65% having less than 50 links and 11% with between 50 and 100. The final 23% kept over 100 links in their bookmarks and these users revealed that they did not organize their list very frequently. They used the bookmarking more as a history feature and used the chronological order of the sites for later recall. Most users (69%) keep their bookmark lists in an organized fashion, creating high level sub-groups for storing links.

Results from the Media Metrix [103] survey (discussed in Chapter 5), showed that overall, Yahoo was the most popular search engine followed by MSN, Netscape, and InfoSeek. We asked our subjects which search engines they used to find information on the web and the top search engine, used by 52% of our subjects, was Yahoo. In our main study, AltaVista was the second most popular search engine following closely behind Yahoo, with 50% of subjects using it. In comparison, AltaVista ranked seventh in the Media Metrix survey. Other popular search engines included InfoSeek (23% of subjects) and Excite (20% of subjects). The rest of the search engines (i.e. HotBot, Lycos, Google, Northern Lights, etc.) were used by less than 10% of our subjects.

In our pilot study, Yahoo subjects complained that they did not have enough time to complete their task. For the main study, we allotted extra time to ensure that subjects were not pressured to finish quickly. This seems to have solved the problem because 93% of the subjects said they had more than enough time.

Another question that we asked was whether or not subjects were confident that the sites they selected were the best set of sites to represent their topic. Subjects rated their confidence on a scale of 1 to 7, 1 being very confident and 7 being not confident at all. TopicShop subjects' average confidence was 2.25, compared to Yahoo subjects' 2.5. Recall from the pilot study that TopicShop subjects stated a 4.5 confidence level versus a 4.75 confidence for Yahoo subjects. Overall, subjects were more confident with their selections in the main study, probably due to the fact that they had more time. Once again, TopicShop subjects had slightly more confidence in their collection of sites because they had additional information about each site.

We looked at the usefulness of the TopicShop site profile data (in-links, out-links, title, URL, images, audio files, pages, and keyword relevance), and found similar results to the pilot study. Subjects ranked all site parameters from most useful (ranked as 1) to least useful (ranked as 8). By comparing average ranks of each parameter, we found that the top parameters once again included in-links (average rank of 3.2), title (average rank of 3.7), and number of pages (average rank of 4.2). The keyword relevance parameter represents the new parameter that categorized sites into domain specific types such as lyrics, episode guides, and discographies. This parameter was also ranked as one of the most useful parameters with an average rank of 4.0. The other 4 parameters had an average rank greater than 5. So, once again the number of endorsements (in-links, ranked first by 40% of the subjects) and the size of the site (number of pages, ranked first by 20% of the subjects) were two of the most important indicators of quality. In CHAPTER 9: we will confirm that subjects are accurate in their utility judgements by showing that both in-links and number of pages are good predictors of site quality. Fifty percent of subjects stated that they computed composite parameters using two or more site properties when making their decisions about which sites to visit first. We also asked Yahoo users to rate usefulness of the annotations, from most to least useful on a scale of 1 to 7. Their average score was 3.75, with only one subject rating the annotation very useful. Typically, annotation does not accurately represent site content and therefore is not useful in making judgements about a site.

TopicShop has four unique features that help users evaluate and organize web sites: annotation, coloring (grouping), sorting, and spatial organization. We asked subjects to rank the usefulness of these 4 features from most to least useful on a scale of 1 to 4 to see which feature helps the most. Sorting turned out to be the most useful feature, ranked first by 60% of all subjects and getting an average rank of 1.8. The other three features had an average rank of approximately 2.6. Since sorting gives users access to the site parameters mentioned previously, it is obvious that it would be more useful to users for selecting quality sites.

#### **7.7.10 Qualitative observations**

During the post-session interview we received many comments from users about the task they were performing and the interface they were using. Here are some comments made by TopicShop subjects commenting on 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.*

Yahoo subjects had many comments about the lack of information presented to them in their interface. They 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 (so as to be able to judge the worthiness).*

The spatial organization capabilities of TopicShop helped users in organizing their collections. In addition the memory aids provided in TopicShop give users a simple way to recall sites. The following are design implications we discovered during the experiment and some user quotes supporting them.

- **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].*

Another major advantage of TopicShop was the ability to seamlessly integrate the sub-tasks of organization and evaluation. Several comments showed that subjects appreciated this capability, as well as having the state of their sub-task made visible.

- **Linked views helped users integrate evaluation and organization sub-tasks. In particular, they could evaluate within groups they created.**

*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.*

## 7.8 DESIGN SUMMARY

The design of TopicShop was informed along the way by user comments and evaluation from the two studies. The initial version of TopicShop had some design flaws that were uncovered during the pilot test and resolved for the second version of TopicShop. Many of the incorporated design updates were validated as useful by comments and results from the main user study. In addition, we have evidence of the approval for these new features by both Yahoo subjects who indicated their desire for them and TopicShop subjects who confirmed their utility.

### 7.8.1 Spatial Organization in Work Area

We quickly realized that spatial organization was essential to the task of topic management. While this feature was provided in the original version of TopicShop, it was not incorporated with other areas of the interface. To preserve screen real estate, we forced users to select between two views: one for spatial organization and one for site investigation. Without automated cooperation in the interface between these two views, it was difficult to combine the work performed in each view. Users felt switching views

was too painful and confusing. One user commented, “I really want to organize the large icons, but don’t want to lose the detailed information, so I have to settle for the details view only.” Our solution to this problem was to implement linked views that mirrored each other’s activity. By providing side-by-side icons and details views, we enabled users to easily see critical site-specific data while simultaneously arranging sites in their collection. Linked views proved to be very effective in helping subjects complete the experiment. Comments from TopicShop users supported this design change; they felt the direct manipulation metaphor used for spatial organization facilitated grouping of sites.

The process used in completing the sub-tasks we presented assumed different forms for users in each condition. In contrast to Yahoo users, who tended to perform the separable sub-tasks of selection and organization serially, users of TopicShop integrated spatial organization into the browsing and selecting sub-task because the interface provided direct support for organization without interfering with the users’ main goal.

## **CHAPTER 8: COMPARISON OF STUDIES**

The main study validated the preliminary results discovered in the pilot study and has given us much more insight into the nature of the task and necessary features for a Topic Management interface. The structure of the main study changed in response to problems we saw in the pilot study, but many of the results were still analogous. In this section we show a comparison of the major results across the two studies.

### **8.1 RESULTS**

#### **8.1.1 Finding Quality Sites**

Helping users find quality sites was one of the most important goals of this work. In the pilot study, we measured quality by comparing each subject's set of web sites to the expert intersection for their topic. Recall that experts in the pilot study looked exhaustively at all sites for a topic and selected a set of 20 sites that they considered the best sites. The intersection of all experts was considered to be a set of high quality sites and we reported the number of user sites that intersected with it. For the main study, our initial collections of web sites were much larger and it was infeasible to have experts evaluate all of them. We chose a subset based on user selections and generated a combined score of experts' ratings for each site in the subset. By selecting the top 15 of these sites based on the average combined score, we get an expert

intersection set similar to the one in the pilot study. We once again used the size of the intersection between each user and this expert set as our quality metric. Table 8.1 shows intersection results from the two studies.

	Pilot Study			Main Study		
	TopicShop	Yahoo	% Increase	TopicShop	Yahoo	% Increase
<b>Buffy</b>	7.5	5.0	50%	7.25	3.5	107%
<b>Overall</b>	8.4	4.6	83%	7.4	4.5	76%

**Table 8.1: Expert intersection comparison across studies**

As shown in Table 8.1, results of these two analyses are very similar. TopicShop users had a much larger intersection with expert sets. The percent increase of the TopicShop intersection over the Yahoo intersection is about the same. Subjects within the topic Buffy the Vampire Slayer (the only topic that remained constant across the two studies) had slightly different results. While TopicShop subjects selected about the same number of quality sites in both studies, Yahoo subjects selected 30% less quality sites in the main study. This can probably be explained by the fact that the number of Buffy sites increased from 60 to 258. For TopicShop users this was no problem because the interface gave them the tools they needed to sift through a large collection and find the best sites. But Yahoo subjects only had a simple alphabetical list of sites with a high level sub-category structure imposed and were probably overwhelmed by the sheer number of sites and thus were not able to find as many quality sites.

### 8.1.2 User Effort

There are two metrics involving the amount of effort a user expended that we can compare between the two studies: task time and number of sites browsed. Recall that task time was constrained to 45 minutes in the pilot study, but not in the main study. Table 8.2 shows overall task times for the two interface conditions across the two studies. TopicShop subjects spent 10% less time in the main study than they did in the pilot and Yahoo subjects spent 13% more time. In the main study, the 45 minute time limit was removed and subjects were allowed to spend as long as necessary to complete their task, which explains why Yahoo subjects took more time during the main study. In the new iteration of TopicShop used in the main study, we made the most widely used features easily accessible to the users. By providing

fast access to important interface components, we enabled TopicShop subjects to finish their task more efficiently in the main study despite the increased number of web sites.

	<b>TopicShop</b>	<b>Yahoo</b>	<b>% More Time</b>
<b>Pilot Study</b>	41.5	46.6	12%
<b>Main Study</b>	37.69	52.53	39%
<b>Time Difference</b>	10% <i>less</i> time	13% <i>more</i> time	

**Table 8.2: Task time comparison across studies**

The gap in the number of sites browsed between TopicShop and Yahoo users also widened. In the pilot study, Yahoo subjects browsed to 8 more sites on average than TopicShop users (Table 8.3). The results from the main study show that this increased to 13.2 sites. In both cases, subjects browsed fewer sites in the main study than in the pilot study, even though there were more sites for each topic and unlimited time. When there was a smaller number of sites, it was fairly easy to consider all sites while selecting the best. But when there are hundreds of sites, even if users can immediately throw away half, using title and other information, it is still not feasible to browse to all remaining sites. Therefore when presented with a larger list of relevant sites, making comprehensive investigation of every site impossible, users will browse to only the sites that appear most promising.

	<b>TopicShop</b>	<b>Yahoo</b>	<b><u>Difference (in # of sites)</u></b>
<b>Pilot Study</b>	36	44	8
<b>Main Study</b>	27.05	40.25	13.2
<b>Difference in number of sites browsed</b>	25% less sites	8% less sites	65% larger gap

**Table 8.3: Comparison of number of sites browsed**

With our improvements to the TopicShop interface, targeted at improving the user’s ability to quickly select high quality sites and effectively organize them into meaningful sub-categories, we were able to decrease the amount of time and effort spent by subjects even though the size of the collections increased substantially between the pilot study and the main study.

### 8.1.3 Questionnaire

There are also two interesting similarities in responses we received on the questionnaires completed during the study. We asked subjects in both studies whether or not they were confident that sites they selected were the best set of sites to represent their topic. The ratings were on a 1 to 7 scale, with 1 being very confident and 7 being not confident at all.

	<b>TopicShop</b>	<b>Yahoo</b>	<b>% Increase</b>
<b>Pilot Study</b>	4.5	4.75	5.5%
<b>Main Study</b>	2.25	2.5	11.1%

**Table 8.4: User confidence from questionnaire**

Overall the scale shifted, with all subjects being more confident that they had a good collection of sites in the main study, probably because they had unlimited time to complete their task. TopicShop subjects had slightly more confidence in their collection in both studies, as seen in Table 8.4. Again, this is probably due to the fact that they had more information on which to base their decisions. The larger quantity of sites in the collections could also have an effect. In a substantial set of web sites, mediocre sites usually far outweigh high quality sites. In the main study, subjects probably browsed a large number of similar, bland sites and this might have influenced their confidence in the quality of the sites they had already selected.

	<b>Pilot Study</b>	<b>Main Study</b>
<b>In-Links</b>	2.0	3.2
<b>Title</b>	2.75	3.7
<b>Pages</b>	3.0	4.2
<b>Others</b>	>5.0	>5.0

**Table 8.5: Site Parameter rankings**

Subjects in the TopicShop condition were asked to rank from most to least useful the site profile data they used during the experiment. In both studies the top 3 most useful parameters were: in-links, title, and pages, as shown in Table 8.5. The other parameters had average ranks greater than 5 (on a scale of 1-7) in both studies. In both studies, in-links was chosen as the most useful metric. Title is usually a very useful property because it describes the contents of a site and serves as a handy label that users associate with the site for later recall. Number of pages can also be used as a somewhat crude metric for discerning the quality of a site. Sites with more pages are usually those where their authors spent more time and effort, thus producing a higher quality site. In-links are the most popular because they extract their utility

from the rich hyperlink structure of the World Wide Web representing the number of endorsements to each site. In the next chapter we perform an analysis that indicates that both in-links and number of pages are good predictors of site quality and validate the utility judgments of our subjects.

## CHAPTER 9: PREDICTING QUALITY SITES

The in-link metric used throughout our work is one of a number of similar metrics based on hyperlink structure. We have shown that these metrics have great potential to identify high quality items and that our metric of simply counting the links to a document is one that yields an estimate of that document's quality. More sophisticated algorithms have also been developed that build on this intuition (See chapter 2). However, there has been little empirical evaluation of these algorithms. This leaves a fundamental issue unresolved: do link-based metrics correlate with human judgements of quality? We are 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 discuss our investigation of content-based as well as link-based metrics. With the data we already have from our previous experiments, we can make a comparison to other well-known metrics by applying them to our existing data.

We considered several other issues before we determined how accurately these metrics could predict quality. First, we needed to know to what extent topic experts agree on the quality of sites within a topic. If human quality judgements vary widely, this suggested 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 concept and understanding of quality even exists. Conversely, if experts did tend to agree in their quality judgements,

our confidence in the concept of quality would be bolstered, even if it is difficult to give a precise definition.

Second, we needed to know whether there were any significant differences among various link analysis algorithms; for example, would one person score documents  $D_1$ ,  $D_2$ , and  $D_3$  highly, while another person scored  $D_4$ ,  $D_5$ , and  $D_6$  more highly? If there were no such differences, then an algorithm could be chosen for other factors, such as efficiency.

Recall from the related work section (chapter 2) that Kleinberg formalized the quality of documents within a hyperlinked collection using the concept of *authority* [52]. An authoritative document is one that many other documents link to. This notion can be strengthened by observing that links from all documents are not 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).

## 9.1 EXPERIMENT

The collections of URLs that we created using our crawler/analyzer for the two studies can be used to investigate whether there are web site metrics that predict quality of a site. For any of the collections we have already created we can think of the hyperlink structure as a graph whose nodes are sites and whose edges represent a hyperlink from one site to another. This neighborhood graph can be used to evaluate metrics that are based on the hyperlink structure of a set of sites.

From these graphs, we computed 5 link-based features: in and out degree, Kleinberg's authority and hub scores[52][53], and the PageRank score [16][74]. 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 & Henzinger [10] by counting only links between URLs on different sites and averaging the contribution of links from all URLs on one site to an 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, our crawler computes text similarity scores. Although we consider relevance and quality to be different notions, we wanted to test whether relevance would help predict quality. The crawler uses the Smart IR system [18] to generate a centroid – a weighted vector of keywords – from the content of the seed items for each topic. A relevance score is computed for each item by computing the inner product similarity of the item’s text to the centroid. For each site, the relevance score of the root page, the maximum relevance score of any contained page, and the average relevance scores of all contained pages are recorded.

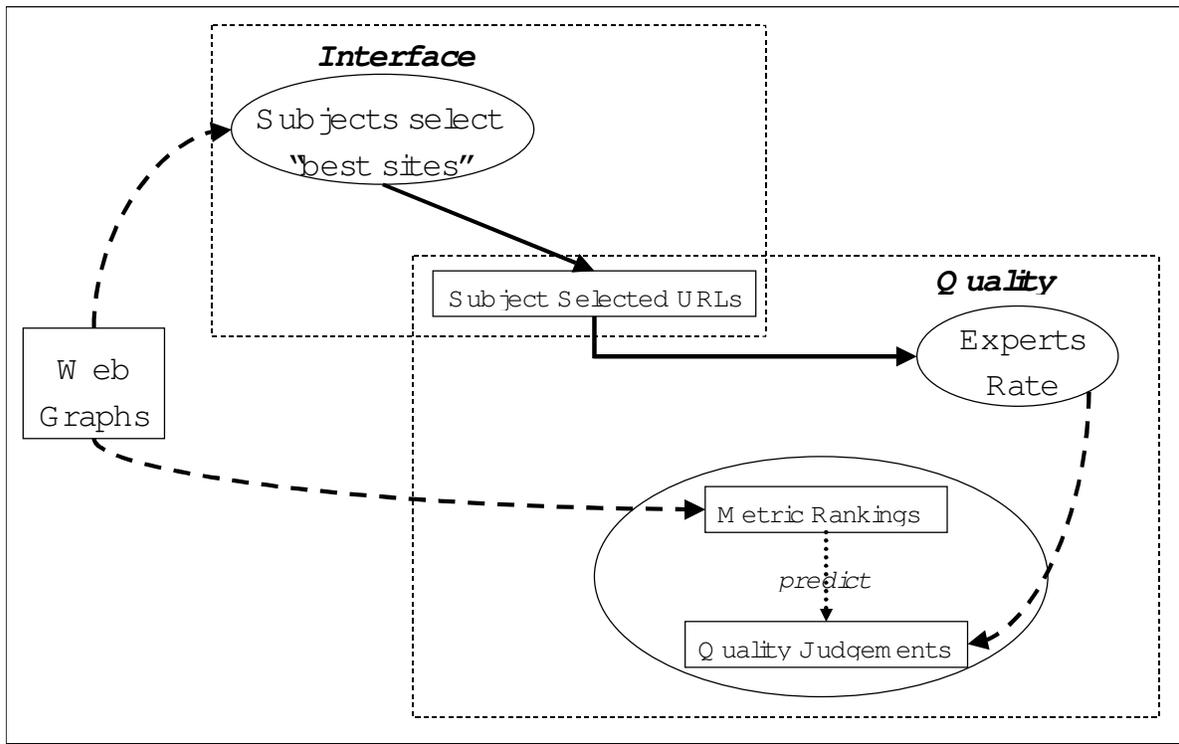
Each of the features, listed below, induces a ranking of the items in our dataset. In subsequent analysis, we examined how well the various rankings matched human quality judgements. Following is a list of all nine features we used (each feature can be applied at both the page and site level, resulting in 18 total features). They are a combination of the features we used in TopicShop and features used by other popular systems:

- In degree – number of sites that link to this site,
- Kleinberg’s Authority Score,
- PageRank Score – link-based score used in Google [16],
- Out degree – number of sites this site links to,
- Kleinberg’s Hub Score,
- Text relevance score – similarity to topic seed text,
- Size (number of bytes and number of contained pages),
- number of images, and
- number of audio files.

### **9.1.1 Data**

In analyzing how well we can predict human quality judgments, we built on some of the results that we had already obtained. For this analysis, the important components of our two studies were sites selected by our subjects and expert ratings given to each of those sites. We combined these data with profiles collected by our crawler and attempted to predict the quality of sites according to our experts.

Figure 9.1 shows the relationship of the previous two interface studies, discussed in sections 6 and 7, to this analysis.



**Figure 9.1: Data for quality experiments**

## 9.2 RESULTS

### 9.2.1 Expert Agreement

We first investigated how much experts agreed in their quality evaluations. 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 two computations to measure agreement. First, we correlated scores assigned to sites by each pair of experts for each topic. (Recall there were 4 experts to rate The Simpsons sites and 3 for all other topics.) We used the Pearson product-moment correlation since expert averages represent interval data ranging from 1 to 7. Table 9.1 presents the results. It shows that almost all pairs of experts were highly correlated in their judgements of item quality (all correlations were significant,  $p < 0.01$ ).

Topic	Correlations between pairs of experts						
	1-2	2-3	1-3	1-4	2-4	3-4	Average
Babylon 5	0.91	0.92	0.76				<b>0.87</b>
Buffy	0.75	0.79	0.83				<b>0.79</b>
Smashing Pumpkins	0.80	0.73	0.69				<b>0.74</b>
Tori Amos	0.61	0.63	0.50				<b>0.58</b>
Simpsons	0.52	0.59	0.50	0.75	0.59	0.59	<b>0.59</b>
Total							<b>0.71</b>

**Table 9.1: Expert agreement using correlations**

We did a second analysis that abstracted expert judgements a bit. Rather than using exact scores that experts assigned to sites, we categorized each site into one of two bins: “good” sites were those that an expert rated 5, 6, or 7, and “other” sites were all the rest. For each topic, we computed the set of sites that all experts assigned to the same category, as well as pair-wise agreement between each pair of experts.

Table 9.2 presents the results, which are quite similar to the correlations presented above. On average, across topics, all experts agreed on the category for 65% of items. Pairs of experts agreed 78% of the time. Since ratings within each topic were put into two categories, “good” and “other”, we computed Cohen’s kappa using an extension by Fleiss to test the agreement of multiple raters ( $\kappa=0.469$ ,  $p<0.000001$ ). Kappa values for each topic are shown in the last column of Table 9.2. This table shows that there was more agreement in some topics than others. Note that this agreement measure is lower than the other two presented in the table. Cohen’s kappa is a more accurate measure of agreement for categorical data because it computes and removes agreements to be expected by chance.

Topic	# Items	# Times Users Agreed	% Agreement	Average Pairwise #Agreement	Average Pairwise %Agreement	Cohen’s kappa (w/ Fleiss Extension)
Babylon 5	40	31	0.78	34.0	0.85	0.691 ( $p<.000001$ )
Buffy	41	28	0.68	32.3	0.79	0.540 ( $p<.000001$ )
Simpsons	39	24	0.62	30.7	0.79	0.420 ( $p<.0003$ )
Smashing Pumpkins	41	28	0.68	32.3	0.79	0.366 ( $p<.02$ )
Tori Amos	42	21	0.50	28.0	0.67	0.281 ( $p<.006$ )
Average	40.6	26.4	<b>0.65</b>	31.5	<b>0.78</b>	<b>0.469 (<math>p&lt;.000001</math>)</b>

**Table 9.2: Expert agreement, using categories**

These results suggest that experts generally agree on the nature of quality within a topic, and therefore that expert judgments can be used to evaluate rankings obtained by algorithms. However, there is some variation between topics; Babylon 5 experts agreed the most, Tori Amos experts the least, and the other three topics were in the middle. Some lack of agreement 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 relevant to the topic. Some experts' quality judgments may be influenced by the relevance. Second, some variation in opinions is inevitable, particularly in the area of popular entertainment, where there is no objective quality standard. 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 filtering approach ultimately may be necessary. To get the best information for *you*, you may have to inform the system about your preferences, so it can find experts with similar preferences, and recommend items that they like.

### 9.2.2 Link-Based Metric Comparison

The second issue we investigated was whether the three link-based metrics – in degree (number of links to a particular site), authority, and PageRank – ranked items differently.

Since 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 9.3 presents the results. The correlations were extremely high (and were all significant,  $p < 0.01$ ). We also computed Kendall's coefficient of concordance (Kendall's W) across all three metrics to verify their similarity. The average coefficient of concordance across all topics was 0.65 ( $p < .0001$ ).

Topic	Inlinks/ Authority	Inlinks/ PageRank	Authority/ PageRank	Kendall's W
Babylon 5	0.97	0.93	0.90	0.68
Buffy	0.92	0.85	0.70	0.82
Simpsons	0.97	0.99	0.95	0.51
Smashing Pumpkins	0.95	0.98	0.92	0.51
Tori Amos	0.97	0.92	0.88	0.73
Average	<b>0.96</b>	<b>0.93</b>	<b>0.87</b>	<b>0.65</b>

**Table 9.3: Metric similarity, using Spearman's Rho Correlation. Kendall's coefficient of concordance is included at the right for reference**

Second, we computed intersections between the top 5 and top 10 items as ranked by the three metrics. Table 9.4 presents the results. Again, there is great agreement. For example, in-degree and authority have an average intersection of 8.4 of the top 10 items, and all three metrics agree on an average of 6.4 of the top 10 items.

Topic	Inlinks/Auth 5	Inlinks/PR 5	Auth/PR 5	All 5	Inlinks/Auth 10	Inlinks/PR 10	Auth/PR 10	All 10
Babylon 5	5	4	4	4	9	7	6	6
Buffy	4	4	3	3	7	5	5	4
Simpsons	3	3	3	2	8	8	7	6
Smashing Pumpkins	5	4	4	4	9	9	9	9
Tori Amos	5	4	4	4	9	9	8	7
<b>Total</b>	<b>4.4</b>	<b>3.8</b>	<b>3.6</b>	<b>3.4</b>	<b>8.4</b>	<b>7.6</b>	<b>7</b>	<b>6.4</b>

**Table 9.4: Metric similarity, intersection of top 5 and 10**

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 is not sufficient and that all links are not equal. While our results do not prove this assumption false, they definitely indicate it needs further investigation.

By starting with items from Yahoo, we almost guaranteed that in the neighborhood graph we constructed would be relevant to the topic. In contrast, other evaluations of Kleinberg’s algorithm [52] have begun with much noisier neighborhoods. Typically, Kleinberg’s algorithm has started with a base set of web pages 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 while 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 do not form a coherent topic. In such cases, the authority/hub algorithm will assign low scores to some items with high in-degree.

Two processes are occurring: (1) obtaining a set of relevant items and (2) rating the quality of in this set. As commonly conceived, the authority algorithm helps with both. However, our additional data analysis shows that if one already has a set of relevant items, in-degree alone may be just as good a quality measure. Many manually constructed collections of topically relevant are available from general-purpose or topic-oriented directories.

Further, the in-degree metric we are using is *site* in-degree. By aggregating links to the site level, we avoid the problems Bharat & Henzinger [10] identified (links between pages that belong 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.

### 9.2.3 Predicting Quality

We tested how well rankings induced by each of the 9 features (at the site level only) listed in section 9.1 matched expert quality judgements. The first analysis we performed was a simple stepwise linear regression to see if any of the metrics could be combined to predict quality effectively. The set of predictors that were available for inclusion in the stepwise regression were: in-degree, out-degree, authority score, hub score, PageRank score, text relevance score, size, number of images, and number of audio files. The best model combined in-links, authority score, PageRank score, and number of images. All the link-based metrics that we considered were included in the model, which again showed that evaluating link structure can assist users in finding high quality web sites. Table 9.5 shows coefficients and significance for the linear model. The  $R^2$  value for the model was 0.328 ( $p < 0.001$ ), so the best linear model does not do a very good job at predicting quality.

Metric	Coefficient	t-value	Significance
Constant Term	4.238	31.407	0.000
Authority Score	1.338	1.067	0.289
In-Degree	0.052	2.291	0.024
Images	0.0002	1.352	0.180
PageRank Score	-0.225	-1.477	0.143

**Table 9.5: Linear Model for Predicting Expert Average**

We wanted to compute the precision of each ranking; to do this, we needed the set of good (high quality) items for each topic. As in previous analysis, we defined good items as those that a majority of experts rated as good (i.e., scored 5, 6, or 7). Table 9.6 shows the total number of for each topic, number of good, and proportion of good. The proportion of good serves as a useful baseline; it tell us that, across all topics, if a set of 10 items were picked at random, about 3 would be expected to be high quality.

<b>Topic</b>	<b>Total</b>	<b>Number good</b>	<b>Proportion good</b>
<b>Babylon 5</b>	40	19	0.48
<b>Buffy</b>	41	15	0.37
<b>Simpsons</b>	39	10	0.26
<b>Smashing Pumpkins</b>	41	7	0.17
<b>Tori Amos</b>	42	13	0.31
<b>Average</b>			<b>0.32</b>

**Table 9.6: Number and proportion of good**

For ease of presentation, we show results for 10 metrics of the 18 total metrics defined in section 9.1. 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 included the best – maximum relevance score – for the sake of comparison.

Using the set of good, we computed the precision at 5 and at 10 for each metric (recall that precision is the standard information retrieval metric measuring the number of “good” items that are actually considered “good” by (in our case) experts in our study). Table 9.7 presents results, with metrics ordered by average precision at 5. The table shows that the top four or five metrics all performed quite well. (Recall that TopicShop subjects performed better than all of these metrics, while Yahoo subjects performed worse. See section 7.7.8.) For example, the in-degree metric has a precision at 5 of 0.76: on average, which means nearly 4 of the first 5 documents it returns would be rated good by the experts. This is more than double the number of good documents expected from selecting 5 at random from the expert dataset that we have been using throughout our analysis. And recall that most of the items in the expert dataset probably are of good quality, since they were selected by multiple subjects in phase 1 of our experiment in the main study. Thus, we speculate that in a larger dataset, the improvement in quality obtained by using these metrics is even greater.

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

**Table 9.7: Precision at 5 and 10**

Since we showed that 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 our dataset of expert judgements, 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 is not bad, either). We speculate that the number of pages on a site is an indication of how much effort the author is devoted to the site. And more effort may indicate higher quality. Interestingly enough, recall that subjects in both our studies rated number of pages as a useful metric.

The precision analysis abstracted away from the site scores by treating the entire range of good sites equally, rather than acknowledging that some sites in this set are better than others. This could result in significant differences being hidden. For example, suppose that two metrics have identical precision. In principle, they could return completely different sets of items; further, one metric could return the best – highest ranked – of the good items, while the second returned the worst of the good items. Thus, we performed another analysis using site scores to check for this possibility.

We experimented with two different scoring schemes: the average of all expert scores and a majority score (# of experts rating item as good / # of experts rating the item). The two methods yielded similar results, and for the sake of consistency with previous analysis, we used majority score.

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

**Table 9.8: Majority Score at 5 and 10**

Table 9.8 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, and 68% at 10. So, precision analysis was consistent with this more fine-grained approach; no important distinctions were lost.

We also computed a simple average expert score of the top 10 sites for each of our 18 metrics. Table 9.9 shows results of this analysis (only the 9 metrics computed at the site level are included in the table, but the analysis was performed with all 18 metrics). 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 18x5

two factor ANOVA of metric (the 18 features listed in section 9.1) and topic (the 5 topics in our study) verified that the metric factor is statistically significant ( $F(17,810)=3.775$ ,  $p<.0001$ ) and the interaction between topic and metric was not significant ( $F(68,810)=.683$ ,  $p<.976$ ). For reference, the topic factor was also significant ( $F(4,810)=26.67$ ,  $p<.0001$ ).

	Babylon	Buffy	Simpsons	Smashing Pumpkins	Tori Amos	Average
# Pages on site	5.568	5.767	4.2	4.367	3.701	4.7206
In degree	5.134	5.2	4.625	4.135	4.335	4.6858
Authority score	5.4	5.067	4.425	4.168	4.368	4.6856
# Images	5.668	5.366	4.1	4.1	3.968	4.6404
PageRank score	5.301	4.934	4.45	4.135	3.735	4.511
Out degree	4.367	4.433	4.025	3.767	4.334	4.1852
Hub score	4.334	4.5	3.725	4.034	4.301	4.1788
Max Rel Score	4.735	4.166	3.975	3.3	4.467	4.1286
# Audio files	3.834	3.699	4.05	4.066	4.8	4.0898
Root Page Size	4.134	3.3	3.55	3.4	3.8	3.6368

**Table 9.9: Average Expert Scores of Top 10 sites**

The same metrics – in-degree, authority, page rank, #pages, and #images – are in the top 5 slots in each of the past six analyses (precision/majority score at 5/10, expert average, and expert position), although their order varies a little. We wondered whether there were any significant differences among the metrics. We performed a Tukey HSD post-hoc analysis, but found that there were no significant differences in any of the parameters listed above at the 0.05 level. So results are not statistically significant, but may prove interesting nonetheless. Further analyzing the data shows that results can be broken into 3 distinct subsets in the data. Tables (Table 9.7, Table 9.8, Table 9.9) have a triple line indicating where these subsets occur. It turns out that the top 5 means are very close in value to each other, and are distinct from the rest of the chart. This is also true of the next four metrics. (out-degree, hub score, maximum relevance score, and root page size). Root page size was the only metric that was much less than every other metric in the table.

We found other interesting results. First, there were virtually no differences between any of the first five metrics. Second, in-degree was better than the rest of the metrics (i.e., other than the top 5). Third, all the top 5 methods performed better than text similarity. Perhaps text similarity fares so poorly because we started with a set of relevant documents; in other words, if there were more variance in relevance, higher relevance might indicate higher quality.

#### **9.2.4 Discussion**

With this analysis, we investigated the utility of various computable metrics in estimating the quality of web documents. We showed that topic experts exhibited a high amount of agreement in their quality judgements; however, enough difference of opinion existed to warrant further study. We also showed that three link-based metrics and a simple content metric do a very good job of identifying high quality items.

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

## CHAPTER 10: SUMMARY AND CONCLUSIONS

As the amount of information on the web continues to grow, tools that support users in finding and managing collections of topical resources are becoming increasingly significant. The focus must move from compiling collections to helping users comprehend and manage them. Our main goal is to reduce the time users must spend sifting through “relevant” – but poor quality – sites and increase the amount of time they can devote to exploring high-quality information. We achieved this goal by mining the rich data that already exist in the structure of web sites and content of their pages, and have shown that our system, TopicShop, helps users quickly identify small, manageable, high-quality subsets of web sites. During the course of this research we wanted to develop a Java web crawler that efficiently gathers collections of relevant web sites about a topic using the hypertext structure of the web, iteratively design a visualization and management user interface for viewing and maintaining these collections, and show through empirical studies the effectiveness of this task-specific user interface for the task of topic management.

We started by investigating the task of topic management, the task of gathering, evaluating, and organizing relevant information resources for a given topic. Once we understood how users currently perform this task, we developed an efficient automatic method of gathering high quality collections of web sites about a particular topic and building site profiles of these sites (our web crawler). These collections are necessary to allow us to explore how users manage topically relevant resources. We then began designing TopicShop based on what we had learned about the task of topic management.

The pilot study evaluated the initial implementation of TopicShop. It was a small 2x2 study with 16 subjects comparing TopicShop to Yahoo + Bookmarks across two entertainment topics. We found that subjects using TopicShop were able to find 80% more high quality sites, while browsing only 81% as many sites and completing their task in 89% of the time. The site profile data that TopicShop provides - in

particular, the number of pages on a site and the number of other sites that link to it – was the key to these results, as users exploited it to identify the most promising sites quickly and easily.

Following an extensive redesign of the TopicShop interface, incorporating design changes that addressed issues brought out in the pilot study, we conducted a second empirical study to further assess the effectiveness of our interface. The main empirical study was a 2x5 design with 40 subjects again comparing TopicShop to Yahoo + Bookmarks, but with 5 entertainment topics. The results further validate the conclusions of the pilot study, that subjects using TopicShop were able to find significantly more high quality sites, in less time and with less effort. We were also able to show that TopicShop subjects spent half the time organizing sites, yet still created more groups and more annotations, and agreed more in how they grouped the sites. TopicShop subjects also tightly integrated the sub-tasks of evaluating and organizing sites.

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 systems improve people's ability to solve these problems. It provides information and interaction techniques that help people select the best sites from large collections of web sites. Two user studies have demonstrated that users can select better sites, more quickly and with less effort. It also offers 2D 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.

## CHAPTER 11: REFERENCES

- [1] Abrams, D., Baecker, R., and Chignell, M. Information Archiving with Bookmarks: Personal Web Space Construction and Organization. *Proceedings of CHI'98* (Los Angeles CA, April 1998), ACM Press, 41-48.
- [2] Ackerman, M. and McDonald, D. (1996) Answer Garden 2: Merging Organizational Memory with Collaborative Help. *Proceedings of CSCW '96* (Boston MA, November 1996) ACM Press.
- [3] Allen, R. (1995) Two Digital Library Interfaces which Exploit Hierarchical Structure.
- [4] Allen, R. (1990) User Models: Theory, Method, and Practice.
- [5] Amento, B., Hill, W., Terveen, L., Hix, D. and Ju, P. An Empirical Evaluation of User Interfaces for Topic Management of Web Sites. *Proceedings of CHI'99* (Pittsburgh PA, May 1999) ACM Press.
- [6] Amento, B., Terveen, L., and Hill, W. Does "Authority" Mean Quality? Predicting Expert Quality Ratings of Web Documents, in *Proceedings of SIGIR 2000* (Athens Greece, July 2000), ACM Press.
- [7] Amento, B., Terveen, L., Hill, W, and Hix D. TopicShop: Support for Evaluating and Organizing Collections of Web Sites, in *Proceedings of UIST' 2000* (San Diego CA, November 2000), ACM Press.
- [8] Baldonado, M.Q.W., and Winograd, T. An Information-Exploration Interface Supporting the Contextual Evolution of a User's Interests, in *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 11-18.
- [9] Bederson, B.B., Hollan, J.D., Perlin, K., Meyer, J., Bacon, D., and Furnas, G. Pad++: A Zoomable Graphical Sketchpad for Exploring Alternate Interface Physics. *Journal of Visual Languages and Computing* 7, 3-31, 1996.
- [10] Bharat, K. and Henzinger, M.R. Improved Algorithms for Topic Distillation in a Hyperlinked Environment. ACM SIGIR Conference on Research and Development in Information Retrieval 1998.
- [11] Bharat, K. and Broder, A. A Technique for Measuring the Relative Size and Overlap of Public Web Search Engines. *Proceedings of the Seventh International World Wide Web Conference* (Brisbane Australia April 1998).

- [12] Botafago, R., Rivlen, E., and Schneiderman, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems*.
- [13] Bowman, C., Danzig, P., and Manber, I. (1994) Scalable Internet Resource Discovery: Research Problems and Approaches.
- [14] Brewer, R and Johnson, P. (1994) Collaborative Classification and Evaluation of Usenet. University of Hawaii TechReport ICS-TR-93-22
- [15] Brewer, R. and Johnson, P. (1994) Toward Collaborative Knowledge Management within Large, Dynamically Structured Information Systems.
- [16] Brin, S. and Page, L. (1998) The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Proceedings of the Seventh International World Wide Web Conference* (Brisbane Australia, April 1998).
- [17] Brothers, L., Hollan, J., Nielsen, J., Stornetta, S., Abney, S., Furnas, G., and Littman, M. (1992) Supporting Informal Communication via Ephemeral Interest Groups. *Proceedings of CSCW '92* (Toronto ON, 1992), ACM Press.
- [18] Buckley, C. Implementation of the SMART Information Retrieval System, Department of Computer Science, Cornell University, 1985, TR85-686.
- [19] Card, S.K., Robertson, G.C., and Mackinlay, J.D. The Information Visualizer, an Information Workspace, in *Proceedings of CHI'91* (New Orleans LA, April 1991), ACM Press, 181-188.
- [20] Card, S., Robertson, G., and York, W. The WebBook and the Web Forager: An Information Workspace for the World Wide Web. *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press.
- [21] Carroll, J.M. and Rosson, M.B. Getting Around the Task-Artifact Cycle: How to Make Claims and Design By Scenario. *ACM Transactions on Information Systems* 10(2), 181-212, 1992.
- [22] Carriere, J. and Kazman, R. (1997) WebQuery: Searching and Visualizing the Web through Connectivity. *Proceedings of the Sixth International World Wide Web Conference* (Santa Clara, April 1997).
- [23] Chakrabarti, S., Dom, B., Gibson, D., Kleinberg, J., Raghavan, P., Rajagopalan, S. Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *Computer Networks and ISDN Systems* 30 (1998), 65-74
- [24] Chakrabarti, S., Dom, B., Raghavan, P., Rajagopalan, S., Gibson, D., and Kleinberg, J. (1998) Automatic Resource Compilation by Analyzing Hyperlink Structure and Associated Text. *Proceedings of the Seventh International World Wide Web Conference* (Brisbane Australia, April 1998).
- [25] Chen, C. and Czerwinski, M. From Latent Semantics to Spatial Hypertext – An Integrated Approach. *Proceedings of Hypertext 1998* (Pittsburgh PA, June 1998).
- [26] Cho, J., Garcia, H., and Page, L. (1998) Efficient Crawling Through URL ordering. *Proceedings of the Seventh International World Wide Web Conference* (Brisbane Australia, April 1998).
- [27] Cohen, W. (1995) Learning to Classify English Text with ILP Methods. *Advances in Inductive Logic Programming*. IOS Press, 1995, IOS Frontiers in AI and Applications series.
- [28] Constant, D., Sproull, L., and Kiesler, S. (1996) The Kindness of Strangers: On the Usefulness of Weak Ties for Technical Advice. *Organizational Science*, 1996. Vol. 7. Number 2.

- [29] Durand, D. and Kahn, P. MAPA: A System for Inducing and Visualizing Hierarchy in Websites. *Proceedings of Hypertext 1998* (Pittsburgh PA, June 1998).
- [30] Fischer, G. and Stevens, C. Information Access in Complex, Poorly Structured Information Spaces. *Proceedings of CHI'91*.
- [31] Furnas G. (1995) Effectively View-Navigable Structures. *Human Computer Interaction Consortium Workshop 1995 (HCIC95)* (Snow Mountain Ranch CO, February 1995).
- [32] Furnas, G. (1985) Experience with an Adaptive Indexing Scheme. *Proceedings of CHI '85*.
- [33] Furnas, G. (1997) Effective View Navigation. *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press.
- [34] Furnas, G.W. Generalized fisheye views. in *Proceedings of CHI' 86*(Boston, MA April 1986), ACM Press, 16-23
- [35] Garfield, E. (1979) Citation Indexing. Institute for Scientific Information, Philadelphia, PA.
- [36] Gaver, W. (1995) Oh What a Tangled Web We Weave: Metaphor and Mapping in Graphical Interfaces. *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press.
- [37] Gibson, D., Kleinberg, J., and Raghavan, P. (1998) Inferring Web Communities from Link Topologies. *Proceedings of Hypertext 1998* (Pittsburgh PA, June 1998).
- [38] Gibson, D., Kleinberg, J., and Raghavan, P. (1998) Clustering categorical data: An approach based on dynamical systems. *Proceedings of the 24th International Conference on Very Large Databases*, 1998.
- [39] Goldberg, D., Nichols, D., Oki, B., and Terry, D. Using Collaborative Filtering to Weave an Information Tapestry. *CACM* 1992, Vol. 35, Number 12.
- [40] Grudin, J. Social Evaluation of the User Interface: Who Does the Work and Who Gets the Benefit? *Proceedings of Interact '87*.
- [41] Hartson, H.R., Castillo, J., Kelso, J., Kamler, J., and Neale, W. (1996) Remote Evaluation: The Network as an Extension of the Usability Laboratory. *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press.
- [42] Hightower, R., Ring, L., Helfman, J., Bederson, B., and Hollan, J. Graphical Multiscale Web Histories: A Study of PadPrints. *Proceedings of Hypertext 1998* (Pittsburgh PA, June 1998).
- [43] Hill, W., Hollan J., Wroblewski, D., and McCandless, T. (1992) Edit Wear and Read Wear. Human factors in computing systems: Striking a Balance. *Proceedings of CHI'92*.
- [44] Hill, W. and Hollan, J. (1993) History-Enriched Digital Objects. *Proceedings of CFP '93*.
- [45] Hill, W. and Terveen, L. (1996) Using Frequency-of-Mention in Public Conversations for Social Filtering. *Proceedings of CSCW '96* (Boston MA, November 1996).
- [46] Hill, W., Stead, L., Rosenstein, M., and Furnas, G. (1995) Recommending and evaluating choices in a Virtual Community of Use. *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press.
- [47] Hjørland, B. (1997) Information Seeking and Subject Representation: An Activity-Theoretical Approach to Information Science. Greenwood Press, Connecticut.

- [48] Jackson, M. (1997) Assessing the Structure of Communication on the World Wide Web. *Journal of Computer-Mediated Communication*. Volume 3, Number 1. June 1997.
- [49] Jansen, B. J., Spink, A., Bateman, J., and Saracevic, T. Searchers, the Subjects They Search, and Sufficiency: A Study of a Large Sample of EXCITE Searches, submitted to *WebNet' 98*
- [50] Kandogan, E and Schneiderman, B. (1997) Elastic Windows: A Hierarchical Multi-Window World Wide Web Browser. *Proceedings of UIST '97*.
- [51] Keller, R.M., Wolfe, S.R., Chen, J.R., Rabinowitz, J.L., and Mathe, N. A Bookmarking Service for Organizing and Sharing URLs. *Proceedings of the Sixth International World Wide Web Conference* (Santa Clara CA, April 1997).
- [52] Kleinberg, J.M. (1998) Authoritative Sources in a Hyperlinked Environment. *Proceedings of 9th ACM-SIAM Symposium on Discrete Algorithms* (San Francisco CA, January 1998), ACM Press.
- [53] Kleinberg, J., Papadimitriou, C., and Raghavan, P. Segmentation problems: A micro-economic view of data mining. *Proceedings of the 30th ACM Symposium on Theory of Computing*, 1998.
- [54] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. (1996) Applying Collaborative Filtering to Usenet News: The GroupLens System.
- [55] Lamping, J., Rao, R., and Pirolli, P. (1995) A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies. *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press, 401-408.
- [56] Liechti, O., Sifer, M., and Ichikawa, T. (1998) Structured Graph Format: XML Metadata for Describing Web Site Structure. *Proceedings of the Seventh International World Wide Web Conference* (Brisbane Australia April 1998).
- [57] Levi, M. and Conrad, F. (1996) A Heuristic Evaluation of a World Wide Web Prototype. *Interactions* July/August 1996
- [58] Maarek, Y., Jacovi, M., Chtalhaim, M., Ur, S., Zernik, D., and Shaul, I. (1997) WebCutter: A System for Dynamic and Tailorable Site Mapping. . *Proceedings of the Sixth International World Wide Web Conference* (Santa Clara, April 1997).
- [59] Mackinlay, J.D., Rao, R., and Card, S.K. An Organic User Interface for Searching Citation Links. *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press, 67-73.
- [60] Malone, T.W., Grant, K.R., and Turback, F.A. (1986) The Information Lens: An Intelligent System for Information Sharing in Organizations. *Proceedings of CHI'86*.
- [61] Malone, T., Grant, K., Turbank, F., Brobst, S., and Cohen, M. (1987) Intelligent Information-Sharing Systems. *CACM* May 1987 Vol. 30, Number 5.
- [62] Maltz, D. and Ehrlich K. (1995) Pointing the Way: Active Collaborative Filtering. *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press.
- [63] Maltz, D. (1994) Distributing Information for Collaborative Filtering on Usenet News.
- [64] Mander, R., Salomon, G., and Wong, Y.Y. A 'Pile' Metaphor for Supporting Casual Organization of Information, in *Proceedings of CHI'92* (Monterey CA, May 1992), ACM Press, 627-634.
- [65] Marchiori, M. (1997) The quest for correct information on the Web: Hyper search engines. *Proceedings of the 6<sup>th</sup> International World Wide Web Conference* (Santa Clara, CA, April 1997).

- [66] Marshall, C., Shipman, F., and Coombs, J. VIKI: Spatial Hypertext Supporting Emergent Structure, in Proceedings of ACM ECHT ' 94, (Edinburgh, Scotland, September 1994). ACM Press,13-23.
- [67] Miller, R. and Bharat K. (1998) SPHINX: A Framework for Creating Personal, Site-Specific Web Crawlers. *Proceedings of the Seventh International World Wide Web Conference* (Brisbane Australia April 1998).
- [68] Morita, M. and Shinoda, Y. (1994) Information Filtering Based on User Behavior Analysis and Best Match Text Retrieval. SIGIR '94.
- [69] Mukherjea, S. and Hara, Y. (1997) Focus+Context Views of World-Wide Web Nodes. Proceedings of The Eighth ACM Conference on Hypertext and Hypermedia.
- [70] Mukherjea, S., and Foley, J.D. (1996) Visualizing the World-Wide Web with the Navigational View Builder. *Proceedings of the Fifth International World Wide Web Conference* (Paris France, May 1996).
- [71] Mukherjea, S., Foley, J., and Hudson, S. Visualizing Complex Hypermedia Networks through Multiple Hierarchical Views. *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press.
- [72] Newman, W. (1997) Better or Just Different? On the Benefits of Designing Interactive Systems in terms of Critical Parameters. *Proceedings of DIS '97*.
- [73] Nardi, B. and Barreau D. Finding and Reminding: File Organization from the Desktop. *ACM SIGCHI Bulletin*, 27, 3, July 1995.
- [74] Page L., Brin S., Motwani R., and Winograd T. The PageRank Citation Ranking: Bringing Order to the Web. *Stanford Digital Libraries Working Paper*.
- [75] Pirolli, P., Pitkow, J., and Rao, R. Silk from a Sow's Ear: Extracting Usable Structures from the Web. *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press 118-125.
- [76] Pirolli, P., Schank, P., Hearst, M., and Dieh, C. Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press 213-220.
- [77] Pirolli, P., and Card, S.K. (1995) Information Foraging in Information Access Environments. *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press.
- [78] Pitkow, J., and Pirolli, P. Life, Death, and Lawfulness on the Electronic Frontier. *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 383-390.
- [79] Rao, R., and Card, S. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information.
- [80] Recker, M. and Pitkow, J. (1994) Predicting Document Access in Large, Multimedia Repositories.
- [81] Resnick, P. and Varian, H. (1997) Communications of the ACM: Special issue on Recommender Systems. 40,3 (March 1997).
- [82] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., and Riedl, J. (1994) GroupLens: An Open Architecture for Collaborative Filtering of Netnews. *Proceedings ACM Computer Supported Cooperative Work (CSCW) 1994*.
- [83] Rhodes, B. and Starner, T. Remembrance Agent: A continuously running automated information retrieval system. *Proceedings of PAAM '96*.

- [84] Robertson, G., Czerwinski, M., Larson, K., Robbins, D.C., Thiel, C., van Dantzich, M. Data Mountain: Using Spatial Memory for Document Management, in *Proceedings of UIST'98* (San Francisco CA, November 1998), ACM Press, 153-162.
- [85] Schwartz, M., Emtage, A., Kahle, B., and Newuman, C. (1992) A Comparison of Internet Resource Discovery Approaches.
- [86] Scott, J. (1991) *Social Network Analysis: A Handbook*. Sage Publications, Inv., Thousand Oaks, CA.
- [87] Shardanand, U. and Maes, P. (1995) Social Information Filtering: Algorithms for Automating "Word of Mouth" *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press.
- [88] Shardanand, U. (1994) Social Filtering for Music Recommendation.
- [89] Shipman, F., Marshall, C., and LeMere, M. Beyond Location: Hypertext Workspaces and Non-Linear Views, in *Proceedings of ACM Hypertext '99* ACM Press, 121-130.
- [90] Spertus, E. ParaSite: Mining Structural Information on the Web. *Proceedings of the Sixth International World Wide Web Conference* (Santa Clara, April 1997).
- [91] Stevens, C. Automating the Creation of Information Filters. *CACM '92*. Vol. 35, Number 12.
- [92] Takano, H. and Winograd, T. Dynamic Bookmarks for the WWW: Managing Personal Navigation Space by Analysis of Link Structure and User Behavior. *Proceedings of Hypertext 1998* (Pittsburgh PA, June 1998).
- [93] Terry, D. (1993) A Tour Through Tapestry. *Proceedings ACM Conference on Organizational Computing Systems (COOCS) 1993*.
- [94] Terveen, L., Hill, W., and Amento, B. Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources. *ACM Transactions on Computer-Human Interaction*, 6,1 (March 1999), 67-94.
- [95] Terveen, L. and Hill, W. Finding and Visualizing Inter-site Clan Graphs. *Proceedings of CHI'98* (Los Angeles CA, April 1998), ACM Press 448-455.
- [96] Terveen, L. and Hill, W. Evaluating Emergent Collaboration on the Web. *Proceedings of CSCW'98* (Seattle WA, November 1998) ACM Press.
- [97] Terveen, L., Hill, W., Amento, B., McDonald, D., and Creter, J. (1997) Building Task-Specific Interfaces to High Volume Conversational Data. *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 226-233.
- [98] Terveen, L., Hill, W., Amento, B., McDonald, D., and Creter, J. (1997) PHOAKS: A System for Sharing Recommendations. *CACM* March 1997, Vol. 40, Number 3.
- [99] Terveen, L., Hill, W., Cherny, L., and Whittaker, S. (1997) Quantifying Online Conversation.
- [100] Wexelblat, A. and Maes, P. (1997) Footprints: History-Rich Web Browsing.
- [101] Whittaker, S. Electronic Collaboration: An Empirical Evaluation of Factors Affecting Mediated Group Interaction. *Proceedings of CSCW '96* (Boston MA, November 1996).
- [102] Wittenburg, K., Das, D., Hill, W., and Stead, L. (1995) Group Asynchronous Browsing on the World Wide Web. *Proceedings of the Fourth International World Wide Web Conference* (Boston MA, December 1995).

[103] <http://www.searchenginewatch.com/reports/mediamatrix.html>

[104] [http://www.gvu.gatech.edu/user\\_surveys/](http://www.gvu.gatech.edu/user_surveys/)

## CHAPTER 12: CURRICULUM VITAE

### Brian S. Amento

37 Headley Avenue  
Morris Plains, NJ 07950  
(973) 656-0886

[brian@research.att.com](mailto:brian@research.att.com)

<http://www.research.att.com/~brian/>

#### Education:

- **Doctor of Philosophy, Computer Science and Applications**
  - September 2001
  - Virginia Polytechnic Institute and State University (Virginia Tech)
  - Blacksburg, VA
  - GPA: 3.7/4.0 Major GPA: 3.7/4.0
- **Master of Science, Computer Science and Applications**
  - December 1995
  - Virginia Polytechnic Institute and State University (Virginia Tech)
  - Blacksburg, VA
  - GPA: 3.6/4.0 Major GPA: 3.6/4.0
- **Bachelor of Science, Computer Science, *Cum Laude* with Math Minor**
  - May 1994
  - Virginia Polytechnic Institute and State University (Virginia Tech)
  - Blacksburg, VA
  - GPA: 3.5/4.0 Major GPA: 3.9/4.0

#### Experience:

- **Senior Technical Staff Member, [AT&T Labs - Research](#)**, May 1996 to present
  - SCANMail - Developed and evaluated a service that uses Automatic Speech Recognition and other information analysis techniques to help users browse and search for voicemail messages based on content. Currently deployed to ~2000 internal employees.
  - Control Shadows - Graphical technique that supports users in the task of specifying preferences relative to their personal history on desktop and mobile devices. Evaluated the design implications of moving complex interfaces to smaller more limited platforms.
  - FGI - Developed a wristband mounted bio-acoustic finger gesture interface based on acoustic signals captured from gentle fingertip gestures through a contact microphone
  - CarThing - Wireless audio system for the car that automatically updates music selections based on past listening history, social filtering, and user preferences. The system takes advantage of high bandwidth (802.11) when available to move large audio files to and from the system and transmits control data only when bandwidth is limited (cell link).
  - WARR - Collaborative music listening environment supporting rich user interaction while listening to shared music lists.
  - Implemented [PHOAKS](#) People Helping One Another Know Stuff.
  - Designed TopicShop Explorer - interface for managing collections of topical resources.
  - Developed [FanCentral](#) to help users explore, organize, and share a collection of Web sites on entertainment topics.
- **Research, [Action-Centered Task Video System](#)**, Jan 1997 to Dec 1998
  - Developed ACTV System to provide more flexible and appropriate visual access to remote environments.
  - Performed empirical evaluation to validate usefulness of system in remote collaborative situations.
- **Research, [MOOsburg](#)**, Sep 1995 to Dec 1998

- Implemented a Text-Based Virtual Environment to support enhanced social interaction over the internet in the form of a MOO (Object Oriented Multi-User environment) named [MOOsburg](#), based on the physical town of Blacksburg, VA.
- **Graduate Research Assistant, [Naval Research Labs](#):** Alexandria, VA Jan 1995 to Oct 1998
  - Development and evaluation of a [new interaction method](#) called [pre-screen projection](#), which is a method of panning and zooming in a virtual world, where the viewing transformation is updated with respect to head position read from a virtual reality head tracker
  - Presented work at CHI ' 95, Denver, May 1995.
- **Graduate Research Assistant, [HCI Research Group, Virginia Tech](#):** Blacksburg, VA Aug 1994 to Dec 1994
  - Continued development of IDEAL, -- Interface Design Environment and Analysis Lattice -- an integrated tool environment that supports user-centered design and the process of formative evaluation of user interfaces
  - Incorporated real time VCR and camera control into IDEAL using the Sony' s VISCA command language.
  - Presented paper & demo of IDEAL at the Mid-Atlantic Human Factors Conference, March 1995.
- **Systems Analyst/Programmer, Consumer Products Group MIS, [Sony](#):** Park Ridge, NJ May 1994 to Mar 1996
  - Updated and managed an automated software upgrade system to install new software on a user' s machine over the network.
  - Developed database system to keep track of users and equipment in windows, and linked it to an SQL server running in OS/2
  - Assisted in administration of several OS/2 servers, and also developed and performed automated remote server updates.
- **Lab Manager/Programmer, PIP Software Support, [Sony](#):** Montvale, NJ May 1993 to Dec 1993
  - Developed Multimedia CD-Rom and CD-RomXA titles for the MMCD Player
  - Evaluated and enhanced user interfaces of existing CD-Rom titles
  - Designed Windows Application for Bitmap conversion, sizing, and maintenance
  - Implemented Unix commands in DOS for employees using both operating systems
  - Produced Electronic Data Discman titles for specialized clients using a markup language
- **Owner/President, [Avatar Multimedia](#):** Blacksburg, VA July 1992 to Present
  - Sell and assemble computer hardware and software for clients
  - Design and develop customized software and integrated systems to fit the needs of the customer
  - Develop custom Multimedia CD-ROM applications
  - Develop Interactive Web Pages for Clients
- **Programmer/Network Administrator, Security Systems, [Sony](#):** Montvale, NJ May 1992 to April 1993
  - Wrote a marketing report system on mainframe including Analysis of Commissionables, Backorder, Future Order, Salesman Order Detail, etc.
  - Set up and maintained network of PCs and Macs for all employees of Security Systems Division
  - Trained employees on the use of the network, packaged software, etc.
  - Controlled purchasing of equipment for Security Systems
- **Head Programmer/Network Manager, [DJR Enterprises](#):** Radford, VA January 1992 to Present
  - Programming application software in Dataflex Language (Both procedural and Object Oriented styles)
  - Consult with company employees and maintain network and hardware
- **Computer Lab Consultant, [Virginia Tech, Dept. of CS](#):** Blacksburg, VA January 1992 to May 1994
  - Assist students working on programming projects
  - Maintain computer systems in lab
- **Software Developer, H & H Production and Machining:** Sparta, NJ May 1992 to August 1992

- Developed integrated system for invoicing, inventory, and tracking of all orders
- **Personal Applications Programmer**, MIS, [Sony](#): Park Ridge, NJ May 1991 to August 1991
  - Designed schedule rotation system, integrated with a color plotter, for tracking data to be sent off-site
  - Performed major updates to MORS (MIS Operations Reporting System)
  - Worked in large data center in Networking, Operations, and Tape Library
- **Computer Operator/Programmer**, [TRW Inc.](#): Fairfield, NJ July 1990 to August 1990
  - Designed and maintained system to manage new account records consisting of sales, pricing, and payments between PCs and mainframe
- **Undergraduate Research in Human Computer Interaction**
  - Development of an array of tools to aid in the evaluation of user interfaces
  - Integration of Audio & Video into interpreted language Tcl/Tk, developed at UC Berkeley
  - Continued Development of QUANTUM -- Quick User Action Notation Tool for User-interface Management - which facilitates and automates the use of the UAN in GUI evaluation
  - Development of a world wide web server for Virginia Tech HCI Research
- **Undergraduate Research in C++**
  - Developed Object Oriented Graphic Windowing System in an X Windows environment
  - Designed Interfaces to enable Pascal students to use Graphics Windowing in class
  - Presented at the Virginia Computer Users Conference 1993

## Computer Skills:

- Computer Languages:
  - C, C++, MFC, Java, Tcl/Tk, Perl, TeX/LaTeX, JavaScript, Pascal, X Windows, HTML, 68000&80x86 Assembly, DataFlex, Lisp, Basic, REXX, FORTRAN, COBOL, FOCUS, JCL.
- Operating Systems/Hardware:
  - Unix (Solaris, Ultrix, BSD, Minix, Linux), Windows, MSDOS, OS/2, TSO, Macintosh.

## Publications:

- Amento, Brian; Whittaker, Steve. "Semantic Speech Editing." *Proceedings of CHI 2004* (Vienna, Austria)
- Whittaker, Steve; Amento, Brian. "Seeing what you are hearing: Co-ordinating responses to trouble reports in network troubleshooting.", *Proceedings of ECSCW 2003 (Helsinki, Finland)*.
- Amento, Brian; Terveen, Loren; Hill, Will; Hix, Deborah; Schulman, Robert. "Experiments in Social Data Mining: The TopicShop System", in *ACM Transactions on Computer-Human Interaction* (2003).
- Isbell, Charles; Amento, Brian; Whittaker, Steve; Bell, Gavin; Helfman, Jonathan. "Ishmail: Managing Massive Amounts of Mail", *Proceedings of UIST 2002* (Paris, France).
- Amento, Brian; Terveen, Loren; Hill, Will. "From PHOAKS to TopicShop: Experiments in Social Data Mining", *Interacting With Social Information Spaces*, Springer-Verlag, October 2002.
- Amento, Brian; Hill, Will; Terveen, Loren. "The Sound of One Hand: A Wrist-mounted Bio-acoustic Fingertip Gesture Interface", *Proceedings of CHI 2002* (Minneapolis, MN).
- Terveen, Loren; McMackin, Jessica; Amento, Brian; Hill, Will. "Specifying Preferences Based on User History", *Proceedings of CHI 2002* (Minneapolis, MN).
- Amento, Brian; Isbell, Charles; Whittaker, Steve; Bell, Gavin. "Ishmail: Designing Advanced Email Systems". *CSCW 2002* (New Orleans, LA)
- Whittaker, Steve; Hirschberg, Julia; Amento, Brian; Stead, Larry; Zamchick, Gary; Rosenberg, Aaron; Bacchiani, Michiel; Stark, Litza; Isenhour, Phillip. "SCANMail: A Voicemail Interface That Makes Speech Browsable, Readable, and Searchable", *Proceedings of CHI 2002* (Minneapolis, MN).

- Amento, B., Terveen, L.G., Hill, W.C., Hix, Deborah. "TopicShop: Enhanced Support for Evaluating and Organizing Collections of Web Sites", *Proceedings of UIST 2000* (San Diego, CA)
- Amento, B., Terveen, L.G., and Hill, W.C. "Does ' Authority' Mean Quality? Predicting Expert Quality Ratings of Web Sites". *Proceedings of SIGIR 2000*(Athens, Greece)
- Terveen, L.G., Hill, W.C., and Amento, B. [Constructing, Organizing, and Visualizing Collections of Topically Related Web Resources](#). in *ACM Transactions on Computer-Human Interaction* (1999). 6, 1 (Mar. 1999), Pages 67 - 94
- Amento, B., Hill, W., Terveen, L., Hix, D., and Ju, P. [An Empirical Evaluation of User Interfaces for Topic Management of Web Sites](#), in *Proceedings of CHI' 99*Pittsburgh, PA, May 1999), ACM Press, 552-559.
- Neale, D. C, Amento, B. S., and McGee, M.K. "Telepresence in ACTV Media Spaces". Submission to *HFES' 99*
- Amento, Brian.; Hix, Deborah; Templeman, James; Schmidt-Nielsen, Astrid and Sibert, Linda. "An Empirical Comparison of Interaction Techniques for Panning and Zooming in a Desktop Virtual Environment". Submission to *HFES' 99*
- Terveen, Loren; Hill, Will; Amento, Brian. "Collaborative Filtering to Locate, Comprehend, and Organize Collections of Web Sites." SIGART Bulletin, Volume 9, Issue 3-4 (December 1998).
- Neale, D. C., McGee, M. K., Amento, B. S., and Brooks, P. C. ["Making media spaces useful: Video support and telepresence"](#). (Tech. Report HCIL-98-01). Blacksburg, VA: Virginia Polytechnic Institute and State University, Human-Computer Interaction Laboratory.
- Terveen, Loren; Hill, Will and Amento, Brian, ["PHOAKS: A System for Sharing Recommendations"](#) CACM March 1997 Vol 40, No 3
- Terveen, Loren; Hill, Will and Amento, Brian, ["Building Task-Specific Interfaces to High Volume Conversational Data"](#), in *Proceedings of CHI' 97*Atlanta GA, March 1997), ACM Press, 226-233.
- McGee, M. K., Amento, B., Brooks, P., Harley, H. D. (1997). "Fitts and Virtual Reality: Evaluating Display and Input Devices with Fitts' Law." In Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society. Santa Monica, CA: Human Factors and Ergonomics Society.
- Kies, Jon; Amento, Brian and Struble, Craig, ["MOOsburg: Experiences with a Community-based MOO"](#) (Tech Report HCIL-96-03) Blacksburg, VA: Virginia Polytechnic Institute and State University, Human-Computer Interaction Laboratory.
- Amento, Brian and Struble Craig, "IDEAL: User Interface Analysis" Mid-Atlantic Human Factors 95

## Patents:

- Novel Service Based On Direct User Preference Data-Payment, Advertisement, Metering And Recommendation
- Control Shadows
- System and Method For Tracking Media Usage And Providing Related Information Based On The Media Usage
- System and Method For Administering a Payment Policy Based On Actual Media Usage
- Wireless Adaptive Finger Gesture Interface (WAFGI) For Controlling Wearable Communication Devices And Other Nearby Digital Devices
- Bioacoustic Control System, Method and Apparatus
- System And Method for Improving Multimedia Continuity Among Various Devices
- An Empirical Evaluation of User Interfaces for Topic Management of Web Sites
- Latency Reduction for Automatic Speech Recognition using Partial Multi-Pass Results

## **Honors and Activities:**

- **Technology Chair**, CSCW 2004, Chicago, IL
- **Technology Chair**, CHI 2003, Ft. Lauderdale, FL
- **Technology Chair**, CHI 2002, Minneapolis, MN
- **Student Volunteer Co-Chair**, CSCW 2000, Philadelphia, PA
- **President**, Upsilon Pi Epsilon, National Computer Science Honorary Society
- **Vice President**, Graduate Student Association
- **Graduate Representative to Faculty**, Virginia Tech
- **Member**, Computer Resource Committee, Virginia Tech
- **Member**, Phi Kappa Phi, National Honor Society
- **Instructor**, Governor' s School of Virginia
- **Member**, Golden Key National Honor Society
- **Member**, Gamma Beta Phi Society, Honorary Community Service Organization
- **Member**, University Honors Program, Virginia Tech
- **Member**, Computer Network Access Committee, Virginia Tech
- **Member**, ACM, Association of Computing Machinery, Local and National Chapters
- **Member**, National Eagle Scout Association, Boy Scouts of America
- **Dean' s List** Virginia Tech
- **Investment in Excellence Scholarship**, Virginia Tech