

Evaluating Emergent Collaboration on the Web

Loren Terveen

AT&T Labs - Research
180 Park Avenue
Florham Park, NJ 07932
+1 973 360-8343
terveen@research.att.com

Will Hill

AT&T Labs - Research
180 Park Avenue
Florham Park, NJ 07932
+1 973 360-8342
willhill@research.att.com

ABSTRACT

Links between web sites can be seen as evidence of a type of *emergent collaboration* among web site authors. We report here on an empirical investigation into emergent collaboration. We developed a webcrawling algorithm and tested its performance on topics volunteered by 30 subjects. Our findings include:

- Some topics exhibit emergent collaboration, some do not. The presence of commercial sites reduces collaboration.
- When sites are linked with other sites, they tend to group into one large, tightly connected component.
- Connectivity can serve as the basis for collaborative filtering. Human experts rate connected sites as significantly more relevant and of higher quality.

Keywords

Social filtering, collaborative filtering, computer supported cooperative work, human computer interaction, information access, information retrieval

INTRODUCTION

The field of CSCW sees collaboration as involving people who *know* they are working together, e.g., to edit a document, to carry out a software design review, or to play a role-playing game. CSCW research then focuses on understanding the nature of collaboration and developing computer support to enhance collaboration.

Our focus is different. We have been inspired by the example of a path through the woods. A path results from the decisions of many individuals, united only by where they choose to walk, yet still reflects a rough notion of what the walkers find to be a good path. The path both reflects history of use and serves as a resource for future users. We look for analogous situations in the computational world, attempting to turn implicit "paths through the woods" of particular user communities into explicit, shared resources. We seek situations where groups of people already are producing computational records (such as documents, Usenet messages, or web sites and links) as part of their normal activity. We then identify potentially useful information implicit in these records and invent computational techniques to harvest the information and make it explicit. Finally, we create interfaces to distribute

the information both to the original group members and to a wider community.

The work reported here investigates links between web sites as a potential ground for emergent collaboration. Links from a single web site tell us what the site author thinks are topically relevant, high quality sites. However, we hypothesize that patterns of links among a group of web sites can tell us what a (usually implicit) community of web site authors, united only by their interest in a common topic, think are relevant, high quality sites. This paper reports on an empirical test of this hypothesis.

EMERGENT COLLABORATION: OUR PERSPECTIVE AND RELATED WORK

Our first line of work that took the emergent collaboration approach was history-enriched digital objects [4, 5, 13]. This work was based on the observation that objects in the real world accumulate wear over the history of their use, and that wear — such as the path through the woods or the dog-eared pages in a paperback book or the smudges on certain recipes in a cookbook — informs future usage. *Edit Wear* and *Read Wear* were terms used to describe computational analogues of these phenomena. Statistics such as time spent reading various parts of a document, counts of spreadsheet cell recalculations, and menu selections were captured and then used to modify the appearance of documents and other interface objects in accordance with prior use.

Recently other researchers have begun to explore the use of capturing and visualizing history. Seligman [11] takes the metaphor of paper documents literally, visualizing web pages as if they were paper documents, showing information about their access history as smudges, stains, fingerprints, etc. Wexelblat's Footprints system [14] captures the history of access to sets of web pages and creates visualizations based on this history to aid navigation. Alexa (<http://www.alexa.com>) keeps track of the links users follow from one web page to another and uses this information to recommend the most popular links.

Our second attempt at harvesting information from group records was PHOAKS (People Helping One Another Know Stuff)[6, 12]. PHOAKS is a collaborative filtering system, based on the premise that an effective way of finding 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. It then 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.

There is a significant amount of ongoing research aimed at extracting useful information from the structure of web links. Two projects related to ours are those of Kleinberg and Carrière & Kazman. Kleinberg [7] 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. Carrière & Kazman's WebQuery system [2] sorts pages into equivalence classes based on their total degree (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.

Pirolli, Pitkow, and colleagues at Xerox PARC have done a great deal of work that analyzes web structure and usage data, attempting to categorize and cluster web pages. Pitkow and Pirolli [10] describe clustering algorithms based on co-citation analysis [3]. 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 [9] 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.

Botafogo et al [1] present interesting algorithms for analyzing hypertexts. Although much of their focus was on the analysis and improvement of (navigation through) a single hypertext, some of their techniques undoubtedly could be used to analyze collections of hyper-linked documents.

In summary, our approach is to harvest information that emerges from patterns of individual activity, where the

individuals involved may not be communicating explicitly, but are bound by common artifacts and interests — documents they edit, topics they care about, web sites they maintain and visit, newsgroups they read and post to. Like others, our concern is with deriving useful information from links between web sites. The unique contribution of the work we report here is that it investigates empirically how individual decisions to link to other sites knit topic areas together and how the connectivity of sites within a topic correlates with human judgements about the relevance and quality of sites.

THE DATA

We invited members of our laboratory to supply us with 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. Yahoo categories offer a convenient means of obtaining a set of web pages for a given topic.

30 people responded to our request, 26 with a Yahoo category, and 4 with either a single links page or a set of web sites that they selected to represent their interest. (Since we simply needed sets of sites on specific topics for this experiment, we were happy to accept the 4 submissions that were not Yahoo categories.) In addition, for reasons that we discuss below, we selected another 15 Yahoo categories randomly, 5 from the Business and Economy hierarchy, 10 from elsewhere in the hierarchy. Table 1 shows how the topics are distributed across the Yahoo hierarchy (we were easily able to classify the 4 topics that were not represented as Yahoo categories). Some examples of the topics users selected include Palm Pilots, slide rules, Frank Herbert's "Dune" series, interactive fiction, astronomy, and computer electronics.

Table 1: Topic areas in the experiment

Category	User-supplied	Selected randomly
Arts and Entertainment (A)	11	3
Science (S)	6	4
Recreation (R)	6	0
Business and Economy (B)	5	5
Computers and Internet (C)	1	
Education (E)	1	
Health (H)		3
Totals	30	15

We applied our web-crawling algorithm (described in [13]) to the set of sites for each topic to construct the inter-site graph. The nodes of this graph are the sites, and there is an edge from one site's node to another if and only if (some page of) the first site links to (some page of) the second.

Our algorithm heuristically aggregates pages into sites ([13] gives details). For example, if it encounters a link to the URL <http://a/b/c/page1.html>, and <http://a/b/c/index.html> is a known site, it records this URL as part of the site. Further, if the link was encountered while analyzing the site <http://x/y/z/>, a link is recorded from the site <http://x/y/z/> to the site <http://a/b/c/index.html>. We ran our webcrawler to fetch and analyze 25 pages for each site. The primary reason to fetch a significant number of pages per site is to maximize the chance of obtaining the links page of each site (if there is one; if not, we still have a better chance of getting links that are distributed across the site's pages) so that we construct the inter-site graph accurately. In addition, the algorithm constructs a profile of site content, collecting and counting links to images, audio files, html and text files, and the more pages fetched, the more accurate the profile. The content profile aids both user comprehension and automatic classification of sites.

As we analyzed the data, we noticed important differences between Business and Economy (B&E) topics and the other topics (see below for details). However, since users had supplied only 5 B&E topics, we wanted to obtain a larger sample; therefore, we randomly selected 5 additional B&E topics, as well as 10 other topics to further increase our sample size.¹ Thus, we ended up with 45 topics, which contained an average of 79 sites.

After we built the graph for a topic, we carried out two types of analysis. First, we analyzed the connectivity of the graphs; the goal was to determine whether topics on the web are marked by emergent collaboration (as indicated by significant linking between sites). Second, we investigated whether the connectivity of a site could serve as the grounds for collaborative filtering, i.e., could it allow us to identify high-quality sites. To answer this question, we selected 20 sites at random from each topic and presented the sites to the person who suggested the topic for rating. Sites were presented in an HTML form; subjects followed hot links to sites to investigate their content, then used menus to rank site relevance and quality on a scale of 1 to 5. 23 of the people rated at least some of the sites for their topic.

We discuss each of these analyses in turn.

CONNECTIVITY OF TOPIC GRAPHS: COLLABORATION OR COMPETITION?

We first checked each graph to see how many of the sites were isolated, that is, neither linking to nor linked to by any other sites on their common topic. We were surprised to find a large number of isolated sites: on average, 43% of the sites on a topic were isolated. While the web raises the possibility of creating a multi-author, inter-related, cross-referenced knowledge base for a topic, a large proportion of sites do not follow this ideal. We did two things to follow up on the finding of a large number of isolated sites. First, we analyzed the structure of the non-isolated sites; next, we tried to understand why so many sites were isolated. We will discuss each of these analyses in turn; however, first let us consider two topic graphs.

Figure 1 shows a fairly densely connected graph for an Arts & Entertainment category, "Beat Poets". Of the 32 sites in the topic, only 8 are isolated. There are 44 links, giving a *density* of 0.044 (density is defined below). Even the coarse presentation of Figure 1 reveals some patterns, e.g., a few sites with many outgoing links and one site with many incoming links.

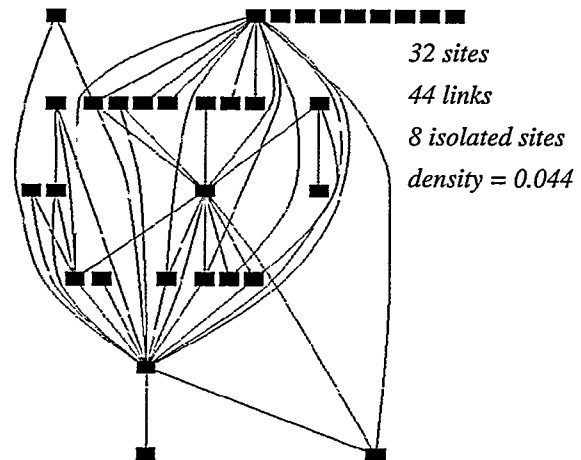


Figure 1: Densely connected graph ("Beat Poets")

Figure 2 shows a more sparsely connected graph for the topic of travel to Italy. The most striking feature is that 42 of the 58 sites are isolated. The density is 0.0048 (an order of magnitude less than the previous graph). We discuss this topic more later.

¹ We should note that there were no significant differences between the topics supplied by our subjects and those we selected at random on any of the measures we report here.

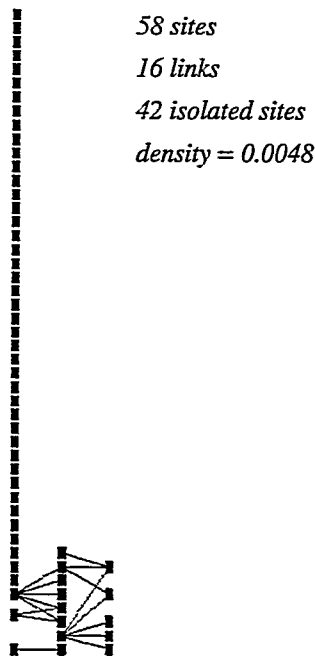


Figure 2: Sparsely connected graph ("Italian travel")

Connected Sites: Giant Components, High Density, Short Paths

We first examined the component structure of connected sites — a component is a subset of the nodes in a graph such that there is a path from any node in the component to any other node and no paths between nodes in different components.² It is completely plausible that the sites could be fairly evenly distributed across a set of components. For example, a topic with 50 non-isolated sites could consist of components of size 20, 15, 9, and 6. However, this was not the case: if sites were not isolated, they tended to be part of one giant component. On average, there were just under 2 components per graph, and the largest component contained 90% of all the non-isolated nodes.

Another metric for measuring the connectivity of nodes in a graph is *density*. Density is simply the number of links that do exist divided by the number of links that could exist (if the graph were fully connected). For a directed graph of

² For the purposes of component analysis, we treated the graph as symmetric; that is, sites *a* and *b* were considered connected if there was a link from either *a* to *b* or from *b* to *a* (or both). And *a* and *c* are part of the same component if there is some (undirected) path between them, e.g., from *a* to *b* and *b* to *c*, even if *a* and *c* do not link directly to each other.

size *N*, a total of $N * (N-1)$ links are possible. Clearly, the Web as a whole is a very sparse graph; the probability that there is a link between any two sites selected at random is nearly 0. However, the graph defined by a particular topic area is much denser: the average density of the 45 graphs we analyzed was 0.056. This means that a graph of size 50 would contain about 138 links.

Another useful way to characterize the structural connections of the non-isolated sites is to examine the *lengths of paths* among them. We applied the following test to each topic 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 "seed". (We proceeded only if at least 20 non-isolated sites remained once we made the selection.) For each set of seed sites, we counted the number of other non-isolated sites that could be reached via a path of length 1 or 2. Figure 3 shows the results. 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.

To understand this experiment thoroughly, let us consider the case of 5 randomly selected seeds in a bit more detail. 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 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 seed sites.

This experiment shows how tightly topically related sites are knit together by inter-site links. It also has interesting consequences for site discovery, showing that one can start with a small set of seed sites known to be on a topic and find a large majority of related sites by following just one or two links. We discuss this point in more detail later.

Non-collaborative Topics: It's Money That Matters

We examined the data further to better understand the large number of isolated sites — were all topics roughly the same in this respect, or were there differences based on the type of topic? We immediately observed a striking difference between Business and Economy topics, such as lawn and garden companies, consumer electronics, and year 2000 software organizations, and other types of topics. For 10 B&E topics, 79% of sites were isolated, for the other 35 topics, only 32% of sites were isolated. (This was a significant difference, using a standard two sample t-Test, $t(43) = -6.02$, $p < 0.0001$.) There was also a large difference in density: the average density for B&E topics was 0.004, and for the other topics it was 0.071. (However, this difference was not significant at the 95% level: using a standard t-Test, $p(43) = 1.87$, $p = 0.068$.) Again, to be concrete, in a graph of 50 sites, this is the difference between 10 links and 174. To sum up, topics dominated by merchants who are competing for the same customers do not exhibit collaboration.

Proportion of sites reachable by paths of length 1 or 2

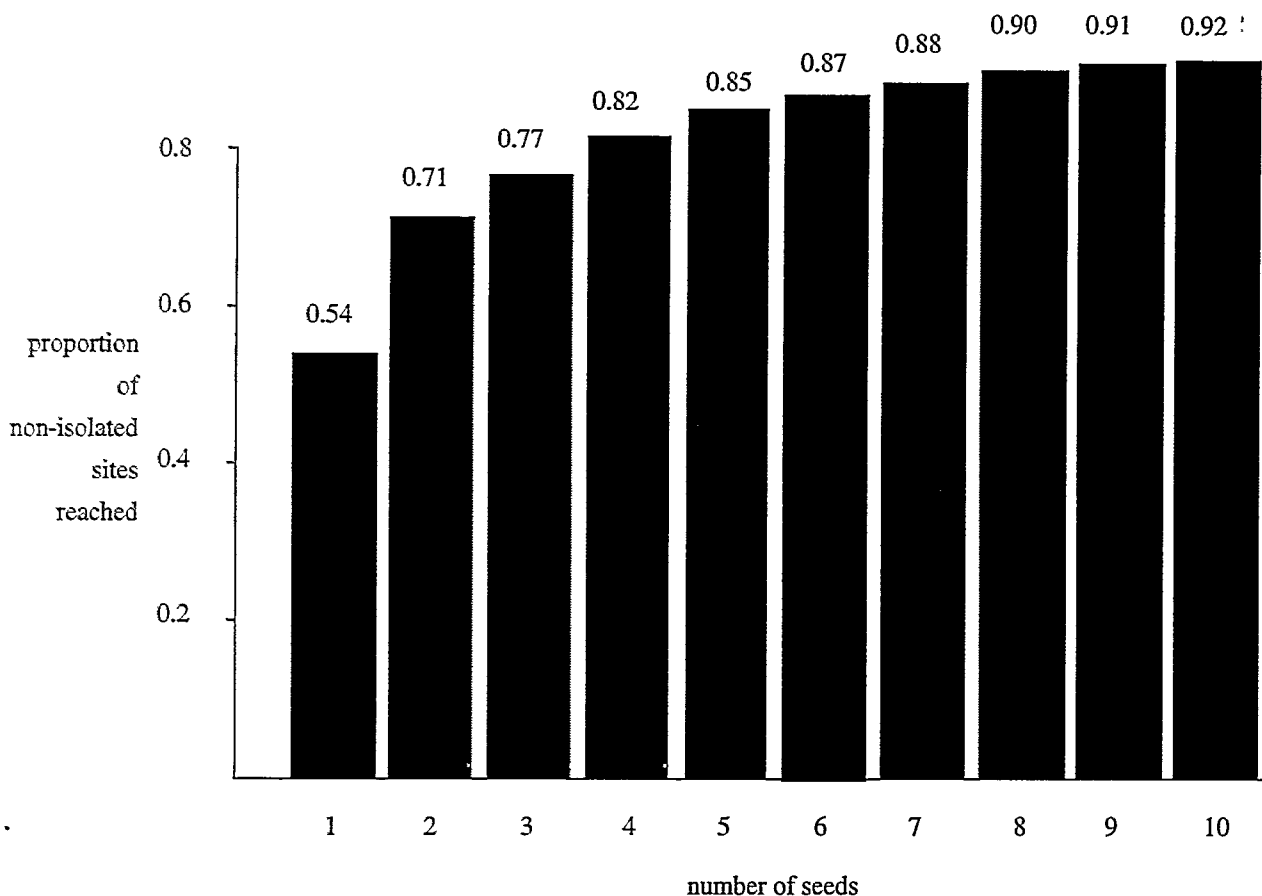


Figure 3: Reachability of sites by paths of length 1 or 2

Figure 4 (see also Color Figure 1) summarizes our findings concerning isolated sites and components. Each bar represents one topic. The gray upper block in each bar represents the number of isolated sites. Lower blocks show the number of sites in each component of the graph. The labels on the X axis indicate the type of each topic (see Table 1 — “A” is Arts and Entertainment, “S” is Science, etc.) For example, the bar with the label “NLP, which represents the topic Natural Language Processing, has one component of size 16 (the lowest block), then components of size 3, 2, and 2, and 22 isolated sites.

Some of the bars in the plot appear to be exceptions to our generalization about Business and Economy topics being non-collaborative. It is worth discussing several cases individually since, upon closer examination, they shed more light on the generalization. “Coin collecting” is a B&E topic, yet has a very small proportion of isolated nodes. Although the majority of the sites are commercial and do not link to each other, there are a few index sites that link to

many of the other sites and thus knit them together into a large connected component. However, the overall density of the graph is still low (190 links in a topic of 149 sites = 0.0086), which shows that very little linking between sites actually is occurring.

Two Recreation topics have a large proportion of isolated nodes. As discussed earlier, in the “Italian travel” topic, 42 of the 58 sites were isolated, and the density was only 0.0048. This was because the topic contained a large number of sites for travel agencies and official city and regional information bureaus, which do not link to each other. The “interactive fiction” topic contained a large number of sites for computer games and software companies, and it is these sites which tend to be isolated from the rest. (108 of 176 sites were isolated, and there were 120 links, for a graph density of 0.004.) Thus, we see that some topics may have a “mixed” nature, containing both commercial sites, which do not link, and sites maintained by enthusiasts or hobbyists, which do.

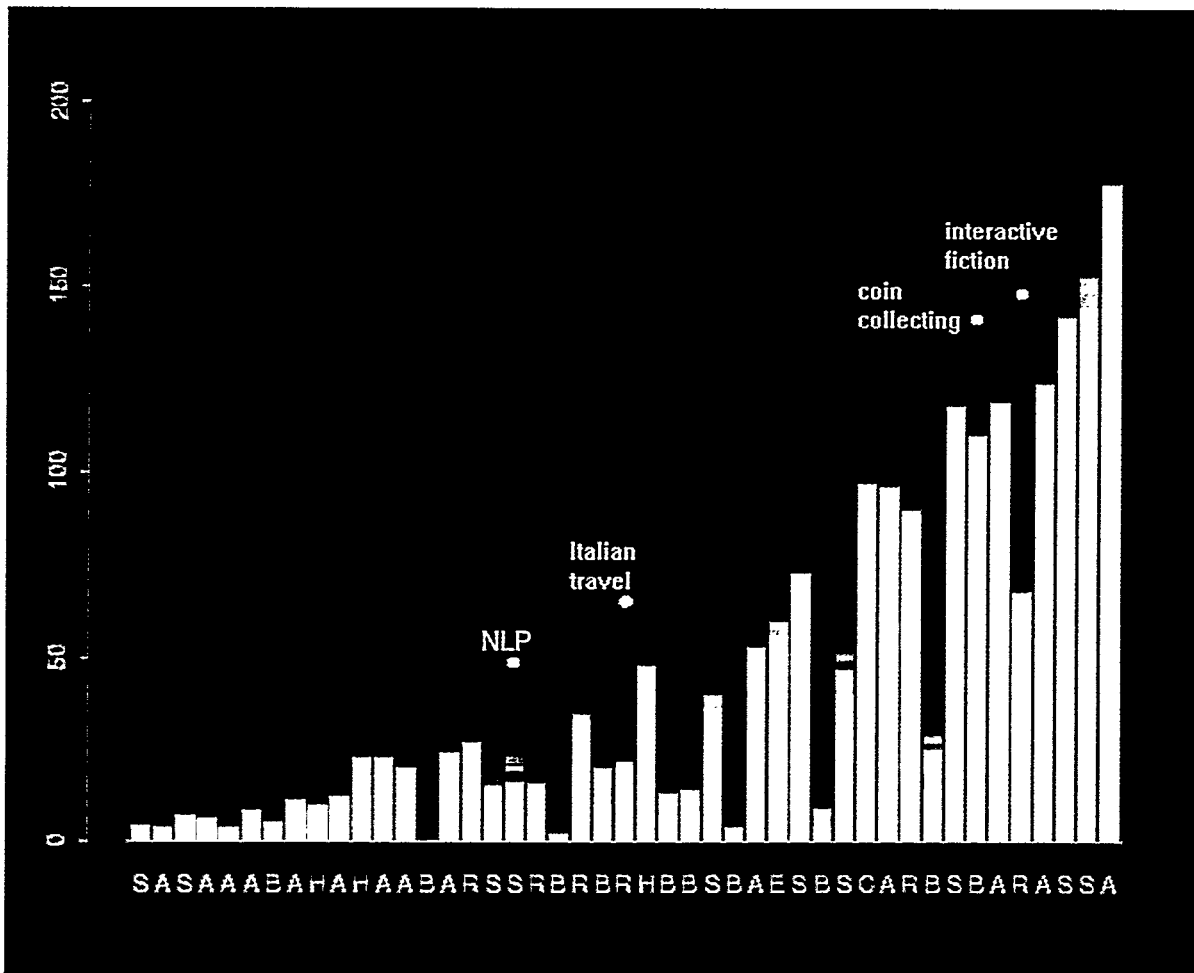


Figure 4: Components and isolated sites

We see two major implications of our findings. First, link-based collaborative filtering can work only for topics with a significant number of inter-site links. We have identified one large topic area where such links are extremely rare and have suggested that it is because commercial sites that are in competition with one another do not link. We will look for other topic areas that are especially collaborative or non-collaborative as we continue our investigations.

Second, we found that nearly all (90%) the sites for a topic that are not isolated group into one large component, and that nearly all of the non-isolated sites are within 2 links of a small, randomly selected subset of "seed" sites. The patterns of connection that emerge from the implicit collaboration among web site authors are potentially useful in two ways. First, connectivity may be used as an estimate of relevance and quality, e.g., in ordering the presentation of sites in an interface. Second, site discovery algorithms can find a vast majority of sites related to a set of seed sites by following just 1 or 2 links from the seeds. We discuss each of these points in detail in the next two sections.

DO LINKS BETWEEN SITES CORRELATE WITH HUMAN RELEVANCE AND QUALITY JUDGEMENTS?

The 23 people who rated topics collectively rated the relevance and quality of 346 sites. The mean relevance of these sites was 3.99, and the mean quality was 3.16 (both relevance and quality were rated on a scale of 1 to 5, with 1 being least relevant/lowest quality and 5 being most relevant/highest quality).

We investigated the relationship of site connectivity to relevance and quality judgements. It was our hypothesis that connectivity would influence these judgements, and the data support this hypothesis. Of the 309 sites, 88 were isolated, and 258 were connected to other sites in their topic. The mean relevance of connected sites was 4.09 and of isolated sites was 3.70. (This was a significant difference, using a standard two sample t-Test, $t(344) = 2.96$, $p < 0.005$). However, there was a much larger difference in the quality of connected and isolates sites. The mean quality of connected sites was 3.33 and of isolated sites was 2.66. (This was a significant difference,

using a standard two sample t-Test, $t(344) = 4.36$, $p < 0.0001$). And there was a modest positive correlation between the in-degree of a site (the number of other sites that linked to it) and its quality of 0.23 ($r(344) = 0.23$, $p < 0.001$). Table 2 summarizes these results.

Table 2: Relevance and Quality

	All sites	Connected sites	Isolated sites
Relevance	3.99	4.09	3.70
Quality	3.16	3.33	2.66

The difference in quality between isolated and connected sites is quite large and worth exploring in more detail. One way we did this was by forming three subsets of sites: isolated sites, those with exactly one in-link, and those with multiple in-links. (Our intuition was that in-links should be a better quality estimator than out-links, since they show that a site's peers found it worth linking to; in-links did in fact prove to be more correlated with quality than were out-links) For each of these sets, we then computed the proportion of low quality (quality ≤ 2) and high quality sites (quality ≥ 4). As the connectivity increases, the proportion of low quality sites decreases, and the proportion of high quality sites increases. For the 88 isolated sites, 39 were bad (44%) and 24 were good (27%). For the 49 sites with just one in-link, 14 were bad (29%) and 18 were good (37%). For the 155 sites with multiple in-links, 39 were bad (25%) and 81 were good (52%). Figure 5 shows how the proportion of good sites increases with connectivity.

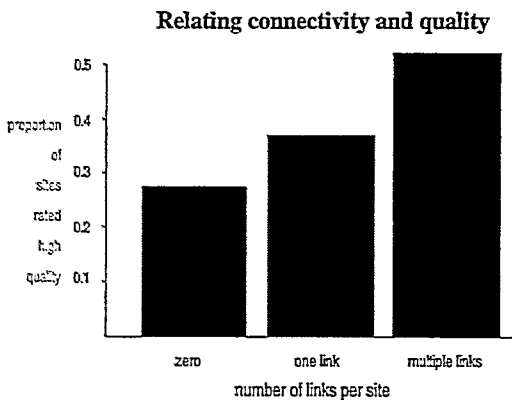


Figure 5: Quality and connectivity

Another way to see the power of connectivity is to consider the correlation between the density of a topic graph and the mean quality rating the subject assigned to sites for that topic. There was a fairly strong correlation between density and mean quality of 0.51 ($r(21) = 0.51$, $p < 0.05$). Therefore, topics that are collaboratively bound together by links tend to contain higher quality information.

These data show that using connectivity data could significantly boost the quality of sites contained in index

sites, whether Yahoo or specialized indices maintained by someone who is interested in a particular topic. For example, it would be a reasonable strategy to order sites by their degree or in-degree, so the proportion of high quality sites will be highest at the top of the list.

We also can observe how manual collaborative filtering can be enhanced by computational techniques. A person might construct an initial index, then apply algorithms to both order the items in the initial index and suggest new items for addition to the index. We discuss this topic of site discovery briefly in the next section.

FROM RANKING TO DISCOVERY: NK CLAN GRAPHS

We have focused so far on analysis of a given set of sites which can be assumed to be at least somewhat relevant to a topic, since they were included in a manually constructed index page. We also mentioned the companion problem of discovering additional sites related to a given set of sites. We want to discuss our approach to this problem briefly in light of the results presented here.

Our webcrawler is part of a more comprehensive system that constructs the *NK clan graph* [13] relative to a set of "seed" sites. The NK clan graph for a seed set S is $\{(v,e) \mid v \text{ is in an N-clan with at least K members of S}\}$. An N-clan [8] 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-clans, that is, the 2K clan graph. The indicated subgraph of Figure 6 shows an example of a 2-clan contained within a larger graph.

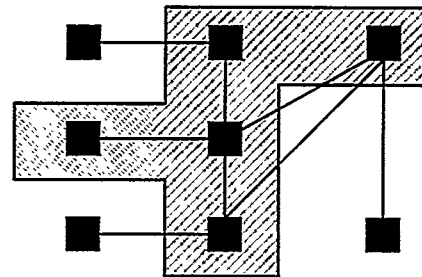


Figure 6: An example 2 clan

We proposed the NK clan graph as a useful formalization of the notion of a topically related locality of web sites (or, more generally, of documents in a hyper-linked collection). Thus, we claimed, if you know of a few sites on a particular topic, our system can construct the NK clan graph for you, and the sites it discovers and adds to the graph are quite likely to be on topic. Our argument in favor of this claim was purely analytic: we pointed out that any documents in a 2-clan also would participate in either collaborative filtering, co-citation, or citation transitivity relationships. Since these relations all group together related items, we claimed that the NK clan graph would tend to do so, too.

This paper adds an empirical argument for the use of the 2K clan graph. We showed that the vast majority of sites

known to be on a topic (known because they occurred in the same index page) can be generated by following 1 or 2 links from a small set of seed sites for the topic. This is precisely how far we look when building the 2K clan graph.

In information retrieval terms, what the empirical evidence shows so far is that the NK clan graph is likely to offer good recall, i.e., it is likely to include the vast majority of web sites related to the seed sites. We still need to prove that it shows good precision, i.e., that it does not discover many irrelevant sites along with the relevant ones. That is the focus of another experiment that we are conducting in parallel with the one reported here.

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

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

CONCLUSIONS

We reported on an empirical investigation into patterns of collaboration emergent from links between web sites. We both presented novel algorithms and offered empirical evidence of their utility. We showed that topics dominated by commercial sites are not collaborative at all, but that, in other topics, most sites are knit together into large, closely connected components. We also showed that connectivity is related to human relevance and quality judgements: sites from more densely connected topics receive higher overall quality ratings, and individual sites that are linked with their peers receive higher ratings than isolated sites. Finally, we argue that these findings suggest a promising approach to discover new web sites related to an initial set of seeds.

ACKNOWLEDGEMENTS

We are deeply grateful to all the members of our laboratory who supplied us with topics and spent the time to give us relevance and quality judgements.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy

REFERENCES

1. Botafogo, R.A., Rivlin, E., and Shneiderman, B. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems* 10, 2 (April 1992), 142-180.
2. Carrière, J., and Kazman, R. WebQuery: Searching and Visualizing the Web through Connectivity, in *Proceedings of WWW6* (Santa Clara CA, April 1997).
3. Garfield, E. *Citation Indexing*. ISI Press, Philadelphia, PA, 1979.
4. Hill, W. C., Hollan, J. D., Wroblewski, D., and McCandless, T., Edit Wear and Read Wear, in *Proceedings of CHI'92*. (Monterey CA, May 1992), ACM Press, 3-9.
5. Hill, W.C., Hollan, J.D. History-Enriched Digital Objects: Prototypes and Policy Issues. *The Information Society*, 10, 2 (1994), 139-145.
6. 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.
7. Kleinberg, J.M. Authoritative Sources in a Hyperlinked Environment, in *Proceedings of 1998 ACM-SIAM Symposium on Discrete Algorithms* (forthcoming).
8. Mokken, R.J. Cliques, Clubs and Clans. *Quality and Quantity* 13, (1979), 161-173
9. 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.
10. 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.
11. Seligman, D.D. <http://medusa.multimedia.bell-labs.com/LWS/>
12. 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.
13. 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.
14. Wexelbat, A. History-Rich Tools for Social Navigation, in submission to *Communityware'98*,. See also <http://footprints.media.mit.edu>.
15. Wroblewski, D., McCandless, T., Hill, W. Advertisements, Proxies and Wear: Three Methods for Feedback in Interactive Systems, in *Dialogue and Instruction*, Beun, R., Baker, M., and Reiner, M. (eds.), 1994, Springer-Verlag.