Life, Death, and Lawfulness on the Electronic Frontier

James Pitkow

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto CA 94304 USA
pitkow@parc.xerox.com

Peter Pirolli

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto CA 94304 USA
pirolli@parc.xerox.com

ABSTRACT

To facilitate users' ability to make sense of large collections of hypertext we present two new techniques for inducing clusters of related documents on the World Wide Web. Users' ability to find relevant information might also be enhanced by finding lawful properties of document behavior and use. We present models and analyses of document use and change for the World Wide Web.

Keywords

Clustering, categorization, co-citation analysis, World Wide Web, hypertext, survival analysis, usage models

INTRODUCTION

The ever-increasing universe of electronic information competes for the effectively fixed and limited attention of people. Both consumers and producers of information want to understand what kinds of information are out there, how desirable it is, and how its content and use change through time. Our work aims to discover empirical regularities of hypertext content, use, and structure, and ways of exploiting these regularities to provide new ways of helping people to find and make sense of information.

Making sense of very large hypertext collections and foraging for information in such environments is difficult without specialized aids. The basic structure of hypertext is designed to promote the process of browsing from one document to another along hypertext links, which is unfortunately very slow and inefficient when hypertext collections become very large and heterogeneous. Two sorts of aids seem to evolve in such situations. The first are structures or tools that abstract and cluster information in some form of classification system. Examples of such would be library card catalogs and the Yahoo! WWW site [18]. The second are systems that attempt to predict the information relevant to a user's needs and to order the presentation of information accordingly. Examples would include search engines such as Lycos [11], which take a user's specifications of an information need, in the form of words and phrases, and return ranked lists of documents that are predicted to be relevant to the user's need.

One aim of our work is to develop methods that automatically categorize and aggregate hypertext information. Another aim is to develop automatic methods that predict needed information in hypertext collections. Unlike other approaches to automatic categorization [7] or search engines [11], we seek to use more than just text content as the basis for our techniques. We think that the browsing patterns of users and the changing structure of hypertext documents and links provide important data that can be exploited to enhance our methods. Just as the regularities of texts and language have been exploited in current search engines, we expect there to be regularities of use and hypertext that can also be exploited.

We are investigating these methods as potential enhancements to an Information Workspace [5] that is connected to the WWW. The Web Forager [6] is an example of such an Information Workspace. It supports a variety of interaction techniques for finding, grouping, and abstracting collections of WWW documents. The Web Book [6] is an example of a structure supported by the Web Forager, but in its current form Web Books are constructed largely by direct manipulation techniques. Automatic categorization and aggregation techniques could be used to rapidly create structures such as Web Books. The ability to predict needed information based on a user's current focus of attention could be used to order and arrange information in the workspace, or in pre-fetching documents that the user is likely to read next. The current sluggishness of page retrieval on the World Wide Web could be greatly alleviated by such pre-fetching.

In this paper, we extend our previous work on clustering WWW documents [14] with new clustering techniques. These techniques are inspired by ones developed for the analysis of the structure of scientific literatures [9] and exploit the link structure of hypertext. We also propose an intrinsic property, called *desirability*, of WWW documents, that might be used to predict future need and use. We present a simple model of the time-course of desirability of WWW pages that explains some statistics of WWW use. This model is, in fact, quite general to information use on many kinds of media. We then turn from the analysis of information consumption to concern ourselves with the analysis of the dynamics of information production. We present a model of the time-course of document creation,

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
GVU Technical Report: GIT-GVU-97-12

J. Pitkow & P. Pirolli

1

change, and deletion on the World Wide Web. This model explains some factors that affect the survival and change dynamics of documents.

CLUSTERING THE WORLD WIDE WEB

One way to approach the automatic clustering of hypertext documents is to adapt the existing approaches of clustering standard text documents [7]. However, there are several impracticalities with such existing text-clustering techniques. Text-based clustering [7] typically involves computing inter-document similarities based on content-word frequency statistics. Not only is this often expensive, but, more importantly, it's effectiveness was developed and tuned on human-readable texts. It appears, though, that the proportion of human-readable source files for WWW documents is decreasing with the infusion of dynamic and programmed pages.

Other attempts at clustering hypertext typically utilize the hypertext link topology of the collection [3]. These clustering methods have been applied to collections with several hundred elements, and do not seem particularly suited to scale gracefully to large heterogeneous collections like the WWW, where over 70 million text-based documents currently exist [11].

In our own previous work [14], we represented each WWW document as a feature vector, with features extracted from information about text-content similarity, hypertext connections, and usage patterns. Clustering was then computed from inter-document similarities among these feature vectors. Unfortunately, any clustering based on usage patterns requires access to data that is not usually recorded in any easily accessible format. In the case of the WWW, while a moderate amount of usage information is recorded for each requested document at a particular WWW site, the log files for other sites are not publicly accessible. Thus while the usage for a particular site can be ascertained, this information is not available for the other 500,000 WWW sites that currently exist [13].

As a potential way of circumventing these difficulties, we decided to try out *co-citation analysis* [9]. Our adaptation of this clustering technique is based solely on the analysis of hypertext link topology. Unlike earlier link-topology techniques, co-citation analysis builds upon the notion that when a WWW document D contains links referring to documents A and B, then A and B are related in some manner in the mind of the person who produced the document. In this example, documents A and B are said to be *co-cited*. It is important to note that links between document A and document B may or may not exist. Given this property of picking up patterns from the implicit topological structure of hypertext documents, we hypothesized that co-citation analysis might be useful in telling us something about the semantic structure of a collection and the thinking of the authoring community.

CO-CITATION ANALYSIS

Citation indexing, the creation of an index that details the explicit linkages of citations between papers, has been employed as a tool to facilitate the searching and the man-

agement of information for over a century, dating back to the legal profession's use of the *Shepard's Citations* in 1873. The field underwent major advances during the post World War II increase in scientific expenditures and subsequent explosive increase in the scientific literature. With the intent of ensuring information exchange among scientists, the United States government initiated a number of projects to generate indexes without human involvement. Citation indexing was found to be a powerful yet simple tool, as it replaces an indexer's subjective judgements with author's citations, thus avoiding many of the semantic problems found in term and title based analyses [9].

It was not until the mid-1970s however that Small and Griffith [16] developed co-citation analysis as a method for measuring the common intellectual interest between a pair of documents. The principal component of co-citation analysis measures the number of documents that have cited a given pair of documents together. This metric is referred to as *co-citation strength*. Unlike other forms of citation analysis, co-citation strength is able to reflect the frequency of items being cited over time, thus enabling deeper insight into the development of certain research fields and other semantic structures within a citation index. We hypothesize and later show that co-citation analysis yields insight into the implicit semantic structures of the WWW.

Algorithm

The original algorithm developed by Small and Griffith [16] takes a citation index as initial input. For all documents in the index, the number of times a document was cited is computed and those documents whose *cited frequency* falls above a specific threshold are kept for further processing. This prefiltering retains the most important (or at least the most popular) documents. Next, the extracted documents are sorted and all pairs of documents that have been cited together by the same source document are formed. The resulting list contains unique co-citation pairs and their associated frequency of co-occurrence.

The final step in co-citation analysis creates a set of clusters whose elements are indirectly or directly related by co-citation. This is accomplished by clustering all documents that have at least one document of the co-citation pair in common with the other elements in the cluster. To start, a pair is selected, say AB, and all pairs that contain A or B are added to the cluster. Next, all pairs that include a document that have been included in the cluster are added. This processes repeats until there are no pairs that have a document in common with the elements in the cluster. At this point, a new pair is selected from the remaining pairs to form a new cluster and the processes repeated until all pairs belong to a cluster.

Application to the WWW

It is interesting to note that the properties that fueled the development of citation and co-citation analysis are similar to those found with the WWW. Hyperlinks, when employed in a non-random format, provide semantic linkages between objects, much in the same manner that citations link documents to other related documents. The resulting topology of a Web site reflects the organization of a community and its

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
GVU Technical Report: GIT-GVU-97-12

J. Pitkow & P. Pirolli
2

knowledge base, similar to the way in which citations in a scholarly paper reflect a scientific community's organization of knowledge.

One might argue that hyperlinks often serve as just navigational aids. Still, the role of hyperlinks for navigation can be viewed as a hypothesis by the hypertext author(s) that the person interested in the current page will also be interested in browsing the linked pages. It was our belief that given the close resemblance of hyperlinks to citations, meaningful structures would emerge as the result of co-citation analysis on WWW ecologies.

During early September 1996 we extracted the hyperlink structure of the Georgia Institute of Technology's Graphic Visualization and Usability (GVU) Center WWW site which contained 5,582 HTML files, 15,139 non-HTML files and 24,768 hyperlinks[1]. This site was chosen because of its loosely structured properties, i.e., the site contained a large number of documents authored by hundreds of people over the course of several years. The co-citation clustering analysis mentioned above was applied using several different citation frequency thresholds (one, three, five, and ten). Table 1 shows the distribution of the size of clusters using different citation frequency thresholds. For example, using citation frequency threshold of three, there were six clusters formed where each cluster contained between 101 and 500 pages. Table 2 shows the number of pages each range of cluster sizes produced. Overall, the six clusters that contained between 101 and 500 pages collectively contained 979 pages. Since co-citation analysis using the citation frequency threshold of three resulted in 2,798 pages being clustered, over a third of the pages are contained in the six medium sized clusters.

Cluster Size (Pages)	Citation Frequency Threshold			
	1	3	5	10
3 - 6	34	4	2	2
7 - 10	12	2	1	1
11 - 20	14	0	1	1
21 - 50	8	1	2	0
51 - 100	4	1	1	2
101 - 500	7	6	3	3
501 - 1000	0	1	0	0
1,001+	1	0	0	0
Total	80	15	10	9

Table 1: For each range of cluster sizes, the total number of clusters formed are given for various citation frequency thresholds.

1. Objects embedded into HTML pages, e.g., images, were not considered hyperlinks for this analysis.

Cluster Size (Pages)	Citation Frequency Threshold			
	1	3	5	10
3 - 6	136	15	8	8
7 - 10	97	17	7	8
11 - 20	213	0	16	14
21 - 50	211	21	92	0
51 - 100	319	95	93	163
101 - 500	1,747	979	687	520
501 - 1000	0	1,671	0	0
1,001+	3,315	0	0	0
Total	6,038	2,798	903	713

Table 2: The total number of pages included in the range of cluster sizes using various citation frequency thresholds.

As one would expect, lowering the number of times a document is cited results in more documents being included into the co-citation analysis. This results in the formation of more clusters as well as the formation of larger clusters. Our analysis included a cited frequency of one to show the effects of including documents that do not necessarily contribute to the definition of a specific area. Since these documents were only cited once, it is likely that the community of authors has failed to reach consensus on the importance of these documents with respect to the ecology of the entire Web. We observe that from the clusters formed from the cited frequency of five and ten, a certain degree of agreement has been reached by the authoring community on the intellectual structure of the set of pages. This is reflected in the similarity of the cluster sizes and their respective elements as well as in the actual elements included in each cluster as determined from random inspection.

Cluster	Number of Pages	Description
Sub Arctic	177	Documentation specific to the Sub Arctic Toolkit
Talk Slides	46	Group of slides for a talk converted to HTML
WWW Surveys	19	Main pages for GVU's WWW User Surveys
GravityWeb	4	The major stories for an online humour publication

Table 3: Examples of the types and sizes of clusters formed by co-citation analysis.

The trend for a significant proportion of the documents to belong to a few large clusters is an effect typically found in traditional co-citation analysis of publications [9]. These

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
GVU Technical Report: GIT-GVU-97-12

J. Pitkow & P. Pirolli

3

large clusters consist of diverse set of pages, reflecting a loose semantic coupling among elements. For example, for the set of clusters formed with the cited frequency set at five, the cluster containing 387 elements was composed of different projects' online documentation, peoples' personal pages, specific project pages, and online presentations. The smaller clusters tend to form sets of tightly related pages, typically composed of all the same types of elements, e.g., a specific project, online book, etc. Table 3 shows examples of the types of clusters formed using this form of co-citation analysis.

Other Co-citation Techniques

Other techniques used in co-citation analysis include hierarchical clustering, multi-dimensional scaling, and factor analysis [12], though these techniques typically use authors as the unit of analysis instead of documents. In the case of the WWW, the author of a document can not be reliably determined across sites. In our version of these techniques, the co-citation matrix is computed as outlined above using a cited frequency threshold to reduce the set of potential candidates for clustering. Rather than use the iterative method of cluster formation described above, the techniques are based on the computation of similarities among co-citation patterns for each document [17].

For our analysis, we took the three, five, and ten co-cited frequency threshold co-citation matrices computed in the above analysis and calculated the euclidean distance matrix on the log transformed co-citation frequencies. The resulting distance matrix was than run through a complete linkage clustering algorithm.

Figure 1 shows the partial results of the clusters formed from the cited frequency five threshold matrix. The number of pages is noted in parenthesis. Several interesting and useful structures emerged. The 'Organization Pages' shows that the clustering pulled the College of Computing and GVU's Organization home pages together into one cluster. This two element cluster was not formed in the iterative clustering method, as all the pages that co-cited with either of these pages were been included in the cluster that contained these two elements. The algorithm successfully clustered the pages of an online class, as well as the online documentation for the SVE library and Sub Arctic (SA) projects. Within the "People" cluster, subclusters were formed that separated out the people from the papers they publish. Further inspection of the clusters revealed many other interesting and well-formed clusters, though no empirical evaluation of the "goodness" of the clusters was performed.

With the goal of facilitating users' ability to make sense of large collections of hypertext by reducing the number of documents necessary to explore, we have presented two new clustering techniques. We now present a general model of document desirability, and extend our analyses of WWW interaction to issues concerning the life histories of WWW documents (birth, changes, and death). The lawful properties discussed below can be used to further refine the structure of information presented to users.

THE DESIRABILITY OF DOCUMENTS

In previous work we analyzed patterns of WWW use to induce retrieval structures that could be used by a spreading activation mechanism [1] to predict documents relevant to a
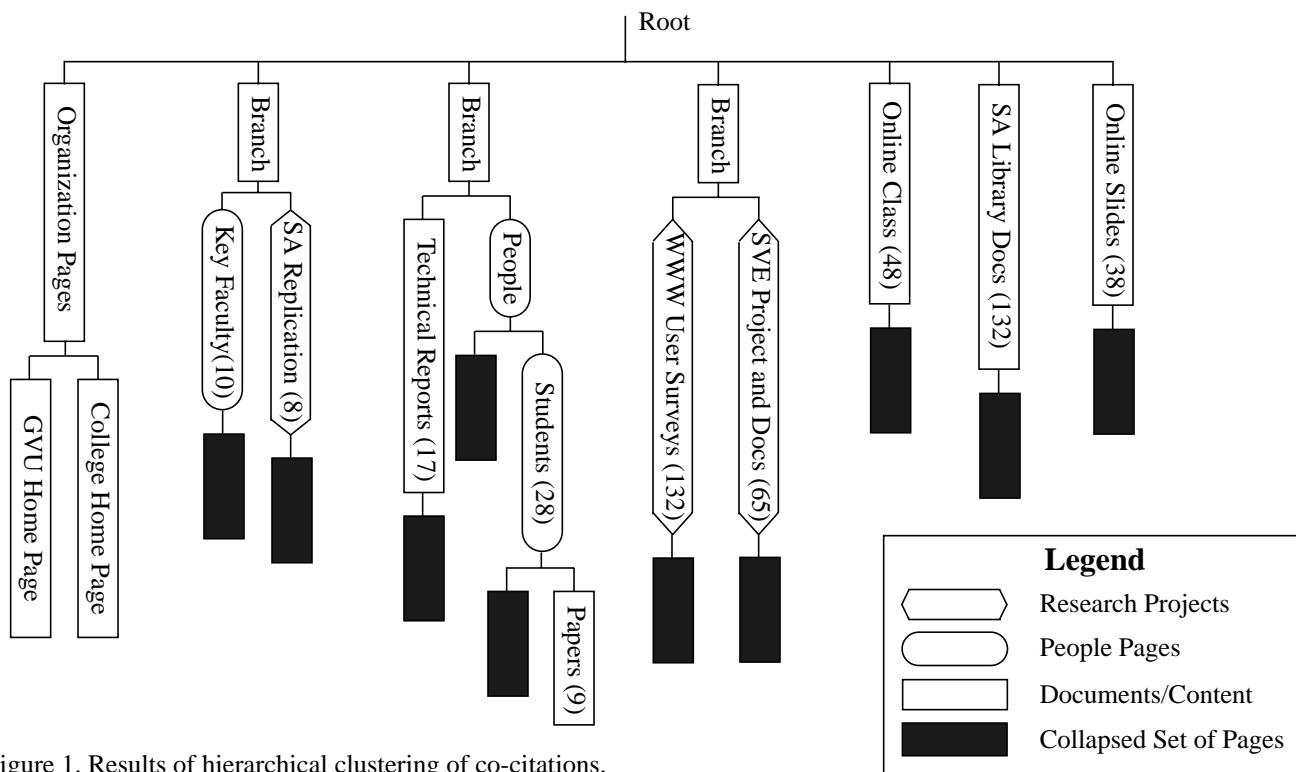

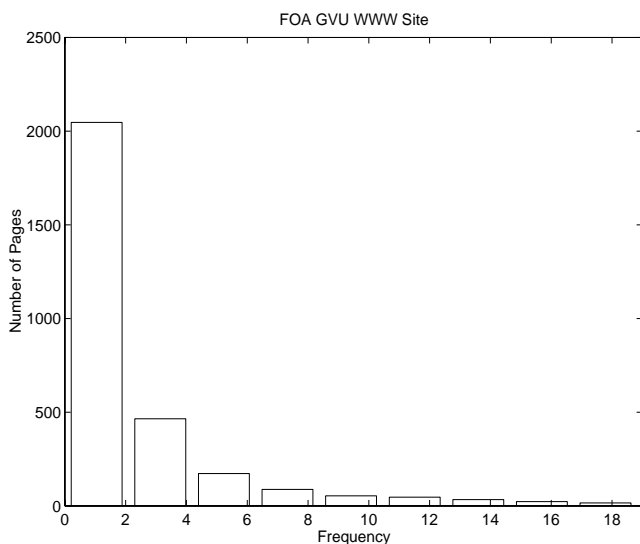Figure 1. Results of hierarchical clustering of co-citations.

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
GVU Technical Report: GIT-GVU-97-12

J. Pitkow & P. Pirolli
4

Figure 2. Typical frequency of access (FOA) curve for a WWW ecology, which in this case is GVU's WWW site.

users interest. In essence, we were predicting the need probability or *desirability* of information. A broadly applicable and approximate model of the distribution and time-course of information desirability is, what we call, the Burrell Gamma-Poisson (BGP) model [4]. In its original form [4] it was used to model the circulation of library holdings. In modified form, it has been found to model the recurrence of information use (or need) in child language, newspaper headlines, and electronic mail [2]. The applicability of the BGP model to the WWW is partly justified by the key observation that the distribution of frequency of access (FOA, or page hits) across WWW pages is approximately a negative binomial distribution (see [10] for discussions of how to judge the appropriateness of the negative binomial as a model of an empirical distribution).
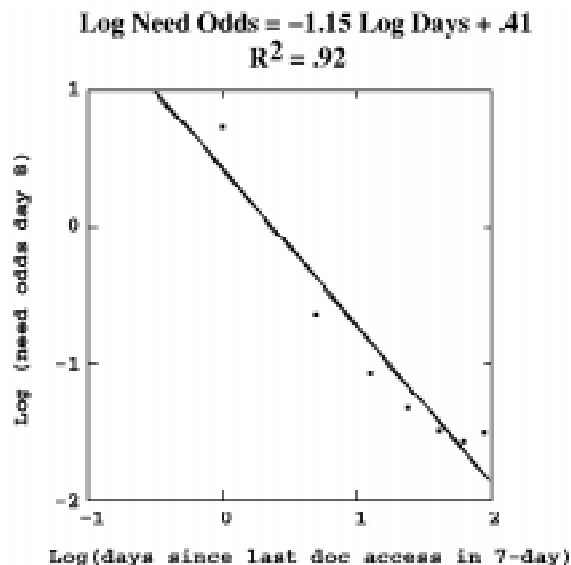

Figure 3. Relationship between recency and need odds.

Figure 2 is a plot of the FOA for the GVU Center's WWW site on April 27, 1996. Basically the BGP model assumes that accesses to information are Poisson events, and the desirability of individual information elements (e.g., WWW pages) is modeled by the Possion parameter $\lambda(t)$. In Poisson models, this parameter determines the average wait time between events. In this case, it models the average time between accesses of a particular WWW page.

The Poisson parameter, $\lambda(t)$, is time dependent (nonhomogenous) on time t. This is because the desirability of information changes through time. Figure 3 shows the typical average decay in odds of accessing WWW pages. This power law relationship between odds of access and time appears to be ubiquitous [2][15]. Conceptually, the desirability of faddish WWW pages, long-lasting popular pages, and occasionally used but enduring reference material is presented in Figure 4.

The BGP model addresses patterns of information use, but we are also interested in modeling the way that documents themselves change through time. These changes occurring over the lifetime of document probably interact with the patterns of use. Towards this end, we hypothesized that survival analysis would augment our existing models by identifying principal factors determining the life-cycles of documents on the WWW. From this analysis, we can gain a better understanding of the relationships that exist between the attention documents receive from the users of the information as well as the attention from the authors of the information.
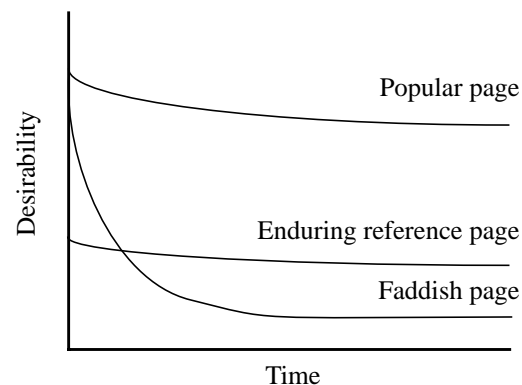

Figure 4. Different desirability curves across time.

SURVIVAL ANALYSIS

Survival analysis models the time to death of specific entities. In the case of the WWW, these entities are items within a given WWW collection. The probability that a particular item will be deleted at particular time forms the basis of the model. It seems reasonable to suspect that highly desirable documents are not likely to be removed from WWW sites. If also seems reasonable that as the desirability of a document increases so does the likelihood that the document will change accordingly.

Survival analysis is a well-developed field of statistical research [8]. A survival function defined over time t is the probability that an entity survives at least to time t. More

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
GVU Technical Report: GIT-GVU-97-12

J. Pitkow & P. Pirolli
5

formally, let T be a positive random variable with the distribution function $F(t)$ and density $f(t)$. The survival function $S(t)$ is:

$$S(t) = 1 - F(t) = P\{T > t\}$$

and the hazard rate $\lambda(t)$ is:

$$\lambda(t) = \frac{f(t)}{S(t)}$$

which translates to the probability that a page will be deleted in the next unit of time, Δt, given hat the page has survived to time t.

Unlike other forms of statistical analysis, survival analysis handles items that are still alive, where the true survival time is only known to be greater than the end of the observation period. These observations are said to be *censored*. As one would expect, for WWW ecologies the amount of censored data is quite large, as the majority of items are not deleted once published on the Web. Like other forms of statistical analysis, the survival observations can be stratified and the presence of statistically significant differences between the strata computed.

Methodology

Software was implemented to collect the changes to the objects within the GVU WWW site on a daily basis. As mentioned for the co-citation analysis, the GVU sites was chosen for its loose, large structure and the presence of many authors. The changes observed by the software included the addition and deletion of material as well as any modifications made to the material. Properties about the file, e.g., author, filesize, etc., were recorded as well as the structure of the hyperlinks between items. From this data, the age and censoring status of each item on the site was computed as well as the probability that an item will change on a daily basis. The probability of change can be viewed as a direct component of the attention given by the producer of the information. Data was collected for a 215 day period, starting January 15, 1996.

In trying to weave the life story of items on the Web, the attentional processes of the consumers of the information are also important. That is, how much attention is an item receiving and to what community does the information primarily serve? These attentional processes could originate from within the organization's intranet, originate from consumers on the Internet, or be driven mutually by both internal and external communities. We hypothesized that information that is driven primarily by the external community would have the greatest likelihood of survival, followed by mutually driven and internally driven information. Attention in this case can be viewed as a form of desirability as explained in the previous section.

As a hypothetical example, the life story of an item reads that the item receives attention initially from the producer of the information as reflected in the creation and subsequent

flurry of modifications. At this point, the information is primarily being consumed internally by the author. Until some notable change occurs in the attention given to the item by internal and/or external sources, the item will continue to age, following a path of decreased modifications by the author, and have a rather good chance of being deleted sooner than later. If members of the internal and/or external communities discover the item, the survival probability for the item may change significantly.

Given that the perceived utility of the item has increased, as measured by increased consumer attention, we hypothesized that the item would be less likely to be removed, especially if the information is widely attenuated by the global Internet community. In order to model the origin of consumer attention, the GVU WWW site's access logs files for a 226 day period beginning January 15, 1996. Requests that originated within the 'gatech.edu' subdomains were considered internal requests with those originating outside of 'gatech.edu' considered to be external requests.

Figure 5 shows the access history for internal and external requests to the GVU sites. Inspection of Figure 5 reveals that for external accesses, a steady positive trend is observed, while the internal accesses show a positive trend until around day 150, when the Spring quarter ended. The spike around day 60 resulted from an internal process that repeatedly refreshed a set of pages.

Information on a per file basis was also recorded (over 700,000 entries for the sample period). As one would expect, the access rates per file vary considerably. Despite this difference in visitation rates, both internal and external access curves show asymptotic behavior at high access rates for a few selected pages. This suggests that certain information's desirability will be driven by internal sources, while other information's desirability will be driven by external
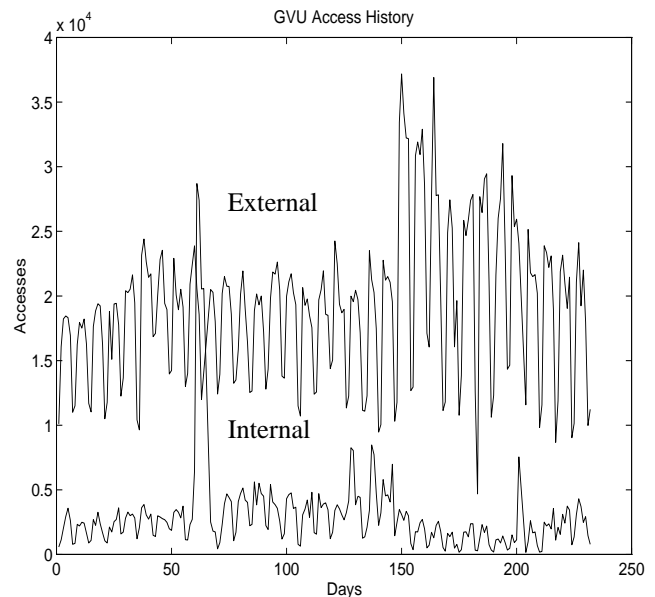

Figure 5. Number of accesses to the GVU site split by external and internal requests.

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
GVU Technical Report: GIT-GVU-97-12

J. Pitkow & P. Pirolli

6

sources. Naturally, the remaining pages will mutually service both internal and external desirability forces. Figure 6 reveals that indeed such a distinction exists. The curve in Figure 6 was derived from the following formula:

$$DifferenceVector_{(file,day)} = \log\left(\frac{file_{(external,day)} + 1}{file_{(internal,day)} + 1}\right)$$

$$= \log(file_{(external,day)} + 1) - \log(file_{(internal,day)} + 1)$$

which subtracts internal requests from external requests for each file for each day, resulting in the difference vector D. In Figure 6, D is sorted by value. The mean for the difference vector D is 0.51, with a standard deviation of 1.33. Negative values in Figure 6 reflect information that is primarily of internal interest, near zero values indicate information that is of both internal and external value, and strong positive values reflect pages that are of primarily external interest This stratification provides a useful metric to categorize the origin of desirability of information in WWW ecologies.

For each file, it's age, censor status, probability of change, and the origin of attention were included into the survival analysis. Additionally, the media type (HTML vs. non-HTML) files was incorporated into the analysis, as we hypothesized that the life stories of content would differ from other media given the different creation and modification costs and roles of each within a WWW ecology.

Results

For the survival analysis, we use the non-parametric Kaplan-Meier estimate of the survival distribution, the Flemming Harrington method for comparing the difference between survival curves for each strata, and the Cox proportional hazards model for regression modeling of the differ-
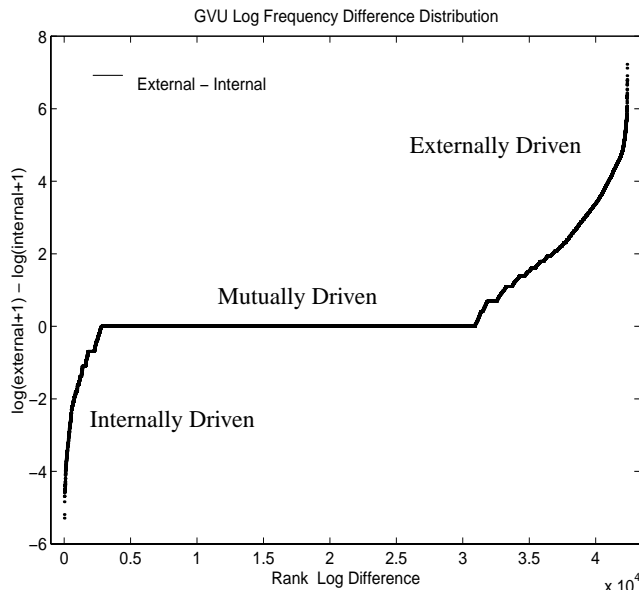
Figure 6. In semilog scale, the difference between external and internal requests is displayed on a per file basis sorted by value across the *x-axis*.

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
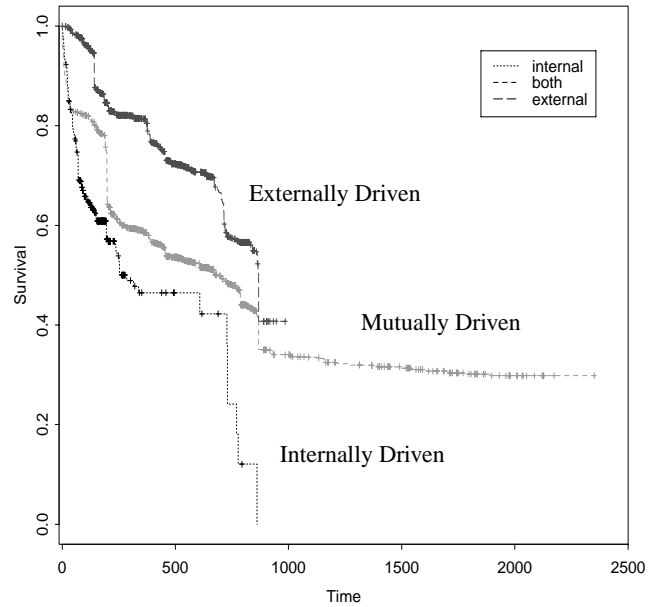GVU Technical Report: GIT-GVU-97-12

Figure 7. Three different survival curves stratified by the primary origin of requests.

ent factors influence on survival likelihood [8]. It is important to note that while the results reported below are statistically significant, they have not been replicated on other sites. This remains an area for future research.

Figure 7 shows the survival curves of the documents within the GVU WWW site stratified by origin of access. The highest curve represents externally driven pages, the middle curve, mutually driven pages, and the bottom curve, internally driven pages. From Figure 7 it is clear that internally driven pages are the most likely to be deleted whereas externally driven pages are the most likely to remain accessible to the Internet community. The difference between the survival curves was found to be highly significant across strata (χ^2 867.2, df=2, p <0.001), indicating the origin of access is a robust predictor of the likelihood of survival.

For the probability of change, three strata were created corresponding to the low, medium, and high likelihood of a document changing. These strata where then analysis and found to be significantly different (χ^2 1273.0, df=2, p<0.001). That is, items that receive a lot of attention by their author as reflected by a high rate of change have different survival properties than items that are less likely to change. Additionally, media type was also found to have a reliable impact of the survival distribution, with HTML files being more likely to be deleted than non-HTML files (χ^2 42.2, df=1, p<0.001).

DISCUSSION

From the standpoint of an information consumer, our clustering techniques are aimed at abstracting the content and likely usage of large collections on the World Wide Web.

These abstractions can be users to facilitate searching, navigation, and browsing, as they reduce the amount and complexity of the displayed inforamtion.The analysis of desirability and survival patterns of documents can also provide users with knowledge about documents: whether they are faddish-like vs. reference-like, whether they are for public consumption or internal consumption, and so on.

The techniques may also inform information producers maintainers (e.g., "webmasters"), or members of the community of authors on a World Wide Web site. The co-citation clustering can provide insight to the invisible colleges of people interacting on the World Wide Web and who they conceive of the information that they produce and manipulate. The desirability analyses can provide predictions about when information will no longer be need. The survival analyses and factoring into internal vs. external communities of interest can inform about the kinds of consumers that are attracted to specific documents. Extensions of such analysis could provide more fine-grained views of who the readers are and their intellectual tastes.

ACKNOWLEDGMENTS

REFERENCES

1. J. R. Anderson and P. L. Pirolli. Spread of activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:791–798, 1984.

2. J. R. Anderson and L. J. Schooler. Reflections of the environment in memory. *Psychological Science*, 2(6):396–408, 1991.

3. R. A. Botafogo, E. Rivlin, and B. Schneiderman. Structural analysis of hypertexts: Indentifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180, 1992.

4. Q. L. Burrell. A simple stochastic model for library loans. *Journal of Documentation*, 36(2):115–132, 1980.

5. S. Card, J. Mackinlay, and G. Robertson. The information visualizer: An information workspace. In *Conference on Human Factors in Computing Systems (CHI 91)*, New Orleans, April 1991.

6. S. Card, G. Robertson, and W. York. The webbook and the web forager: An information workspace for the world-wide web. In *Conference on Human Factors in Computing Systems (CHI 96)*, Vancouver, British Columbia, Canada, April 1995.

7. D. R. Cutting, D. R. Karger, J. O. Penderson, and J. W. Tukey. Scatter/gather: A cluster-based approach to browsing large document collections. In *The 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 318–329, August 1992.

8. T. R. Fleming and D. P. Harrington. *Counting Processes and Survival Analysis*. Wiley, Net York, New York, 1981.

9. E. Garfield. *Citation Indexing*. ISI Press, Philadelphia, Pennsylvania, 1979.

10. N. L. Johnson and S. Kotz. *Distributions in Statistics: Discrete distributions*. Houghton Mifflin, Boston, Massachusetts, 1969.

11. Lycos Inc., 1996. http://www.lycos.com.

12. K. W. McCain. *Mapping Authors to Intellectual Space: Population Genetics in the 1980s*, pages 194–216. Sage Publications, Newbury Park, California, 1990.

13. Netcraft. Inc. Netcraft survey of http servers, 1996. http://www.netcraft.co.uk/Survey/Reports/.

14. P. Pirolli, J. Pitkow, and R. Rao. Silk from a sow's ear: Extracting usable structures from the web. In *Conference on Human Factors in Computing Systems (CHI 96)*, Vancouver, British Columbia, Canada, April 1996.

15. M. Recker and J. Pitkow. Predicting document access in large, multimedia repositories: A www case study. *ACM Transactions on Computer Human Interaction*, (to appear), 1997.

16. H. Small and B. Griffith. The structure of scientific literatures, i: Identifying and graphing specialities. *Science Studies*, 4(17):17–40, 1974.

17. H. D. White. *Author Co-citation Analysis: Overview and Defense*, pages 84–106. Sage Publications, Newbury Park, California, 1990.

18. Yahoo Inc., 1996. http://www.yahoo.com.

Life, Death, and Lawfullness on the Electronic Frontier
Conference on Human Factors in Computer Systems (CHI 97)
GVU Technical Report: GIT-GVU-97-12

J. Pitkow & P. Pirolli
8

Life, Death, and Lawfullness on the Electronic Frontier   J. Pitkow & P. Pirolli
Conference on Human Factors in Computer Systems (CHI 97)   9
GVU Technical Report: GIT-GVU-97-12