Constructing, organizing, and visualizing collections of topically related web resources

Loren Terveen, Will Hill, and Brian Amento[1]

AT&T Labs - Research
180 Park Avenue, P.O. Box 971
Florham Park, NJ 07932-0971 USA
+1 973 360 {8343, 8342}
{terveen, willhill, s_brian}@research.att.com

ABSTRACT

For many purposes, the Web page is too small a unit of interaction and analysis. Web sites are structured multimedia documents consisting of many pages, and users often are interested in obtaining and evaluating entire collections of topically related sites. Once such a collection is obtained, users face the challenge of exploring, comprehending, and organizing the items. We report four innovations that address these user needs.

- We replaced the web page with the web *site* as the basic unit of interaction and analysis.
- We defined a new information structure, the *clan graph*, that groups together sets of related sites.
- We augment the representation of a site with a *site profile*, information about site structure and content that helps inform user evaluation of a site.
- We invented a new graph visualization, the *auditorium visualization*, that reveals important structural and content properties of sites within a clan graph.

Detailed analysis and user studies document the utility of this approach. The clan graph construction algorithm tends to filter out irrelevant sites and discover additional relevant items. The auditorium visualization, augmented with drill down capabilities to explore site profile data, helps users to find high-quality sites as well as sites that serve a particular function.

KEYWORDS

Social filtering, collaborative filtering, information access, information retrieval, information visualization, human-computer interaction, computer supported cooperative work, social network analysis, co-citation analysis.

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 time 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 multimedia documents relating to the topics.

Our goal is to address these user needs. We replaced the Web page with the *site* — a structured collection of pages, a multimedia document — as the basic unit of interaction and analysis. A site is more appropriate for several reasons. (1) A site usually contains a coherent body of content on a given topic (e.g., song lyrics, episode guides for a TV show, current weather conditions), divided into pages to

[1] Also with the Department of Computer Science, Virginia Tech

ease navigation and download time. Thus, users want to know what's available at a given site, not a single page. (2) Most hyperlinks *to* a site point to the "front door" page, while most links *from* a site come from the site's index page. Thus to analyze inter-site structure appropriately (which is our goal), we must correctly group pages into sites.

Second, we defined a new information structure, the *clan graph*, to represent collections of densely connected sites. The clan graph has a clear intuitive motivation based on concepts from social network analysis, social filtering, and co-citation analysis. A clan graph is defined in terms of a user specified set of seed (example) sites and is constructed by following hypertext links from the seeds. It is easy for users to specify seeds, e.g., they may get them from their bookmarks file, from an index page they found on the web, or from a search engine. And the clan graph construction algorithm is tolerant of "noise" in the seeds: a few off-topic seeds will not affect the quality of the graph.

Third, as our algorithm adds sites to the clan graph, it also constructs *site profiles*. These profiles contain information about the amount and type of content contained on each site as well as the links between sites. The profile data helps to inform user evaluation of the quality and function of a site.

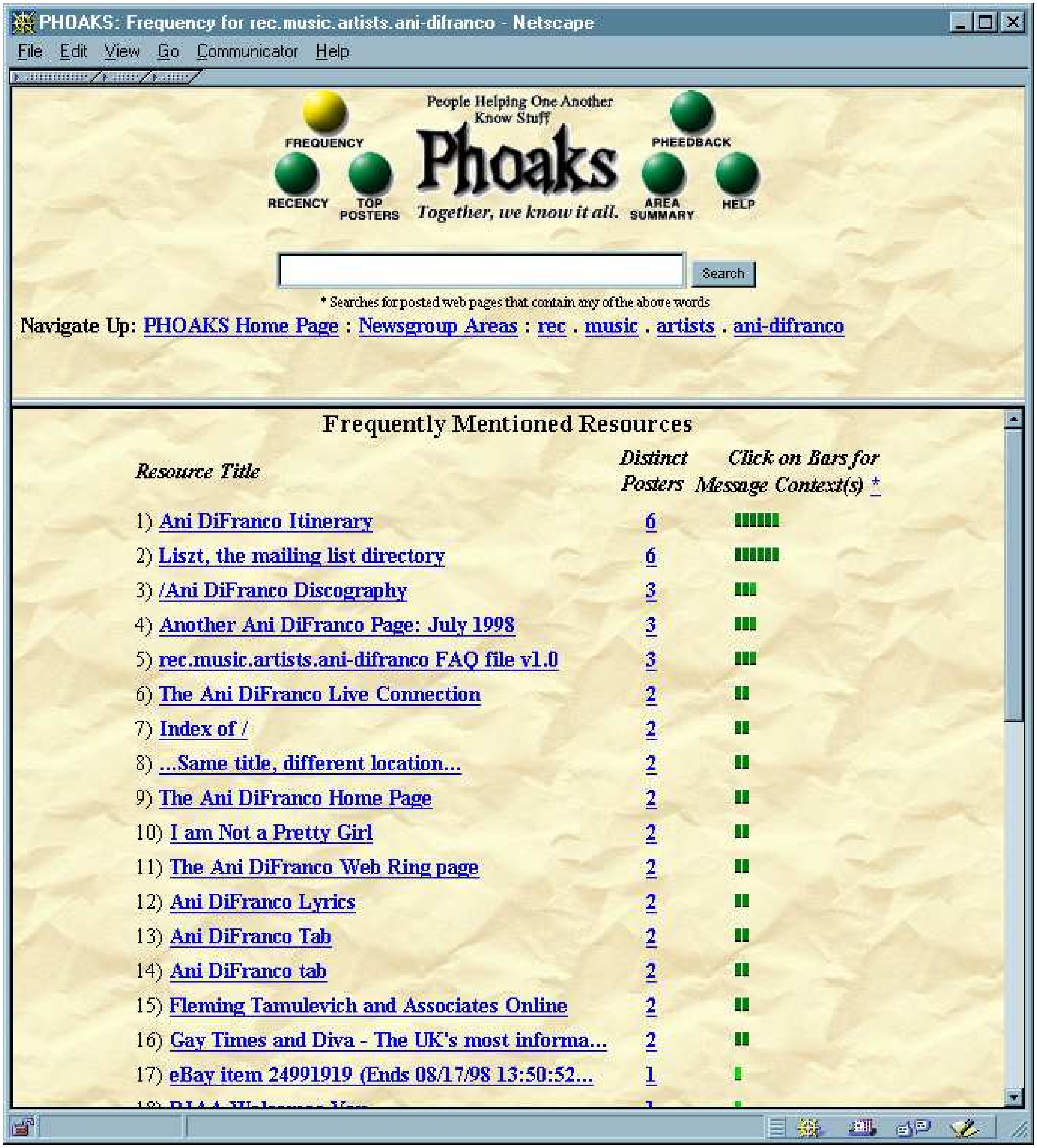
Fourth, to enable users to comprehend and manage the information we extract, we have developed the *auditorium visualization*, which communicates key information such as whether a site is structurally central or peripheral, whether a site is more of a content provider or index, important internal structure of a site, and how sites link together. Figure 4 (which we discuss in a later section) shows an example auditorium visualization.

Our system is implemented in Java. We have built and analyzed clan graphs for dozens of topics, performed experiments to evaluate our algorithms, and iterated our interface design significantly in response to user feedback.

PREVIOUS WORK: PHOAKS

PHOAKS (People Helping One Another Know Stuff)[14,29] was the first system we created to address the problem of constructing collections of web pages. It is based on the premise that an effective way to find good items of information about a given topic is to ask the experts. One problem is that one often does not know any experts for a particular topic. PHOAKS looked to Usenet news for the solution. It searches messages in thousands of newsgroups for mentions of web pages and applies a set of rules to identify those mentions that were done for the purpose of recommending a web page. Web pages are ranked within a topic by the number of different individuals who recommended them. 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 was some correlation between the number of recommenders of a web page and other metrics of web page quality.

Figure 1 shows the PHOAKS page for the newsgroup rec.music.artists.ani-difranco, which covers the music of Ani DiFranco. 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 can help users judge the quality of a recommended web page and sometimes can give an indication of what a page is good for (i.e., the type of information it contains).


PHOAKS: Frequency for rec.music.artists.ani-difranco - Netscape
File Edit View Go Communicator Help
People Helping One Another Know Stuff
Phoaks
Together, we know it all.
FREQUENCY
RECENCY
TOP POSTERS
PHEEDBACK
AREA SUMMARY
HELP
Search

* Searches for posted web pages that contain any of the above words

Navigate Up: PHOAKS Home Page : Newsgroup Areas : rec . music . artists . ani-difranco

Frequently Mentioned Resources

Resource Title	Distinct Posters	Click on Bars for Message Context(s) *
1) Ani DiFranco Itinerary	6	▮▮▮▮▮▮
2) Liszt, the mailing list directory	6	▮▮▮▮▮▮
3) /Ani DiFranco Discography	3	▮▮▮
4) Another Ani DiFranco Page: July 1998	3	▮▮▮
5) rec.music.artists.ani-difranco FAQ file v1.0	3	▮▮▮
6) The Ani DiFranco Live Connection	2	▮▮
7) Index of /	2	▮▮
8) ...Same title, different location...	2	▮▮
9) The Ani DiFranco Home Page	2	▮▮
10) I am Not a Pretty Girl	2	▮▮
11) The Ani DiFranco Web Ring page	2	▮▮
12) Ani DiFranco Lyrics	2	▮▮
13) Ani DiFranco Tab	2	▮▮
14) Ani DiFranco tab	2	▮▮
15) Fleming Tamulevich and Associates Online	2	▮▮
16) Gay Times and Diva - The UK's most informa...	2	▮▮
17) eBay item 24991919 (Ends 08/17/98 13:50:52...	1	▮

Figure 1: PHOAKS page for Ani DiFranco

Through reflecting on the information in a PHOAKS page, we can identify a set of dimensions for understanding the design space for this type of system.

- *Unit* – what is the basic unit of the items within the collection? PHOAKS collections consist of web pages (i.e., URLs). However, for many purposes, the web *page* is the wrong basic unit. Individuals and organizations create web *sites*, coherent, structured multimedia documents consisting of many pages. Figure 1 contains some examples that illustrate this point. Item 9 is the front page of a site that contains (among various pages) item 5, and item 15 is the front page of a site that contains item 1. We want to group web pages into sites, and use the site as the basic unit for presentation to users. However, at the same time, we must not lose or obscure the fact that certain pages within the site have been singled out as being of particular interest -- after all, 6 different people found item 1 (which provides information about Ani DiFranco tour dates) worth mentioning, while only 2 people mentioned item 15 (the home page of a booking agency that handles Ani DiFranco and other artists).

- *Composition* – How are collections built? How do items get included in the collection? How many items are included? The goal is to get as many relevant items as possible (to maximize *recall*) while including few irrelevant items (maximizing *precision*). An item is included in PHOAKS when (1) it is mentioned in the newsgroup, and (2) the categorization rules classify it as a recommendation. We have identified several situations where items of limited relevance may be mentioned fairly frequently and where items of high relevance are not.

  Item 2 illustrates a common type of peripherally relevant material: since the newsgroup discussion is carried out using internet technology, mentions of and discussions about the technology sometimes occur. However, such discussions don’t concern the newsgroup’s topic directly. Item 16 illustrates another common occurrence: a general publication (covering music or a particular lifestyle, for example) may contain one or two articles about the topic of interest (here, Ani DiFranco) but the publication itself is much more general.

  Conversely, some relevant sites are recommended infrequently or not at all. This occurs for various reasons, e.g., one person such as the site maintainer or the newsgroup FAQ maintainer posts the site regularly so no one else feels the need. However, in PHOAKS, mentions by the site maintainer may not be counted as recommendations, since PHOAKS has a “no self promotion” rule that detects instances of someone recommending his or her own site. We have observed several cases of this occurring in the Bob Dylan newsgroup.

- *Organization* – how are items within the collection organized? Are they clustered or categorized? Are they ranked, along one or more dimensions? These questions are crucial since two primary user concerns are locating the highest quality items and locating items that serve a particular purpose.

  PHOAKS collections are ordered using a single ranking metric, the number of distinct individuals who recommended an item. We have some evidence for the plausibility of this metric. In addition to simply manually examining many topics we were familiar with as a sanity check, we also showed a positive correlation between the number of recommenders an URL received and the likelihood that the URL appeared in the newsgroup Frequently Asked Questions (FAQ) list [29]. Since the FAQ is maintained by a human topic expert, inclusion in the FAQ is evidence for the relevance and quality of a URL.

  However, the number of recommenders metric does have its limits. For example, we already have noted that if an item is posted regularly by the individual who maintains the FAQ, no one else may feel the need to mention it. So, even though the item may be of high quality, it ends up with a low rank. Further, not all recommenders are created equal: anyone who has participated in Usenet or other online discussion forums knows that participants vary wildly in their expertise, credibility, and helpfulness, but a simple tally of recommenders does not highlight this. Finally, often many items have two or three recommenders, thus the metric provides little or no discrimination among them.

  Within PHOAKS, the category or function of items (i.e., what an item is good for) can be inferred from titles and from the message context surrounding item recommendations. Titles may or may not communicate what a page is good for. "Ani DiFranco Discography", "Ani DiFranco Lyrics", and "Ani DiFranco Tab" all make the type of information they contain clear, while "I Am Not A Pretty Girl" does not. (Interestingly, the latter page is the front door of a comprehensive site that includes lyrics, reviews, and biographical information. This reinforces our point that the site, not the page, is a better basic unit.) It is our goal to construct profiles of site content and structure that make it easy for users to evaluate sites, helping them to determine both site quality and function.

Trying to solve the problems highlighted in this analysis was the initial motivation for the work we report in this paper. Before we describe this work, we analyze other systems for finding and exploring web information, first discussing widely used web tools, then research prototypes and techniques. We shall argue that these systems suffer from many of the same shortcomings as PHOAKS.

RELATED WORK: STATE OF THE ART WEB TOOLS

The main tools people use to find information on the web are search engines like AltaVista, general directories like Yahoo, and topic-specific directories (e.g., http://www.execpc.com/~billp61/boblink.html ("Bob Dylan – Bob Links") – is a comprehensive index of Bob Dylan information on the web). We will analyze these three methods and PHOAKS using the dimensions of unit, composition, and organization; we also will note the cost – that is, the effort (human or computational) involved in constructing a collection.

	Unit	*Composition*	*Organization*	*Cost*
PHOAKS	Recommended web page (URL)	The 40 pages recommended by the most distinct individuals are included. The composition is thus dependent on the pages people choose to mention and the quality of the PHOAKS categorization rules.	Pages are ordered by the number of recommenders. Titles and message context can be used to infer page function.	Computational cost of processing messages and storing data.
Search Engines	Any indexed page	All pages that match (in the information retrieval sense) the terms in the query are included. Typically, 1000s or 10,000s of items are returned. Queries are run against an index of the web created by a web crawler. Composition is thus dependent on the web crawler and the IR algorithm.	Pages are ordered based on similarity to the query (as computed by the IR algorithm). Users infer page function based on page titles and the first line or two of text from the page.	Computational cost of indexing the web, storing data, and processing queries.
General Directories	Human selected page	A human "librarian" selects the pages to include within a particular category. The librarian may seek out pages as well as accept pages submitted by the public. The involvement of a human librarian means that included pages are likely to be on topic, but some relevant items are likely to be missed. Typically, Yahoo categories contain from 10s to 100s of pages.	In Yahoo, pages are ordered alphabetically, with the additional organizing principle that categories may be divided into sub-categories. A brief textual annotation is included for many pages; this may help users to determine page function.	The major cost is the human effort involved in finding, categorizing, annotating pages, and maintaining collections.
Topic Specific Directories	Human selected page	Similar to a general directory, items are included because the maintainer of the directory selects them. However, topic-specific directories generally are maintained as "labor of love" for someone who cares about the topic, so they are apt to work harder to find all relevant pages. Again, typical size is 10s to 100s of pages.	Pages are organized and annotated at the discretion of the site maintainer.	Same as above.

Table 1: Analysis of search engines and directories

Let us summarize the chief strengths and weaknesses of search engines and (both general and topic-specific) directories. Search engines offer much more data; thus it is possible they may include relevant items missed in a directory. However, this comes at a price: search results nearly always contain much irrelevant information (due to the ambiguity of most queries). Further, search results are not very well organized and often contain duplicate pages and dead links. Most important, users have very little information they can use to evaluate the quality and function of the pages.

Therefore, directories like Yahoo appears to have the edge. They give smaller collections, with all items likely to be relevant to the topic – after all, the items were selected by a human being specifically for their relevance to the topic. Further, they often include some information that can help users in evaluating item quality and function, such as short textual annotations and explicit categories. However, most users do not consult more than a handful of pages, so trying to make sense of a collection of even 50 items is a fairly daunting task. Further, we have observed informally that Yahoo categories often are missing some important pages for a topic – no surprise when one thinks of the effort it takes to keep abreast of all the information on the web. Surprisingly, we also found in an experiment that users did not find Yahoo textual annotations very helpful [1]. Most important, though, is the sheer amount of human effort required to construct and maintain a good collection of items for a topic.

Our goal remains: we seek computational means to construct high-quality collections of items for specified topics, where the items are organized and augmented with information that users can use to evaluate the quality and function of the items. We next consider efforts in the research arena that have addressed some of these challenges.

RELATED WORK: RESEARCH INTO EXTRACTING AND VISUALIZING HIGH LEVEL STRUCTURES FROM THE WEB

Many researchers have sought to define useful, higher-level structures that can be extracted from the web (or, more generally, any hyperlinked collection of documents), such as "collections" [26], "localities" [24], "patches" or "books"[6]. In terms of the dimensions we have introduced, this work has concentrated mostly on means of organizing collections of items, in particular on algorithms to extract useful structures. Less attention has been paid to the composition of collections – often algorithms are applied directly to the results returned by querying a search engine. If more than that is done, it usually consists of a simple expansion of the query results, typically by following one level of links out from the resulting pages. Finally, the basic analytic unit almost universally remains the web page.

However, one additional concern is of great importance in the research community: visualization and interface techniques to help users construct, evolve, explore, and organize collections. Where the web-based tools we surveyed previously were limited by the constraints of HTML and typically simply presented a linear list of items, research prototypes labor under no such limits. Thus, algorithms for organizing collections of web pages and interface techniques for exploring them will be the two main themes discussed in this section.

Algorithms to organize collections

Carrière & Kazman [7] and Kleinberg [17] both build collections by issuing a query to a search engine and augmenting the results with all pages that link to or are linked to by any page in the original set of results. Carrière & Kazman argue that page connectivity is a crucial resource for organizing collections and judging page quality. Their WebQuery system sorts pages into equivalence classes based on their total degree (the number of other pages in the collection they are connected with) and displays the pages in a "bullseye" layout, a series of concentric circles each containing pages of equal degree

Kleinberg defines algorithms that identify *authoritative* and *hub* pages within a hypertext. Authorities and hubs are mutually dependent: a good authority is a page that is linked to by many hubs, and a good hub is one that links to many authorities. An equilibrium algorithm is used to identify hubs and authorities in a hypertext collection. Kleinberg also showed that his algorithm could be used to cluster pages within a collection, in effect disambiguating the query that generated the collection. For example, a query on "Jaguar" returned items concerning the animal, the car, and the NFL team, but Kleinberg's algorithm splits the pages into three sets, corresponding to the three meanings. Other researchers have followed up on Kleinberg's work, experimenting with and improving the original algorithm. One interesting development that is relevant to our work is that the algorithms of both Chakrabarti et al [8] and Bharat & Henzinger [4] implicitly incorporate a notion of a unit above the individual web page. For example, Chakrabarti et al's algorithm diminishes the weight assigned to a link (in conferring authority) if both the source and target page are on the same "logical website". As we understand it, the logical website is equivalent to the domain, that is, the part of the URL before the first "/", e.g., cnn.com, att.com, microsoft.com, salonmagazine.com, yahoo.com, etc.

Gerhart's twURL [32] system attempts to move beyond the page as the basic unit, exploring such concepts as "site" (e.g., nsf.gov) and "host" (e.g., ccr.cise.nsf.gov), based on the controlling organization. twURL also clusters items based on pre-defined vocabularies and organizes items into outlines based on properties such as site, host, and number of incoming links.

Spertus[28] experimented with algorithms that analyze links between web pages to find pages related to a given set of pages and to infer the topic and function of pages. Marchiori [21] developed an algorithm that used information about links (and the contents of linked-to pages) to reorder the results returned by search engines.

Pitkow and Pirolli [26] report cluster algorithms based on co-citation analysis[11]. The intuition is that if two documents, say A and B, are both cited by a third document, this is evidence that A and B are related. The more often a pair of documents is co-cited, the stronger the relationship. They applied two algorithms to Georgia Tech's Graphic Visualization and Usability Center web site and were able to identify interesting clusters.

Pirolli, Pitkow, and Rao [24] defined a set of functional roles that web pages can play, such as "head" (roughly the "front door" of a group of related pages), "index", and "content". They then developed an algorithm that used hyperlink structure, text similarity, and user access data to categorize pages into the various roles. They applied these algorithms to the Xerox web site and were able to categorize pages with good accuracy.

Mukherjea et al [23] and Botafogo et al [5] report on algorithms for analyzing arbitrary networks, splitting them into structures (such as "pre-trees" or hierarchies) that are easier for users to visualize and navigate.

Interfaces for constructing, exploring, and organizing collections

Card, Robertson, and York [6] describe the WebBook, which uses a book metaphor to group a collection of related web pages for viewing and interaction, and the WebForager, an interface that lets users view and manage multiple WebBooks. In addition to these novel interfaces, they also presented a set of automatic methods for generating collections (WebBooks) of related pages, such as recursively following all relative links from a specified web page, following all (absolute) links from a page one level, extracting "book-like" structures by following "next" and "previous", and grouping pages returned from a search query.

Mackinlay, Rao, and Card [20] developed a novel interface for accessing articles from a citation database. The central UI object is a "Butterfly", which represents one article, its references, and its citers. 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 a particular relationship to the current article.

Many other research efforts have proposed novel ways to view and navigate information structures. The Navigational View Builder [22] combines structural and content analysis to support four viewing strategies: binding, clustering, filtering and hierarchization. Through the extensive use of single user operations on multiple windows, the Elastic Windows browser [16] provides efficient overview and sense of current location in information structures. Lamping et al [18] explored hyperbolic tree visualization of information structures. WebCutter [19] builds a collection of URLs based on text similarity metrics, then presents the results in tree, star, and fisheye views. Furnas [10] presents a theory of how to create structures that are easy for users to navigate.

Still other researchers have created interfaces to support users in constructing, evolving, and managing collections of information resources. SenseMaker [2] focuses on supporting users in the contextual evolution of their interest in a topic. It attempts to make it easy to evolve a collection, e.g., expanding it by query-by-example operations or limiting it by applying a filter. Scatter/Gather [25] supports the browsing of large collections of text, allowing users to iteratively reveal topic structure and locate desirable documents. Hightower et al [12] addressed the observation that users often return to previously visited pages. They used Pad++ [3] to implement PadPrints, browser companion software that presents a zoomable interface to a user's browsing history.

Discussion

There are some similarities between these research efforts and ours. We are focusing on a structural analysis of the relationship between items, like [4,7,17,19,26,28], although we work with links between sites, not pages. We are interested in the functional roles a web page can play, like [17,24]. As in [6], seed sites in our system serve as "growth sites" that form the basis for a particular type of "related reference query" [2] that retrieves a structure of related sites. Finally, like [6] we are interested in citations between documents.

Our work also has important differences. First, we move from the page to the site as the basic unit. Only [32] has been explicit in making this move. (Other work has implicitly used site-like units, for example, the algorithms presented in [4] and [8] and some of the interface actions in [6].) Grouping pages into sites – in other words, determining the boundaries of an online document – is a difficult problem. Our approach differs from Gerhart's by grouping pages based on context rather than by a fixed organizational unit. We present our heuristics for this problem below.

Second, we define a new information structure – the *clan graph* (defined below) – that formalizes the composition of a collection of items related to an initial set, the "seeds". (We thus presume that users can obtain an initial collection somehow, whether from a search engine or directory. As we shall see, our algorithms can take seed sets that are both noisy and incomplete and still generate high-quality collections, exploiting multiple seeds to achieve a kind of "triangulation" effect.) Most previous work either takes the collection as a given (e.g., all the web pages rooted at a particular URL like www.xerox.com), defines the collection as an augmentation of the results of a search engine query, or provides interface techniques that essentially leave the construction of a collection up to the user. Card et al [6] do offer some automated techniques for creating collections, but the basic unit out of which their collections are built is a single web page. Thus, the resulting collections are more local than our clan graphs; in particular, some of them are more or less a single site.

Third, our system builds a rich profile of each site within a collection, including its links with other sites in the collection, the amount and type of content it contains (text, images, audio, video, and movie files), as well as indexing sites and pages by domain-specific and user-supplied phrases. Users can exploit these profiles to determine the quality and function of sites within a collection (and of specific pages within a site). Other work has gathered some of this information for pages. For example, WebQuery[7] orders pages by their degree, while Kleinberg [17] takes link analysis much further to categorize and disambiguate pages. WebCutter[19] organizes pages based on text similarity metrics. However, the profiles our system builds are more comprehensive and are provided both at the site level and for pages within sites.

Finally, we have created a novel visualization, the *auditorium view*. This is an exploratory interface that lets users quickly identify sites by a graphical thumbnail image, pick out sites that are most central to the topic (based on interconnectivity with other sites), arrange sites by any of the properties in the site profiles, and to explore the profile information for sites and their constituent pages. The goal is to support users in determining site function and quality.

CLAN GRAPHS: CONCEPTS AND ALGORITHMS

The primary information structure we use is the clan graph. A clan graph is a directed graph, where nodes represent content objects (such as documents) and edges represent a citation of or reference to the contents of the target node by the source node. Before we can describe how we construct and visualize clan graphs, we define our terms precisely.

Terminology

Universal Graph — the graph of all inter-document (e.g., inter-site) links in the information structure.

Topic Graph — A subgraph of the universal graph that contains documents on the same or similar topics. This is an ideal construct that can only be approximated, e.g., through analysis of structure or similarity of content.

Local Graph — For a specified set of seed documents, this is the subgraph of the universal graph whose nodes are the seeds or are "closely connected" to the seeds.

Observed Graph — It is practically impossible to construct the entire local graph because:

- the web is huge: trying to fetch all the pages on a site and to follow all the links off a site takes a long time;
- the web is unreliable: some sites always are down.
- the web is constantly changing, so the universal and local graphs are moving targets.

Thus, the observed graph is the subgraph of the local graph that we observe when we attempt to construct the graph.

Local clan graph: a formal definition

As we just stated, our goal is to compute (approximately) the local graph for a set of seed pages[2]. However, we still must define precisely what it means to be "closely connected" to the seeds. This is where our notion of the *NK* local *clan* graph comes in. After experimenting with several definitions, we converged on a simple, appealing definition building on concepts from social network analysis [15,27], co-citation analysis [11], and social filtering [9]:

- the NK local clan graph for a seed set S is

  {(vertices v, edges e) | v is in an N-clan with at least K members of S, e is an edge between two vertices v_1 and v_2}.

An N-clan[27] is a graph where (1) every node is connected to every other node by a path of length N or less, and (2) all of the connecting paths only go through nodes in the clan. We are interested primarily in

[2] The input to our system – the seeds – consists of pages; the system maintains both sites and pages as it operates; system output is in terms of sites (which of course have internal structure, including pages). The roles of pages and sites in our system are clarified in subsequent discussion.

2-clans, that is, the 2K local clan graph. The boxed subgraph of the graph in Figure 2 shows an example of a 2-clan contained within a larger graph.

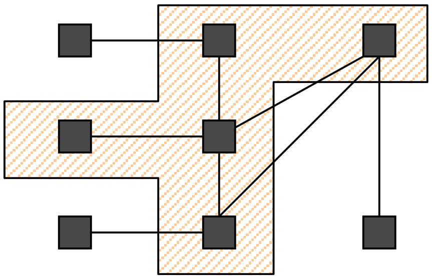


Figure 2: An example 2-clan

The clan graph is a key construct for us; we believe it productively formalizes notions like "collection" and "locality". We now attempt to justify this belief.

Why 2-clans? Figure 3 graphically depicts three types of inter-document relationships. In each case, an interesting relationship between two of the documents can be inferred based on a known relationship between the other two. Co-citation analysis (3a) says that documents B and C are related if A cites them both. Social filtering (3b) says that if documents B and C both refer to A, then B and C may be linked to similar sorts of items in general, and thus deal with similar topics. Figure 3c shows a limited (2-edge) transitivity; we could interpret this as "if C is on a particular topic, and cites A, then A is likely to be on topic; and if A cites B, then B is likely (though somewhat less so) to be on the same topic."

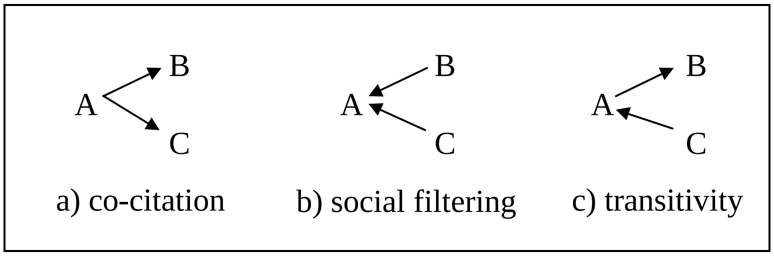


Figure 3: Motivating the use of 2-Clans

These three relationships are the minimal 2-clans. They illustrate why 2-clans are an appropriate formalization of topically related subgraphs within a larger graph. First, any nodes in a 2-clan have an intuitive and appealing interpretation in terms of document relatedness. Second, the existence of a relationship between nodes A and B and between A and C gives us some confidence that B and C also are related.

Notice that 2-clans are defined over undirected graphs: in other words, we take A and B as connected whether A links to B or vice versa. Again, we think this is appropriate. We have observed many sites that are topically central but that have only in-links (content sites) or out-links (index sites). A measure that required bi-directional paths between nodes would underrate some important sites. Therefore we first establish connectivity; subsequent analysis takes directionality into account in order to identify important structures like sinks and sources.

Why K seeds? By requiring that sites be related to a certain number of seeds, we ensure that we find not just dense graphs, but graphs in which a certain number of the seeds participate. Since we assume that the seeds (at least most of them) deal with a specific topic, this is a way to stay on topic. This is how we operationalize the “triangulation” effect we mentioned earlier. And making K larger is a simple way to get smaller, more tightly connected graphs. This usually means that the sites in the graph are more likely to be both on-topic and central to the topic. Of course, the optimal value of K depends in part on the number of seeds. For example, one almost certainly would want to pick a larger value of K for a seed set of size 20 than for one of size 5.

Notice that the NK clan graph definition does not require that all sites be in an N-clan with the same K seeds. For example, if we start with 10 seeds and K = 5, one site might be in an N-clan with seeds 1, 2, 3, 4, and 5, while another site is in an N-Clan with seeds, 2, 4, 7, 8, 9. If we start with 10 seeds, there are 10! /(5! (10-5)!) =252 subsets of size 5. Thus, there are many different N-clans that a non-seed site can participate in to make it into the NK clan graph.

Constructing the observed clan graph

Now that we have defined the local graph formally, we need an algorithm that can approximate it, that is that computes the observed graph. We designed a heuristic algorithm for this purpose. The algorithm is not guaranteed to produce the complete NK local clan graph; however, what it does produce always is a subgraph of the NK clan graph. In the next subsections, we discuss the role of the seeds, sketch the algorithm and describe the scoring function for site selection, and discuss how sites are defined.

Input: the seed

The observed graph we obtain depends on the properties of the seed pages we start with. Our experience is that users are able to choose good seeds. Good seeds have three properties. First, the seed set must cohere: if the seeds have few outgoing links or link to few of the same sites, the observed clan graph will be small or even empty (i.e., there is no N-clan that contains at least K seeds). This implies that the seeds do not participate in a significant dense subgraph within the universal graph. Second, the seeds must cover the topic: a poorly chosen seed set may lead to an observed graph that is a small subset of the topic subgraph. This can be the case if there are too few seeds, or the seeds are not well distributed across components in the topic graph. Finally, the seeds must be accurate: if some of the seeds are off-topic, then the clan graph may contain off-topic sites. However, if most of the seeds are on-topic, this is not a problem in practice. The parameter K plays an important role here: because any site added to the graph must be in a 2-clan with at least K seeds, as long as fewer than K off-topic sites are themselves related, sites they link to will not make it above this threshold.

The algorithm

We needed a type of web crawler, which fetches html pages, follows (some of the) links found on the pages and induces sites from pages. Pages that are linked-to (by already fetched-and-analyzed pages) are stored on a queue and become candidates for expansion (fetching and analysis). The major decisions the algorithm must make are to select the pages from the queue to expand and to aggregate pages into sites. Here is a sketch of the algorithm:

```
queue ← seed pages
while there is a queue element with a score above threshold do
    get the highest scored page from the queue
    construct a site containing this page, adding other pages from the queue as appropriate
    expand this site
    add the expanded site to the observed graph
```

```
        merge new pages and links from the expanded site into the queue
        re-organize and re-score the pages on the queue
    end
```

Scoring pages on the queue

We need a scoring metric that estimates the likelihood that a page on the queue is in the local graph with the seeds, i.e., that it is in a 2-clan with at least K seeds. The metric must be efficient to compute, since it must be applied to each page on the queue, and the queue typically contains thousands of pages.

We currently use the following scoring metric:

- score of page P = the number of seeds that are linked to P by paths of length 2 or less.

This metric is cheap to compute. It also is a reasonable heuristic, since 2-clans are composed of 1 and 2-paths. Thus, if a page has a score of (say) 5, then it already is known to be in a 2-clan with 5 seeds. We are in the process of experimenting with and evaluating this heuristic and considering other heuristics at different points along the accuracy/efficiency continuum.

Sites

A key claim we have been making is that the web site is a more appropriate basic unit than the web page. In this section, we describe precisely what we mean by a site and how we group pages into sites.

A site (multimedia document) 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 thing as a domain: for example, thousands of sites are hosted on www.geocities.com. And what counts as a site may be context dependent. For example, if one is taking a survey of research labs, www.media.mit.edu might well be considered a site, while if one is investigating social filtering projects, individual researchers' sites hosted on www.media.mit.edu are probably the proper units.

The last observation suggested a way to operationalize the definition of a site that suits our needs. When building a clan graph, the relevant context is the set of URLs that have been linked to by the expanded sites. The intuition is that if sites in the clan graph 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[3]. 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.

To illustrate how this works, let us return briefly to the Ani DiFranco example. The following URLs all were encountered, either as seeds or because they were linked to by other pages:

- http://www.flemtam.com
- http://www.flemtam.com/ad.html
- http://www.flemtam.com/ad-itin.html

As soon as one of these URLs was selected for expansion, the other two were determined to belong to the same site, and the first URL (flemtam.com) was set as the site root page.

[3] Notice that our notions of site and clan graph are interdependent: a site is defined in terms of links from within the graph, and the graph is constructed by following links from sites.

In addition to this rule, our system uses a number of heuristics, including general ones such as the convention that the "~" character indicates different user sites, so http://www.research.att.com/~terveen and http://www.research.att.com/~willhill are usually distinct sites, as well as specific heuristics about the structure of large hosting sites like geocities or tripod. As mentioned above, the site-aggregation rule and heuristics are applied when a page is selected from the queue for expansion.

The rule and heuristics can fail — two URLs that belong to the same site will not be merged if no common ancestor in the directory structure (the "real" root page) has been linked to, and two URLs from distinct sites can be merged if there is a link to a more general host (that is *too* general to be a site in this context). We are refining this rule with site-splitting heuristics based on the idea that when some "internal" pages are linked to significantly more often than is the (supposed) root page, then the heavily linked-to internal pages may be separate sites. And we are considering site-merging heuristics based on the idea that if (supposedly) distinct sites point to many of the same pages in the same domain, they may be part of the same site.

We also must decide whether a link from a page is within the site or to another site. We classify links based on their relationship to the root page of the site. If a link is contained within the directory that contains the root page, then we classify it as internal; otherwise, we classify it as a link to an external site. Internal links are added to a site-internal queue of candidate pages to be fetched.

Finally, we must specify how many pages to fetch from a site, i.e., what it means to expand the site. The primary reason for fetching pages is to find links to other sites, which are the building blocks of the clan graph. For this purpose, finding a site's index page presumably would yield most or all such links, so we could stop expanding the site then. Indeed, we try to find index pages first by sorting pages on the site-internal queue by name, preferring pages whose names contain words like "links", "pages", "sites", "web", and "internet".

There is another important reason to fetch pages, namely to build a profile that can be used to evaluate a site. The more pages we fetch, the more accurate a site profile we can create. Therefore, to serve both goals, we introduce a parameter P (default = 25) that controls how many pages to fetch from a site. We discuss the makeup of site profiles next.

COMPUTING SITE PROFILES TO INFORM USER EVALUATION

The clan graph consists of a small to medium sized collection of sites, typically 30 – 200. However, as we argued before, this is still too many for a user to consider. Therefore, we want to generate additional information to help users evaluate the quality and function of each site. This is the role of the site profile.

As we fetch pages for each site, we gather various types of information about each page. The page data is aggregated to construct a site profile. The site profile includes the following information:

- title (of the site's root page);
- a thumbnail image (of the site's root page);
- links to and from other sites;
- media content of pages and sites, including images, audio files, and movie files;
- internal pages of the site, along with a count of links to each page from other sites within the graph; this is a way to identify internal pages that the topic community has found worth endorsing; we can provide access to these internal pages in the interface, thus offering "shortcuts" to interesting content.

- a count of the occurrences of domain-specific indexing phrases in the title and body of each page, as well as in the anchor text of incoming links to the page; for example, for a musical performer, the phrases might include “lyrics”, “chords”, “tour dates” , etc.; this is the primary way of helping users to evaluate the function of sites and pages.

After the graph is complete, we compute the number of 2-clans each site participates in. This is another metric that can be used to judge the structural centrality of a site for its topic.

EVALUATING THE NK CLAN GRAPH

We must verify experimentally that the NK local clan graph is a useful construct. Beginning from a set of seed pages, the construction algorithm should obtain most relevant and few irrelevant sites. The results should not be too dependent on precisely which pages are selected as seeds. The site profiles should be shown to be useful in evaluating site quality and function.

Initial informal inspections of dozens of graphs made us optimistic that these conditions were satisfied. We followed up with a number of in-depth analyses of the results for several topics and several user studies. We first describe several of the analyses we did, then summarize the results of one of the user studies.

Ani DiFranco example

We motivated the clan graph idea by identifying some shortcoming in PHOAKS data, using the Ani DiFranco newsgroup as a example. Let us now return to the Ani DiFranco data from PHOAKS and investigate what happens when we take these items as seeds for the clan graph algorithm.

We ran the algorithm with K=5 and set a limit of 100 sites for the crawler. After 100 sites were completed, we computed all 2-clans in the graph, then deleted all sites that were not in at least one 2-clan with 5 seeds (of course, this can lead to seeds being deleted, too, if they are isolated from the rest of the seeds). The process iterates until no more sites are deleted; in this case, 72 sites remained in the final (2-5) clan graph.

The first observation we made was that peripheral items among the seeds were deleted. Liszt, the mailing list directory, The UK Gay Times online magazine, and the eBay online auction site all vanished.

Second, we are interested in the relevance and quality of the sites that remained in the graph. To get an indication of this, we ordered the sites by the number of links from seed items, using the number of links from non-seed sites to break ties. (Note that this is just one way to order the sites; in real exploration tasks, users almost always experiment with multiple orderings, using different attributes from the site profile.) Table 2 shows the top 10 sites according to this ordering.

Title	**In Links** **(seeds, non-seeds)**	**Phoaks Rank**
Ani DiFranco Tab	8, 2	13
Fleming Tamulevich and Associates Online	6, 14	15
Ani DiFranco		
Ani DiFranco Itinerary		1
Dan Bern		
Dan Bern Tour Itinerary		
OoooOoOOooOoOo - a simple man	6, 6	

Another Ani DiFranco Page	6, 2	4
The Ultimate Band List	5, 6	29
UBL: Ani DiFranco Links, Biographies…		
UBL: Andy Stochansky Links, Biographies…		
Quick Stop: ani difranco	5, 4	
i don’t know who you were expecting	5, 2	
The Ani DiFranco Web Ring page	4, 7	11
Sarah’s Ani DiFranco Page		
Dirty Linen Magazine	4, 4	
Just Another Disgusting Shrine Dedicated To The Great One	4, 4	
Caught On Tour		

Table 2: Results of applying clan graph algorithm to PHOAKS Ani DiFranco URLs

5 of the top 10 sites were added by the algorithm. 4 of these sites are wholly dedicated to Ani DiFranco while the 5th, “Dirty Linen Magazine”, is the online presence of a folk music magazine that has written about Ani DiFranco and related artists. A number of the sites contain multiple pages that were linked to in this context. The Ultimate Band List is an interesting example. It appeared in the seed set, yet it is a general rock music site. However, the other sites within the clan graph linked primarily not to the general site, but to the Ani DiFranco area. In addition, there were a few links to a page for Ani DiFranco’s drummer. Something similar happened with the Fleming Tamulevich site (recall this is a site for a booking agency that handles folk music artists). In addition, to the general page and the tour dates page for Ani DiFranco, there also were links to the general and tour dates page for Dan Bern, another artist whom Ani DiFranco has toured with and produced.

Thus, the clan graph algorithm succeeded in removing noise from the initial seed set, discovering additional relevant sites, and in grouping pages into sites.

Grateful Dead example

We wanted to examine the sensitivity of the clan graph construction algorithm to the initial seed set. To that end, we did an experiment on the topic of the rock group The Grateful Dead. We used 63 URLs obtained from Yahoo as a starting point for the experiment. We randomly divided these URLs into sets of size 5, 10, and 20. We used these as seed sets for our clan graph construction algorithm, and also experimented with different values of K.

Analysis of the results confirmed some of our intuitions. First, larger seed sets (size 10 or 20) tend to result in graphs that better cover a topic than do smaller seed sets (size 5). Using a search engine or a directory like Yahoo makes it easy to obtain a large set of seeds for many topics.

Second, increasing the parameter K results in smaller, more tightly focused graphs, while decreasing K leads to larger, but perhaps not as accurate graphs. We have not quantified this effect precisely yet; this remains a topic for future research.

Third, sites with large numbers of in-links almost always are discovered by the clan construction algorithm regardless of the sites in the seed set. Therefore, the algorithm does not appear overly sensitive to the choice of seeds.

Finally, when we ranked sites within a graph by in-degree, the top ranked sites (i.e., those most cited by their "peers") always were on-topic. We did find that "the topic" may be somewhat broader than we initially had supposed. For example, many Grateful Dead sites link to The Electronic Frontier Foundation and various tape-trading and tape-tracking sites. Although these sites are not about the Grateful Dead per se, clearly they are part of what the online Grateful Dead community considers important and relevant.[4] This community is defined by but not limited to interest in The Grateful Dead.

User Studies

We carried out a user study (see [31] for details) that shed some light on the strengths and limits of the clan graph algorithm and (to some extent) of any approach based on link analysis. We invited members of our laboratory to suggest topics they were interested in. We asked that topics be represented by a Yahoo category; for example, astronomy is represented by the Yahoo category Science:Astronomy. 30 people responded to our request: 26 provided a Yahoo category, while 4 supplied either a topic-specific directory page or a set of sites that they had gathered to represent their interest. We also selected another 15 Yahoo categories at random.

For each topic, we constructed the 2-K local clan graph. We analyzed the results in two ways. First, we computed various structural properties of the graphs. Second, we presented 40 randomly selected sites from each graph to the person who had suggested the topic (20 of the sites were seeds, i.e., were from the original Yahoo category, and 20 were discovered by the algorithm). The person then rated the relevance and quality of each site on a scale of 1 (worst) to 5 (best).

Some of the important results from the study were:

- Many topics were tightly interconnected by links; however, topics dominated by commercial sites were not. Commercial sites tend not to link to other sites on the same topic (even allowing for links through one other, intermediate site). This suggests that any link-based approach will not work very well in topics dominated by commercial sites. However, other topics are characterized by extensive inter-site linking – topics dominated by hobbyists or fans are particularly striking in this regard – and our techniques thus work very well for these topics.
- Most sites within a single topic are linked (directly or indirectly) with each other (57% of sites in all topics, 68% of sites in non-commercial topics); the vast majority of these sites are reachable by following 1 or 2 links from any small, randomly selected subset of the sites.

  We conducted a simple experiment to determine precisely how many sites on a particular topic lie within 2 links of a small set of seeds. We applied the following test to each graph with at least 25 non-isolated sites (this gave us 20 topics). We selected 20 different random subsets of size 1, 2, … , 10 of the non-isolated sites to use as a "test seed". (We continued only if at least 20 non-isolated sites remained once we made the selection.) For each set of test seeds, we counted the number of other non-isolated sites that could be reached via a path of length 1 or 2. Table 3 shows the results.

[4] This is a subtle point. We are not claiming that *anything* that the seeds link to is, by definition, part of the topic. There are obvious counter-examples: many sites point to the Netscape site (as a place to download software), but this doesn't mean that Netscape is part of all these topics. What we have seen, however, is that sometimes the linked-to sites clearly are part of the concerns of a topically-oriented community. By finding such sites, our algorithms make explicit the broader, previously submerged interests of a community.

For a seed set of size 2, 71% of the other sites were reached, for a seed set of size 5, 85% were reached, for a seed set of size 10, 92% were reached, etc.[5]

Number of seeds	*Percentage of non-isolated sites reachable*
1	54
2	71
3	77
4	82
5	85
6	87
7	88
8	90
9	91
10	92

Table 3: Reachability of sites

Let us consider the case of 5 test seeds to make the details of this result clear. The topic graphs in this experiment contained an average of 126 sites, of which 40 sites (32%) were isolated. Once we selected 5 sites at random to serve as a test seed, this meant that 81 sites (126 – 40 – 5) were potentially reachable. 85% (69/81) of these sites were reached by following 1 or 2 links from one of the test seeds. What this suggests is that the vast majority of a fairly large set of sites known to be on a given topic (known because they were all included in the same Yahoo category) can be reached by following just 1 or 2 links from a small seed set.

- Connectivity between sites correlated with human quality and relevance judgements. The subjects rated connected sites as significantly more relevant and of higher quality.
- Many relevant, high-quality sites were discovered. This is significant because our seed sets came from Yahoo, a state of the art, human-maintained directory. Slightly more than a third of the discovered sites were judged to be of high relevance and quality. While this number may not appear all that impressive, recall that the sites were presented at random, i.e., the users did not receive the benefit of the site profile data. In a later experiment[1], this data was used and did indeed prove helpful, with the number of in-links and total number of pages proving especially useful in letting subjects identify high quality sites.

VISUALIZING CLAN GRAPHS

The structure of clan graphs that we have observed in the web is complicated and not easy to visualize or understand. For example, Figure 4 is a direct node/edge representation of the Ani DiFranco clan graph

[5] We considered only seeds in this analysis. We did this for two reasons. First, we wanted to be positive that all the sites really were on-topic. Second, since the analysis dealt with links between sites, and our algorithm adds sites to the graph only if they are linked with seed sites, including discovered sites would skew the results.

we have been discussing. The drawing was produced by a sophisticated graph layout tool, *dot*, which minimizes edge crossings, yet the drawing still is complicated. The clutter of edge crossings, edge angles and local node constellations divert visual attention to non-significant graphic elements. A viewer can identify some nodes of high and low degree, but the layout reveals no overall pattern. It is very difficult to visually discern central and peripheral sites or to grasp the entire set of relationships any individual site participates in.

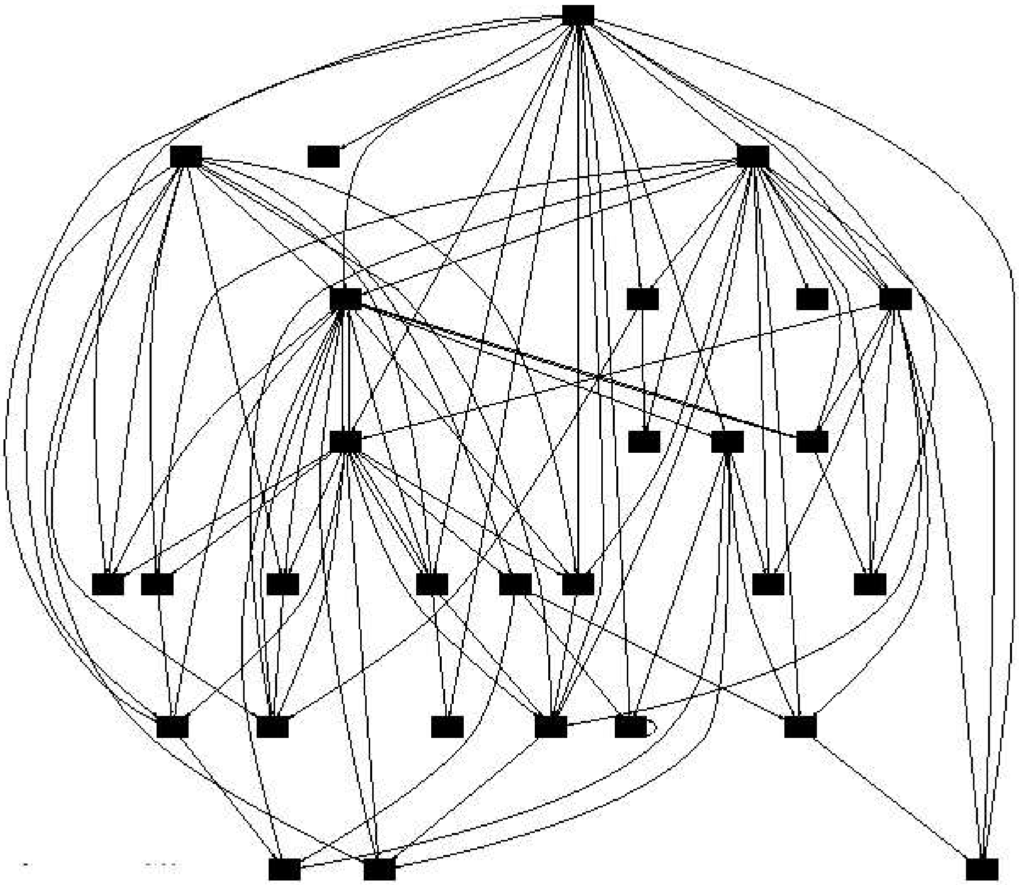


Figure 4: node/link view of the Ani DiFranco graph

For the purpose of revealing node degree, simply collapsing the graph structure into a list of nodes ordered by degree is a better interface: imagine the information in Table 2 presented in a PHOAKS-like format. While such displays are easy to produce and make it easy for users to compare node degree and check quantities, they still hide many important properties of sites and the graph. The list view is linear, so it easily communicates only one dimension. It is textual, so it cannot exploit graphical display properties, either images from the sites or the use of color, position, shape, etc. to communicate site properties. It is static, so there is no dynamic focusing, no hiding and revealing of structure, no sorting by one metric, then another. Only a single metric, the number of in-links, is presented, so users still must evaluate a site based on relatively little information.

We wanted users to see the results of our clan graph analysis in terms of the graph itself. These results include site centrality/peripherality, in-link to out-link ratio, patterns of inter-site links, and how sites rank in terms of properties such as size, number of images, audio and download files.

The Auditorium view: how it satisfies our design goals

To meet these goals, we have iterated cycles of design and usability testing, arriving at a design that we call the *auditorium seating visualization.* The metaphor is to the arrangement of seating in an auditorium: row upon row curved around a center stage. Figure 4 and color plate 1 show the auditorium seating visualization of the Ani DiFranco clan graph. Thumbnails of site “front door” pages serve as iconic representations of sites. The auditorium seating visualization is dynamic. By moving the mouse over a site thumbnail, users switch from a general view of the graph to a view focused around the indicated site. Figure 4 shows the visualization in the site-focused mode. The thumbnail of the focused site is enlarged, and green “in” arrows and red “out” arrows appear on sites that the focused site is linked with. Other unlinked sites are blanked, but their drop shadows are left to note their positions. As the result of user experience with many design versions, we came to use a large number of techniques to communicate information necessary to satisfy user needs. We now discuss these in detail. Please refer to figure 4 or color plate 1 to identify the graphic elements discussed.

- *Concentric semi-circles* of sites group sites into equivalence classes from most to least important on some user-settable property. By default, sites are assigned to rows based on the number of 2-clans they occur in. Thus, the structural centrality of the site is directly represented as a display centrality.

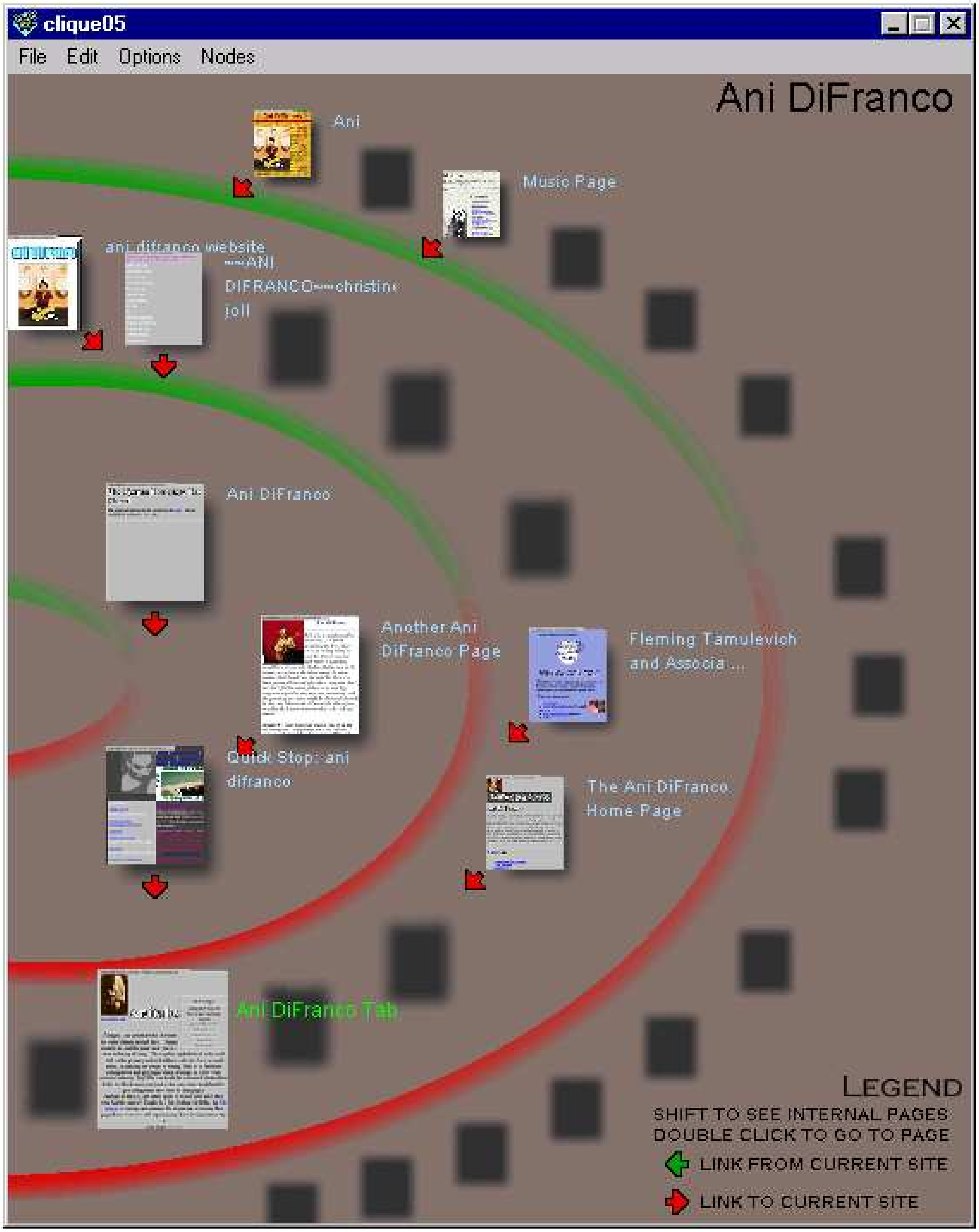


Figure 4: Auditorium view of clan graph

- *Dynamic ordering within semi-circles*. Our original design used concentric circles instead of semi-circles. However, user feedback showed the desirability of ordering sites within each row, and while circles “wrap around”, the properties important for ordering (such as number of in or out links and amount of content) do not. Semi-circles, on the other hand, with their definite end points, were suitable for our purpose.

An important distinction users made was between index sites and content sites. Allowing dynamic ordering of sites within a row by properties like number and proportion of in and out links, and amount of content (audio files, images, or all types of content) makes these distinctions apparent. By default, we order sites within rows by the proportion of in-links to out-links; sites at the bottom have the highest proportion of in-links, sites at the top have the highest proportion of out-links. Thus, linking sites tend to appear at the top of the display, while content sites tend to appear at the bottom.

- *Graded colored bands* aid in interpreting the values of the within-row ordering property of sites. Bands are graded from bright red to bright green, with the color break occurring at the median value of the ordering property. For example, if sites are ordered by the proportion of in links to out links, the break point is a visual cutoff between sites that serve more as indices and sites that serve more as content repositories.
- *Hiding graph spaghetti* — We wanted to reveal the fine structure of inter-site links without producing visual spaghetti as in figure 2. Users typically focused either on all the links from or to a single site or traced the edge between two sites. We designed to support those two visual tasks while removing as many distracting visual elements as possible. We did this with “one-site at a time” dynamic presentation of graph structure. Users move the mouse cursor over a site to focus on it, and only links from or to the focused site are shown. To further reduce clutter, we do not draw complete links between sites, since they draw too much user attention to uninformative crossings and edge angles. Instead, we represent links with small in and out arrows.
- *Linked views* The auditorium view is linked to a web browser; clicking on a thumbnail drives the browser to that web site.
- *Progressive revelation* of greater detail. While a site is in focus, holding down the shift key reveals any internal pages of that site that are linked to by other sites. These are pages that the author of the linking site found worthy of special attention. The link text often is more informative in these cases. In addition, by clicking on a site, the user can access a display of the site profile data to find the amount and type of information the site contains and access significant internal pages.
- *Thumbnail representations* reveal quite a bit of information about sites. Overall design and color scheme can be seen. Ratio of text to graphics on the front door page tells users something about what to expect from a site. Saturated color, positioning and shape of banner ads reveal their presence in thumbnails. If a user has browsed a site previously, a thumbnail usually is sufficient to identify the site.

Early user testing highlighted for us the necessity of relevance feedback, leading to construction of a new observed clan graph. Users can judge sites as on-topic (good) or off-topic (bad). On-topic sites are added to the original seed set, and off-topic sites are added to a stop list. Thus, users can nudge the graph into a somewhat different area, moving it closer to the ideal topic they have in mind.

STATUS AND CURRENT WORK

We have developed a Java applet that gives users interactive access to the clan graph webcrawler. The applet lets users specify the seeds and set various control parameters. As the crawler runs, it presents results immediately, showing thumbnail images, with site profile data and access to the sites themselves just a click away. This lets users begin exploring the topic immediately.

We also have carried out design reviews and detailed user studies [1,31] investigating the utility of our interface. These reviews and studies showed the efficacy of site profile data in helping users identify high-quality sites. They also helped us identify key interface features, which have led to a new interface design. The new design retains many of the same elements, but is somewhat simpler, and makes the site profile data more visible. It also emphasizes techniques that let users organize collections to reflect their own understanding of the area, for example, by grouping and categorizing items.

We have begun experimental comparison of our algorithm to other link analysis algorithms, in particular those based on Kleinberg's algorithm [4,8,17]. Preliminary results do not show that the authority/hub computation tells us much more than simple counts of in and out degree, but we will carry out a more detailed analysis.

Finally, we are making plans to deploy our system. As we have mentioned, our techniques work very well in topic areas dominated by hobbyists or fans. Thus, we are developing a web site containing collections for hundreds of such topics (initially in the area of television/media and rock music fandom), with our software available to download and view these collections. By distributing our software and maintaining a server, we will be able to investigate the social nature and social roles of communities that organize their interests around online information resources.

CONCLUSIONS

The goal of the work reported here is to help people find and manage collections of documents related to topics they care about. We offer a novel information structure, the clan graph, to formalize the notion of a topically related collection of interlinked documents. We present an algorithm to construct a clan graph from a set of seed documents. The algorithm also tackles the hard problem "what is an online document?": it aggregates individual web pages (URLs) into sites (multimedia documents) based on the context of links from other documents. Finally, we introduce and illustrate the auditorium visualization. It gives a graphical overview of the most important several dozen sites for a topic, lets users explore structural relationships between sites and the internal structure of individual sites, and allows dynamic sorting to aid users in understanding the structural role a site plays within the community of related sites. We have moved from informal to formal evaluations of both our algorithms and interface and are moving to deploy our software on a broad scale.

ACKNOWLEDGMENTS

We thank Josh Creter and Peter Ju for their system implementation work. We also thank Harley Manning, Steve Whittaker, Lynn Cherny, and Julia Hirschberg for many useful design discussions. Finally, we thank the TOCHI reviewers for many helpful comments and suggestions.

REFERENCES

1. Amento, B., Hill, W., Terveen, L., Hix, D., and Ju, P. An Empirical Evaluation of User Interfaces for Topic Management of Web Sites, to appear in *Proceedings of CHI'99* (Pittsburgh, PA, May 1997), ACM Press.

2. 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.

3. 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.

4. 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.

5. Botafogo, R.A., Rivlin, E., and Shneiderman, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems 10, 2*, 142-180.

6. Card, S.K., Robertson, G.C., and York, W. The WebBook and the Web Forager: An Information Workspace for the World-Wide Web, in *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press, 111-117.

7. Carrière, J., and Kazman R. WebQuery: Searching and Visualizing the Web through Connectivity, in *Proceedings of WWW6* (Santa Clara CA, April 1997).

8. Chakrabarti, S., Dom, B.E., Gibson, D., Kumar, S.R., Raghavan, P., Rajagopalan, S., and Tomkins, A. Experiments in Topic Distillation. ACM SIGIR workshop on Hypertext Information Retrieval on the Web (1998), Melbourne, Australia.

9. *Communications of the ACM*, Special issue on Recommender Systems, 40, 3 (March 1997). Resnick, P., and Varian, H.R., guest editors.

10. Furnas, G.W. Effective View Navigation, in *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 367-374.

11. Garfield, E. *Citation Indexing*. ISI Press, Philadelphia, PA, 1979.

12. Hightower, R.R., Ring, L.T., Helfman, J.I. Bederson, B.B., and Hollan, J.D. Graphical Multiscale Web Histories: A Study of PadPrints, in Proceedings of Hypertext'98 (Pittsburgh PA, June 1998), ACM Press.

13. Hill, W.C., Stead, L., Rosenstein, M. and Furnas, G. Recommending and Evaluating Choices in a Virtual Community of Use, in *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press, 194-201.

14. Hill, W. C. and Terveen, L. G. Using Frequency-of-Mention in Public Conversations for Social Filtering. in *Proceedings of CSCW'96* (Boston MA, November 1996), ACM Press, 106-112.

15. Jackson, M.H. Assessing the Structure of Communication on the World Wide Web. *Journal of Computer-Mediated Communication , 3, 1,* June 1997

16. Kandogan, E., and Shneiderman, B. Elastic Windows: A Hierarchical Multi-Window World-Wide Web Browser, in *Proceedings of UIST'97* (Banff, Alberta, Canada, October 1997). ACM Press.

17. Kleinberg, J.M. Authoritative Sources in a Hyperlinked Environment, in *Proceedings of 1998 ACM-SIAM Symposium on Discrete Algorithms* (San Francisco CA, January 1998), ACM Press.

18. Lamping, J., Rao, R., and Pirolli, P. A Focus + Context Technique Based on Hyperbolic Geometry for Visualizing Large Hierarchies, in *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press, 401-408.

19. Maarek Y.S., Jacovi, M., Shtalhaim, M., Ur, S., Zernik, D., and Ben Shaul, I.Z. WebCutter: A System for Dynamic and Tailorable Site Mapping, in *Proceedings of WWW6* (Santa Clara CA, April 1997).

20. Mackinlay, J.D., Rao, R., and Card, S.K. An Organic User Interface for Searching Citation Links, in *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press, 67-73.

21. Marchiori, M. The Quest for Correct Information on the Web: Hyper Search Engines, in *Proceedings of WWW6* (Santa Clara CA, April 1997).

22. Mukherjea, S., and Foley, J. D. Visualizing the World-Wide Web with the navigational view finder. *Computer Networks and ISDN Systems 27, 1*, (1995), 1075-1087.

23. Mukherjea, S., Foley, J.D., and Hudson, S. Visualizing Complex Hypermedia Networks through Multiple Hierarchical Views, in *Proceedings of CHI'95* (Denver CO, May 1995), ACM Press, 331-337.

24. Pirolli, P., Pitkow, J., and Rao, R. Silk from a Sow's Ear: Extracting Usable Structures from the Web, in *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press, 118-125.

25. Pirolli, P., Schank, P., Hearst, M., and Diehl, Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection, in *Proceedings of CHI'96* (Vancouver BC, April 1996), ACM Press, 213-220.

26. Pitkow, J., and Pirolli, P. Life, Death, and Lawfulness on the Electronic Frontier, in *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 383-390.

27. Scott, J. *Social Network Analysis: A Handbook*. SAGE Publications, London, 1991.

28. Spertus, E. ParaSite: Mining Structural Information on the Web, in *Proceedings of the Sixth International World Wide Web Conference* (April 1997).

29. Terveen, L.G., Hill, W.C., Amento, B., McDonald, D., and Creter, J. Building Task-Specific Interfaces to High Volume Conversational Data, in *Proceedings of CHI'97* (Atlanta GA, March 1997), ACM Press, 226-233.

30. Terveen, L.G., and Hill, W.C. Finding and Visualizing Inter-site Clan Graphs, in *Proceedings of CHI'98* (Los Angeles CA, April 1998), ACM Press, 448-455.

31. Terveen, L.G., and Hill, W.C. Evaluating Emergent Collaboration on the Web, in *Proceedings of CSCW'98* (Seattle WA, November 1998), ACM Press.

32. *What is twURL?* http://www.twurl.com