Collaborative Filtering To

Locate, Comprehend, and Organize Collections of Web Sites
Introduction

Finding information is a common and fundamental task for Internet users. Search engines such as AltaVista and indices such as Yahoo? are the tools commonly used for this task. Users type in queries using keywords to indicate their interest or navigate through hierarchical directories of topics, eventually obtaining (usually quite large) collections of Web pages.

What then? How do users make sense of the results? No one can wade through hundreds or thousands of pages, and few people can even get through more than a handful. So how does one tell which are the best pages, or what kind of information each page contains, or whether all interesting pages are included? Furthermore, people who are truly interested in a topic usually want to boil it down to a manageable collection of perhaps 10 to 20 key sites and often want to share their results with others. Current tools provide scant support for the process of exploring, comprehending, and organizing collections.

Our research investigates algorithms and interfaces to meet these user needs. We have developed collaborative filtering algorithms that construct collections of topically relevant Web sites and create site profiles that can be used to estimate the relevance, quality, and functionality of a site. We have developed interfaces that make it easy for users to explore and organize collections of sites and a data format that makes it simple to share collections.

The first system we developed is PHOAKS (People Helping One Another Know Stuff) [Hill and Terveen 1996, Terveen et al. 1997]. PHOAKS exploits the thousands of public discussions that take place daily on Usenet newsgroups. One of the things participants in these discussions do is discuss and recommend Web pages related to the newsgroup's topic. PHOAKS processes messages from 6,000 newsgroups and applies rules to identify recommended Web pages. The PHOAKS Web site (http://www.phoaks.com) contains more than 100,000 recommendations and associated contextual information (who recommended the Web pages and what they said about the pages). Whereas PHOAKS mines Usenet messages, our second system, TopicShop [Amento 1999, Terveen and Hill 1998a, Terveen and Hill 1998b], analyzes the structure of the Web itself. Starting from a set of initial "seed" pages for a topic (perhaps taken from PHOAKS), TopicShop uses a Web crawler to group individual pages into sites (structured multimedia documents), build site profiles, and discover new, topically related sites. TopicShop also includes several user interfaces for exploring and organizing sites.

Although PHOAKS and TopicShop are independent systems, the original motivation for TopicShop was to overcome some shortcomings of PHOAKS. Therefore, we begin by discussing PHOAKS, first describing how it works, then examining the set of Web pages PHOAKS found for a particular newsgroup, to highlight both the benefits and limits of its approach. We then will show how TopicShop addresses these limits, describe the TopicShop algorithm, illustrate how it improves on the results available from PHOAKS, and present the TopicShop interface.

PHOAKS

The basic job of PHOAKS is to search messages for mentions of Web pages (URLs) and identify which mentions are recommendations. A URL mention counts as a recommendation if it passes a number of tests. First, the message must not be cross-posted to too many newsgroups. Messages posted to a large number of groups often are worthless (spam), and, even if they are not, they are usually so general that...
they are not likely to be of much interest to any of the groups. Second, if the URL is part of the message signature, it is not counted as a recommendation. Third, if the URL occurs in a quoted section of a previous message, it is ruled out. Fourth, if the textual context surrounding the URL contains word markers that indicate that it is being recommended and does not contain markers that indicate that it is being advertised or announced (i.e., self-promotion), then it is categorized as a recommendation. We have developed rather complicated rules that implement this basic strategy.

We carried out experiments [Hill and Terveen 1996] that provided empirical evidence of the plausibility of our approach. In particular, we showed that Usenet messages are a significant source of recommendations of Web resources,

* Recommendations can be recognized automatically with nearly 90 percent accuracy,

* Some resources are recommended by more than one person, and

* The number of distinct recommenders of a resource is a plausible measure of the quality of the resource.
Figure 1 shows the PHOAKS page for the newsgroup rec.music.dylan, which discusses the music of Bob Dylan. Recommended Web pages are ordered by the number of distinct recommenders. By clicking on the appropriate links, a user can browse to a recommended page or explore the set of posters who recommended a particular page (and then go on to see what else they recommended) or the message context that surrounded the recommendation. This additional contextual information helps users judge the quality of a recommendation and sometimes indicates what a page is good for.

Following is a list of the first 20 recommended pages from rec.music.dylan, including comments that show both the utility and the limits of harvesting Web pages from Usenet newsgroups.

1. Bob Dylan - Bob Links - Fall & Winter
   This is a very useful page from a very useful site; however, this is not the root page of the site.

2. Bob Dylan Boot Database (BDBDB)
   This is not the root page of the site.

3. Bob Dylan - Bringing It All Back Homepage

4. Same title different location...
   This Web site moved from one server to another, and messages mentioning both locations were posted to the newsgroup - this shows that different URLs may refer to the same document.

5. Never played live

6. DYLANOLOGY: THE STUDY OF AN AIDS RIDDEN EX...
   This is a low-quality, low-credibility site; nothing prevents people from making poor recommendations (or even anti-recommendations) although our empirical studies showed that these are extremely rare.

7. Roots Routes and Ramblings
   What is this page about? Can you tell from just the title?

8. ambulances

9. CD Outlets and Other Resources
   This is not the root page of the site, and other pages from the same site are also included in the list.

10. Bob Dylan Chords

11. Frequently Asked Questions
   This is not the root page of the site, and other pages from the same site are also included in the list.

12. CDnow : Main : Homepage
   This page points us toward a related, more general resource; it is not about Bob Dylan, but outlets to purchase music are mentioned occasionally on all music-related groups.

13. Same title different location...
   This is a different URL for the same document.

14. Bob Dylan - Expecting Rain
   This is an important site, with some unique content, yet if rated simply by number of recommenders it does not make it very high in the ordering.

15. Liszt the mailing list directory
   Somewhat similar to the mention of CDnow; since the newsgroup discussion is carried out using Internet technology, mentions of and discussions about the technology sometimes occur, but this is not necessarily related to the newsgroup topic at all.

16. A Cat's Bob Dylan Links

17. EDLIS Parties & Gatherings On Tour

18. bobdylan.com
   This is the official site; it is very rich in content, some of it unique and very valuable, yet again, the number-of-recommenders metric did not rate it highly.

19. BOB DYLAN: A Portrait of the Artist as a Y...

20. MP3.com - The MP3 resource on the Internet
   See note for item 15.

Let us now articulate the issues we have encountered and identify a set of goals for creating more useful collections of sites.

X For many purposes, the Web page is the wrong basic unit. Individuals and organizations create Web sites. Our goal is to group Web pages into sites, without losing sight of or obscuring the fact that certain pages within the site have been singled out as being of particular interest.

X Usenet newsgroups are a somewhat noisy source of recommendations, especially for judging the quality of a Web page. That is, while several people indeed may have recommended a page (at least according to the PHOAKS classification rules), the page may be of low quality or peripheral relevance. Conversely, some relevant, high-quality sites are seldom or never recommended. This can occur for various reasons; for instance, if one person, such as the site maintainer or the newsgroup FAQ maintainer, posts the site regularly (recall that if someone recommends his or her own site, PHOAKS will classify it as a self-promotion and will not count it as a recommendation), no one else needs to. To overcome these problems, we seek additional, more...
TopicShop consists of one main algorithmic component and several user interfaces:

- A Web crawler and site analyzer that follows links from a set of initial "seed" pages, heuristically groups pages into sites, builds profiles of site content and structure, and discovers new sites that are structurally related (and thus likely to be topically related) to the seeds. Other systems with somewhat similar goals are described by Kleinberg [1998] and on the twURL Web site (http://www.twurl.com);
- A Java applet that lets users interactively evoke and control the crawler/analyzer and offers incremental presentation of the results so users can immediately begin to explore sites;
- The TopicShop Auditorium Visualization, an interactive graph visualization that lets users quickly identify topically central sites and investigate relationships between sites; and
- The TopicShop Explorer, a program modeled on Microsoft Windows Explorer, which provides simple, yet powerful means for users to explore, browse, and share collections of sites.

As the crawler works, it heuristically groups Web pages into sites. A site is an organized collection of pages on a specific topic maintained by a single person or group. Sites have structure, with pages that play certain roles (front door, table of contents, index). A site is not the same as a domain: for example, thousands of sites are hosted on http://www.geocities.com or http://www.tripod.com. And what counts as a site may be context dependent. For example, if one is taking a survey of researchers, http://www.media.mit.edu might well be considered a site, whereas if one is investigating social filtering projects, sites of individual researchers sites hosted on http://www.media.mit.edu are probably the proper units.

The last observation suggested a way to operationalize the definition of a site. As the crawler/analyzer is constructing a collection, the relevant known context is the set of URLs that the expanded sites have linked to. The intuition is that if sites in the collection link to two URLs, one of which is in a directory that contains the other, then they are likely to be from the same site. More precisely, if URL A has been linked to and URL A/B has been linked to, then assume that A is the root page of the site and that A/B is an internal URL.

This rule applies recursively, so the URLs A/B/C, A/B, and A would be merged into a site with root page A and internal pages A/B and A/B/C.

In addition to this rule, our system uses a number of heuristics. General heuristics include the convention that the tilde (~) character indicates different user sites (therefore, http://www.research.att.com/~terveen and http://www.research.att.com/~willhill are usually distinct sites). Site-specific heuristics encode knowledge that enables the identification of distinct user sites, hosted on large servers such as GeoCities or Tripod. Of course, the site aggregation rule and our heuristics can fail; therefore, we also have finely tuned quality metrics and techniques to discover new sites that are likely to be on the same topic as a given set of sites.

Finally, page titles often do not communicate what a site is good for. We care about not only quality, but also function. For example, bobdylan.com, the official Sony site, contains RealAudio versions of many otherwise unavailable live performances. "Bob Links" contains a huge number of links to other sites as well as tour dates and concert set lists and reviews. And "Roots Routes and Ramblings" contains information about Dylan's musical roots and influences. Our goal is to construct profiles of site content and structure that make it easy for users to determine not just their quality, but also their function.

These were the motivations that led us to develop the TopicShop system.
explored heuristics to detect when sites should be split or merged.

The crawler analyzes the content of pages it fetches in order to build profiles of site content and structure. Profiles include the following data:

- Title (of the site's root page);
- Thumbnail image (of the site's root page);
- Links to and from other sites;
- List and count of internal HTML pages, images, audio files, and movie files; and
- Count of user-specified or domain-specific keywords.

Let us now consider the results of applying our algorithm to the PHOAKS pages for rec.music.dylan. In the following list, we order the sites within the collection by the number of in-links, that is, the number of other sites within the collection that point to them. Recall that we view links as a form of implicit recommendation, so this can be seen as ordering sites by the number of recommendations they have received from their peers, that is, other individuals who have created Bob Dylan sites. (As we shall soon see, our interfaces make it easy to order sites using any of the profile attributes. Different purposes require different orderings; for example, if you were looking for sound clips from a favorite TV series, you could sort by the number of audio files.) Following are the top 10 sites, with annotations that explain what the crawler/analyser has done for us.

Five of the last six sites in the list were discovered by the crawler. They contain useful information such as lists of cover songs that Dylan has performed and essays exploring the role of Christianity in Dylan's writing and performance.

Finally, what about the items in the PHOAKS top 20 whose presence we questioned? Where did they end up in the ordering? CDnow (Number 12) dropped into the 40s, and DYLANOLOGY (Number 6), "Liszt" (Number 15) and MP3.com (Number 20) fell into the 70s.

Although a simple list of sites, ranked by number of in-links, looks impressive, it still does not meet all users' needs. We designed the TopicShop Explorer (see Figure 2) to support users in exploring, organizing, and sharing collections of sites. It offers two main views. The details view shows site profile information, allowing users to quickly and easily explore and order sites. The icons view allows users to arrange icons spatially, quickly organizing and grouping them even before explicit categories emerge. The views are linked, so selections and operations in one view are reflected in the other. We will explain the user interface further by detailing three main design goals.

The TopicShop Explorer has three main design goals.

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—solely on the basis of titles and, sometimes, brief textual annotations. (A chief complaint of subjects in a study by Abrams et al. [1998] 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, thumbnails 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 the number of in-links, out-links, images, etc.) simply by clicking on the label at the top of the column. Users also can “zoom in” on the internal structure of a site, getting a details view of internal pages of the site (analogous to the view of the sites within a collection shown here). For example, this can take users directly to the links page of a site or to a page that contains audio samples or occurrences of interesting phrases. 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 and Barreau [1995] found that users of graphical file systems preferred spatial location as a technique for organizing their files. We believe that 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 become explicit, users can create folders to contain sites in each of the categories.

3. Integrate the task of managing collections of sites into a user’s normal computing and communications environment. The TopicShop Explorer is based on Microsoft Windows Explorer. This enables Windows users to apply their existing knowledge, resulting in little learning time and high ease of use. Further, the data format for representing sites makes it very easy for collections to be shared. All data for a site are stored in a file, and a collection is simply a folder of these files. Thus collections can be shared in all the normal ways files
are shared, for example, by being zipped and e-mailed.

We conducted an empirical study comparing user performance with TopicShop vs. Yahoo. TopicShop subjects found more than 80 percent more high-quality sites (where quality was determined by independent expert judgments) while browsing only 81 percent as many sites and completing their task in 89 percent 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—were the key to these results, as users exploited it to identify the most promising sites quickly and easily.

Conclusions
Making sense of collections of Web sites is an increasingly important problem. We attacked this problem in two ways. First, we developed algorithms that construct collections of topically related sites and that create profiles of site content and structure. Second, we created interfaces that make it easy for users to explore and organize sites. We built two systems, PHOAKS and TopicShop, that implement this methodology. We conducted empirical studies showing the utility of both systems. PHOAKS has been available on the Web for several years and draws about 5,000 visitors a day. TopicShop was developed more recently, and we are moving ahead with plans to also deploy it on the Web.

References


