

CHARACTERIZING WORLD WIDE WEB ECOLOGIES

A Thesis
Presented to
The Academic Faculty

by

James Edward Pitkow

In Partial Fulfillment
of the Requirements for the Degree
Doctorate of Science in Computer Science

Georgia Institute of Technology
June 1997

Copyright © 1997 by James Edward Pitkow

CHARACTERIZING WORLD WIDE WEB ECOLOGIES

Approved:

Dr. James D. Foley, Chairman

Dr. Stuart Card

Dr. Peter Pirolli

Dr. Mark Guzdial

Dr. Scott Hudson

Date Approved _____

ACKNOWLEDGMENTS

This research has been partially supported by an Intel Foundation Graduate Fellowship, ONR contract number N00014-96-C-0097, the Graphics, Visualization, and Usability Center, and Xerox PARC.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	iii
TABLE OF CONTENTS	iv
LIST OF TABLES	vi
LIST OF FIGURES	viii
SUMMARY	x
CHAPTER I	
INTRODUCTION	1
Desirability and Information Ecologies	1
Information Foraging	6
Categorizing to Optimize Foraging Decisions	8
Spreading Activation to Predict Needed Information	11
Applications	13
Chapter Outline	14
CHAPTER II	
DESIRABILITY	17
The Gamma-Poisson Desirability Model	24
Anderson's Modification	29
Application to WWW Information Ecologies	32
Frequency Analysis	36
Recency Analysis	43
Results from Other Data Sets	47
Summary	47
Applications	51
CHAPTER III	

STRUCTURING	54
Introduction	54
Document Typing	55
Data Sources and Collation	58
Web Categorization	62
Evaluation	66
Clustering	68
Cocitation Analysis	69
Algorithm	70
Application to the WWW	72
Other Cocitation Techniques	75
Evaluation	77
Summary	79
CHAPTER IV	
SPREADING ACTIVATION	81
Networks for Spreading Activation	83
Activation Algorithm	84
Example 1: Predicting the Interests of Home Page Visitors	86
Example 2: Assessing the Typical Web Author at a Locality	88
Combining Activation Nets	88
Evaluation	89
Comparison to Link Topology-based Approaches	90
Summary	93
CHAPTER V	
CONCLUSION	94
Summary	94
Future Directions	95
Information Ecologies	98
REFERENCES	102
VITA	109

LIST OF TABLES

Table	Page
Table 1. Summary of the data sets used in the desirability analysis. For each data set, the dates used in the analysis is given along with the total number of requests, range of requests, the average number of requests per day and the total number of requests. The attentional growth rate is the percentage of documents that were accessed for the first time during the collection period.	32
Table 2. The table shows the accesses of documents A, B, C, and D to an HTTP server across a nine day period. Cells with two letters indicates that the document was requested twice. In this example, the window size is seven days and the pane size is one day, resulting in two windows and two panes for the nine day period.	36
Table 3. For Window 1 and Window 2, the table shows how documents with different access frequencies are grouped into different frequency bins, whether or not the document was requested at least once during the pane, and the summations of each frequency bin. The need probability is computed by dividing the summation of whether the documents in each bin were needed during the panes by the total number of documents in each bin.	39
Table 4. The results of a power law model to frequency and recency data from three data sets. In all cases, recency provides a better predictor of future access than frequency.	47
Table 5. The most popular starting points for people visiting the Xerox Corporate Web site (http://www.xerox.com).	62
Table 6. Node type definitions and precision for linear combination category assignment.	65
Table 7. The top five head nodes as determined by linear combination.	67
Table 8. For each range of cluster sizes, the total number of clusters formed are given for various citation frequency thresholds.	73
Table 9. The total number of pages included in the range of cluster sizes using various citation frequency thresholds.	74
Table 10. Examples of the types and sizes of clusters formed by the all-pairs method of cocitation analysis.	75

Table 11. Results of the experiment to determine the effectiveness of the cocitation algorithm. The cocitation algorithm performed significantly better than randomly formed clusters with respect to the “goodness” of the clusters and well as precision.	78
Table 12. Examples of Web pages selected using spreading activation.	87
Table 13. Results of the experiment to determine the effectiveness of the spreading algorithm. The spreading algorithm performed significantly better than randomly formed clusters with respect to the “goodness” of the clusters and well as precision.	89
Table 14. Web aggregation using link topology as outlined in [Botafogo et al 1991].	91

LIST OF FIGURES

Figure	Page
Figure 1. The breakdown of problems using the Web split by age according to GVU's Sixth WWW User Survey. The x-axis labels refer to: the user's connection speed to the Web being too slow (Speed), the ability to find information that is known to be on the Web (Find Known), the ability to organize gathered information (Organize Info), the ability to find already visited pages (Revisit Sites), other problems (Other), the ability to visualize where the user has been and where they can go (Visualize), and feeling lost (Lost). There is very little effect of age on the problems reported.	3
Figure 2. The Cost of Knowledge Characteristic Function. One goal of information interfaces is to maximize user interaction by increasing the amount of accessible knowledge in shorter periods of time.	7
Figure 3. Uninformed search through an un-categorized Web locality would produce a linear information gain function such as U. Categorizing and ranking WWW pages allows an information forager to rapidly identify high value, ranked, categories (HC) and low value categories (LC) and concentrate on exploiting the HC gain curve.	10
Figure 4. Growth rates for total volumes, circulation, and acquisitions for Purdue Library from 1925 through 1965 plotted on a semi-logarithmic scale.	19
Figure 5. The exponential growth of WWW servers since August of 1992, when fifty servers were known to exist. An exponential curve fit is shown to account for most of the variance ($R^2 = 0.993$).	20
Figure 6. An example of the time-course desirability of items on the Web. Certain items on the Web are popular for only a short period of time (Faddish Pages), while others are used regularly for reference purposes (Enduring Reference Pages) or are continually accessed (Persistently Popular Pages).	33
Figure 7. Frequency of Access curve (FOA) for the Georgia Tech Two data set showing that a large number of files are accessed with low frequency.	34
Figure 8. Calculated probability of a document being accessed on Day 8 as a function of the number of times it was accessed in the previous 7 days (for Frequencies < 100).	41
Figure 9. Transformation of plotting log need odds of a document access on Day 8 as a function of the log number of times it was accessed in the previous 7 days (and linear regression).	41

- Figure 10. Probability of a document access on Day 8 as a function of how long it has been since the document was accessed in the previous 7 days. The data are fit by a power function. 44
- Figure 11. Transformation of plotting log need odds of a document access on Day 8 as a function of the log of how long it has been since the document was accessed in the previous 7 days (and linear regression). 45
- Figure 12. Results of plotting log need odds against log frequency for the Georgia Tech Two data set. Items that were accessed less than 1000 times per window were included in the analysis. 48
- Figure 13. The graph displays the cache hit and miss ratios for a replacement policy of one day and the cache miss ratio for a caching policy of that uses a seven day window. The latter policy assumes that all of the past seven days of accesses are kept in the cache. The x-axis represents day to day performance of the caching policy for all the days in the data set excluding weekends. 52
- Figure 14. Diagram of the process of the all-pairs cocitation algorithm. Initially a seed pair is chosen at random, and all other pairs that contain any of the cluster are added element (initially A or B in the case of Cluster One). This process repeats, until all pairs have been added. Another random seed pair is then chosen and the process repeats until all pairs belong to a cluster. 71
- Figure 15. Manually drawn and labeled representation of hierarchical clustering of cocitation matrix. This algorithm was able to identify interesting sets of pages. For example, it was able to identify the two organizational head pages in the collection and distinguish students as a subclass of people. 76
- Figure 16. Conceptual diagram of the mechanism of spreading activation. Initially, activation is directed into a source node. Via the set of associations between nodes, this initial activation is spread through other nodes. The process is then repeated, pumping more activation into the source node and activation flowing between associated nodes until an asymptotic pattern emerges. The nodes with the highest activation represent the nodes with the strongest association to the source node. We use a combination of usage, content, and topology networks for spreading activation through Web pages. 82

SUMMARY

One of the fastest growing sources of information today is the World Wide Web (WWW), having grown from only fifty sources of information in January of 1993 to over a half million four years later. The exponential growth of information within the Web has created an overabundance of information and a poverty of human attention, with users citing the inability to navigate and find relevant information on the Web as one of the biggest problems facing the Web today. The primary goal of the research presented here is to put forth new techniques and models that can be used to help efficiently manage people's attentional processes when dealing with large, unstructured, heterogeneous information environments. The primary model is based upon the desirability of items on the Web. This research searches for lawful patterns of structure, content, and use. Methods are developed to exploit these patterns to organize and optimize users' information foraging and sense-making activities. These enhancements rely on predicting, categorization and allocation of attention. Several methods are explored for inducing categorical structures for the WWW. Some of these enhancements involve clustering in a high-dimensional space of content, use, and structural features. Others derive from cocitation analysis methods used in the study of scientific communities. A user would also be aided by retrieval mechanisms that predicted and returned the most likely needed WWW pages, given that the user is attending to some given page(s). The approach of this research uses a spreading activation mechanism to predict the needed, relevant information, computed using past usage patterns, degree of shared content, and WWW hyperlink structure.

C H A P T E R I

INTRODUCTION

Desirability and Information Ecologies

What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention, and a need to allocate that attention efficiently among the overabundance of information sources that might consume it.

Herbert A. Simon, as quoted by Hal Varian
Scientific American, September 1995, page 200.

The above insightful quote by Herbert Simon assumes the existence of two things: 1) a system that contains a lot information and 2) that users are unable to pay attention to all the relevant information in the system. Granted, one of the fastest growing significant information systems today is the World Wide Web [Berners-Lee et al. 1992][Berners-Lee et al. 1994], having grown from only fifty sources of information¹ in January of 1993 to over a half million four years later [Netcraft 1996][Gray 1994]. The exponential growth of information within the Web has created a lot of information, resulting in over one hundred million accessible documents [Internet Archive 1997]. To help illustrate the enormity of the Web, if a person were to view one document per minute

continuously, it would take them one hundred and nine years or so just to view all the content of the Web as of January 1997, not to mention all the content that will have been produced since then.

It is not surprising then to find users citing difficulties in navigating and finding relevant information on the Web as one of the biggest problems facing the Web today [Pitkow and Kehoe 1996]. As evident in Figure 1 which shows the top complaints of over 15,000 self-selected WWW users, barring speed, users' primary problems focus around the ability to find, organize, and return to information on the Web—all symptoms of an overabundance of information. It is important to note that the speed complaint is both a product of sub-optimal infrastructure and technologies for accessing the Web at the physical layer as well as not getting the right information to the user in the first place. The primary goal of the research presented is to put forth new analytical techniques and models that can be used to help alleviate these complaints by helping efficiently manage people's attentional processes when dealing with large, unstructured, heterogeneous information environments, like the Web. We refer to these environments as *information ecologies* [Pirolli 1991].

But what exactly is an information ecology? According to Webster's New World Dictionary [Webster 1988], an ecology is "the study of the relationship between and the adjustment of human groups to their geographical and social environments." Bridging off

¹ In the case of the World Wide Web and other distributed systems, information sources are commonly referred to as *servers*, with the recipients referred to as *clients*. For the purposes of this dissertation, the terms servers, sites, repositories, localities, and environments will be used interchangeably, though there are subtle differences between the terms.

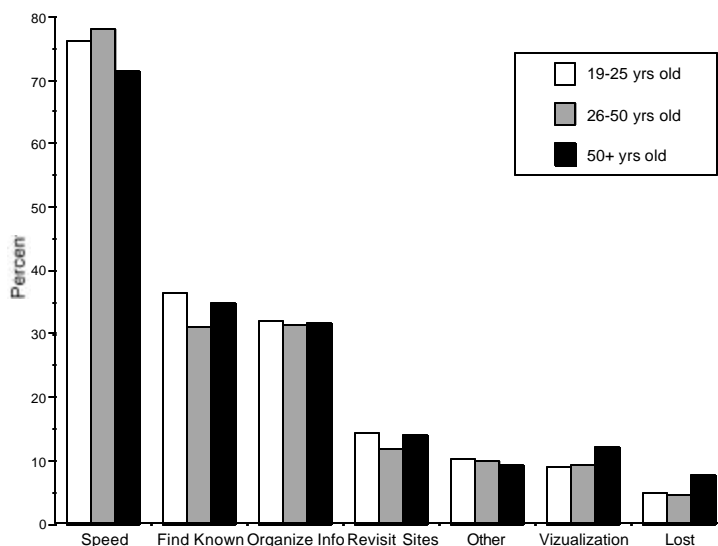


Figure 1. The breakdown of problems using the Web split by age according to GVU's Sixth WWW User Survey. The x-axis labels refer to: the user's connection speed to the Web being too slow (Speed), the ability to find information that is known to be on the Web (Find Known), the ability to organize gathered information (Organize Info), the ability to find already visited pages (Revisit Sites), other problems (Other), the ability to visualize where the user has been and where they can go (Visualize), and feeling lost (Lost). There is very little effect of age on the problems reported.

of this definition, we define an information ecology as “the study of the relationship between the adaptation of human groups to an information environment.” Furthermore, we characterize *dynamic* information ecologies [Pirolli and Card 1995][Pitkow and Recker 1994] as systems that satisfy the following properties:

1. the information in the system expresses a measurable characteristic called *desirability*,
2. the desirability of information varies over time,
3. the information has a life span, that is, it undergoes a birth and eventual death process, and

4. the structure and content changes frequently.

Because individual Web repositories are part of a world-wide, distributed, multimedia information network, the number of documents, their links, and structures are constantly in flux. For example, a Web environment can change internally, as information is supplied by a number of people providing content. A repository can also change externally, as information providers elsewhere on the Internet add and modify pointers to documents located at the Web site. Characterizing this type of chaotic environment is not an easy task.

The notion that information expresses a property called desirability was formally introduced by Quentin Burrell and defined as “the average number of times [an] item is borrowed per unit time” [Burrell 1980 page 118]. With respect to information environments and human attentional processes, we update this term and define desirability as “the likelihood of information receiving attention.” But how does one go about modeling attentional probabilities of information? [Anderson and Milson 1989] outline methods to estimate the likelihood of items receiving attention based upon the item’s previous use (history factor) and the item’s associative strengths with other structures in memory (context factor). The history factor relies upon the frequency and recency of past use to form the estimation, whereas the context factor relies upon the frequency items that have occurred together in the same context, or co-occurred, to form the estimation. As we shall see, a portion of this dissertation develops analytical methods to determine and predict the desirability of items based upon historical and contextual factors.

The first method originates from [Burrell 1980]'s work on library circulation, where the frequency of book circulation was found to be strong predictor of future use. [Burrell 1985] later updated this work to more accurately reflect the data by using a Gamma-Poisson process to predict future circulation patterns as opposed to the original Poisson. [Anderson and Milson 1989] and [Anderson and Schooler 1991] developed a modification of this method, switching from an exponential to a power distribution, to model the frequency and recency characteristics of human memory. [Anderson and Schooler 1991] found that not only does a power law model provide a strong fit to over a hundred years of human forgetting and practice memory data, but the power law model also fits certain environmental sources as well (words in newspaper headlines, email correspondence, and parental speech). That is, for human memory, the more frequently and recently we have been exposed to an item, the more likely we are to remember that item. For environmental sources, the probability of a word appearing in a newspaper headline is based upon how frequency and recently the word has appeared in previous headlines. The same goes for words spoken to children and email correspondence. Given this evidence and the power law distributions behind both human and environmental sources, [Anderson and Schooler 1991] argue that human memory has adapted to the structure of the environment, or at least that the two have co-evolved.

In Chapter Two, the historical notion of desirability is developed in more detail and the power law model is applied and demonstrated to generalize to WWW sites. These findings are shown to hold despite the rather chaotic and diverse nature of WWW sites. Furthermore, it is shown that recency provides a much stronger predictor of future use

(attention) than frequency, which has interesting ramifications on the nature of various Web applications like hit measurement, pre-fetching, caching, etc. Chapter Five contains a more detailed discussion of the applications of these findings.

While the items in information ecologies express the relational property of desirability, attempts to characterize user interactions more completely within these environments is the focus of Information Foraging Theory [Pirolli and Card 1995].

Information Foraging

Foraging theory in anthropology [Smith and Winterhalder 1992] and biology [Stephens and Krebs 1986] attempt to explain how humans and animals optimize their gain of food energy from flux of the physical and organic environment. Similarly, Information Foraging Theory [Pirolli and Card 1995] attempts to explain how human *informavores* optimize their information gain from the flux of the cultural and technological environments. From this perspective, it is useful to conceptualize the WWW as an information ecology in which information-seeking users desire optimized foraging strategies and technologies. The goal of these foraging strategies and technologies is to maximize the amount of information gained from the systems as a function of the costs involved with attaining the information.

Figure 2 presents the Cost of Knowledge Characteristic Function (COKCF) [Card et al. 1993], which graphically represents the cost-benefit relationship between information gain and the costs associated with attainment. From the COKCF it is immediately evident that there are two ways in which one can increase the effectiveness of a user's interaction

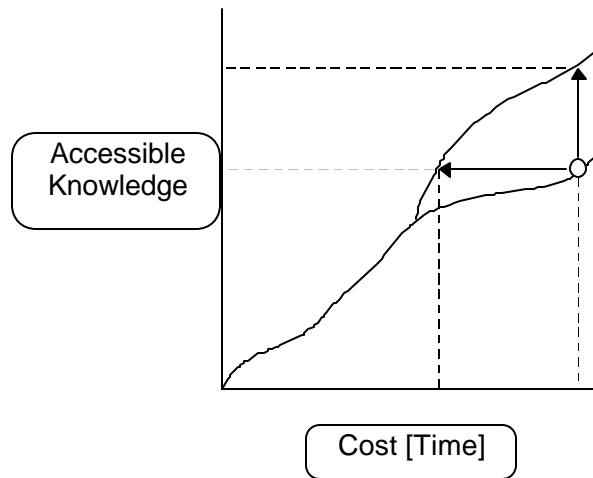


Figure 2. The Cost of Knowledge Characteristic Function. One goal of information interfaces is to maximize user interaction by increasing the amount of accessible knowledge in shorter periods of time.

with an information system (assuming time as the primary measure of the costs incurred in information attainment):

1. increase the amount of information gained in a given amount of time interacting with the system, or
2. decrease the amount of time taken to attain a given amount of information.

Ideally, new strategies and technologies for maximizing information gain will achieve advancements along both dimensions.

Foraging for information in very large hypertext collections like the Web and making sense of 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, unstructured, and heterogeneous. Three 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 categorization techniques are library card catalogs, Table of Contents, and the Yahoo! WWW site [Yahoo 1994]. 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 of such associative retrieval techniques include indices, personal information agents, and search engines such as Lycos [Lycos 1994], 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. The third are systems that attempt to provide visual overviews and analysis of information. Traditional forms of these graphical systems include maps, diagrams, etc. and newer Web related forms like the Navigational View Builder [Mukherjea and Foley 1995] and the Web Forager [Card et al. 1996]. The research being presented attempts to develop both categorization and associative retrieval techniques that can be embedded within graphical interfaces and systems.

Categorizing to Optimize Foraging Decisions

In our efforts to categorize documents on the Web, we assume a scenario in which a user forages for relevant, valuable information at some *Web locality*, meaning some collection of related WWW pages. Perhaps the pages are related because they are at some particular physical site or WWW server, or perhaps related because they have been collected by a particular community or organization. The optimal selection of WWW pages

from the Web locality to satisfy a user's information needs is a kind of *optimal information diet* problem discussed in [Pirolli and Card 1995]. The overall rate of gaining useful information will be improved by eliminating irrelevant or low-value categories of information from consideration. Simply put, to the extent that one can rapidly distinguish junk categories from interesting or relevant ones, a person can allocate time more usefully, thus maximizing their information gain.

Imagine that a user coming upon a Web locality is analogous to a predator such as a lion coming upon an open plain with a teeming array of potential prey species: the optimality of the diet or pursuit sequence chosen by the predator (or user) will depend on the ability to rapidly categorize the prey types (WWW page types), assess their prevalence on the plain (Web locality), assess their profitability (amount of return over cost of pursuit), and decide which categories to pursue and which to ignore. The optimization can be further improved to the extent that the category members can be ranked, so that good examples of a good category could be pursued first. Figure 3 provides a graphical illustration of the improvements provided by categorization and ranking (see [Pirolli and Card 1995] for detailed technical discussion).

Chapter Three will first explore a technique to automatically assign categories to Web pages. Example types include: head pages (the first page to visit in a set of related pages), index pages (pages whose primary purpose is navigation), reference pages (pages that are used repeatedly to explain concepts of content), and content pages (pages that primarily deliver information). Part of the novelty of our approach is that it leverages off of a combination of information to categorize pages, including information about how the

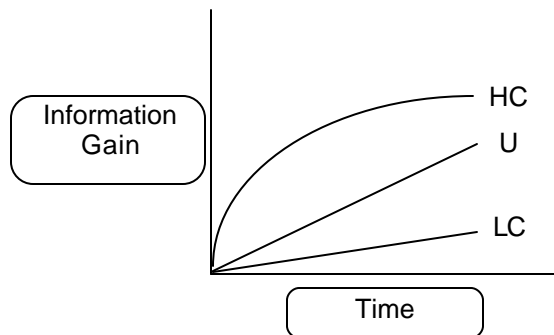


Figure 3. Uninformed search through an un-categorized Web locality would produce a linear information gain function such as U. Categorizing and ranking WWW pages allows an information forager to rapidly identify high value, ranked, categories (HC) and low value categories (LC) and concentrate on exploiting the HC gain curve.

Web is used, the content of the Web, and the Web's topology. We refer to methods that utilize this rich mix of data as *usage+content+structure* techniques. In our approach, we apply a set of human designed linear categorization equations to feature vectors representing usage, the content, and topology data of WWW pages. For each category, a set of weights (+1,0,-1) is assigned to each feature of a WWW page. The solutions to each equation when applied to a list of WWW pages, results in a list of category membership values, with the highest valued documents being more representative of the category than the lower valued documents. This enables the formation of specific categories as well as the ability for documents to belong to multiple categories. This technique is shown to perform well as evaluated by computing precision scores for the top twenty-five pages in each category.

The next section of Chapter Three reviews the application of cocitation analysis to WWW sites to induce semantic clusters of documents. While the above technique performed reasonably well, some of the required data, like usage, is not readily accessible

for the majority of Web sites. Furthermore, the computational complexity of measuring content metrics do not seem particularly suited to scale gracefully to the entire Web. To help alleviate these shortcomings in attempting to structure the entire Web, we use a modified version of cocitation analysis [Garfield 1979] that is based solely upon the hyperlink structure of Web pages. Cocitation analysis builds upon the notion that when a WWW document contains links referring to other documents, say documents A and B, then these documents are related in some manner to each other, even though there may not be a link between documents A and B. In this example, documents A and B are said to be *cocited*.

Two different approaches of cocitation analysis are compared, the first based upon the work of [Small and Griffith 1974] which essentially creates strongly connected components of the most popular cocitation pairs, and the second based upon multidimensional scaling and hierarchical clustering of the most popular cocitation pairs. The latter method is shown to create better formed and more meaningful clusters of documents.

Spreading Activation to Predict Needed Information

[Anderson and Milson 1989] and [Anderson 1990] have recently argued that human memory has adapted through evolution to optimize the retrieval of needed information (memories) based on the current history context of attention. Generally, one could say that their analysis relies on three general sorts of information to compute the *need probabilities* of all stored memories, given a current focus of attention: (1) past usage

patterns, (2) degree of shared content, and (3) inter-memory associative link structures.

These are the historical and contextual factors that were introduced before. More recently, [Anderson 1993] has proposed that spreading activation mechanisms can be used to implement and approximate such computations.

The WWW can be viewed as an external memory and a user-forager would be aided by retrieval mechanisms that predicted and returned the most likely needed WWW pages, given that the user is attending to some given page(s). Chapter Four reviews a kind of spreading activation mechanism [Anderson and Pirolli 1984], another *usage+content+structure* technique, to predict the needed, relevant information. This technique utilizes a combination of past usage patterns, the degree of shared content between pages, and the hyperlink structure of the WWW.

One way to conceptually understand spreading activation is to imagine a system of water reservoirs connected via a set of pipes, with the diameter of the pipes determining the rate of water flowing between reservoirs. When a large amount of water is injected into the system from a particular source reservoir or set of source reservoirs, after a period of time, the water levels in all the reservoirs will settle in a particular pattern. Based upon this final pattern, each reservoir can be inspected and the ones with the most water selected. If one views the flow rates between reservoirs as a measure of their connectedness (association), then the reservoirs with the most water at the end are in a sense the ones more connected (related) to the source reservoir.

In this manner, spreading activation provides a mathematical technique for determining the relatedness of items based upon their degree of association. This technique

produces a variety of related pages based upon the current context of a user in a Web locality. The underlying model enables various weightings to be applied different *usage+content+structure* components, based upon the information requirements of the user. These components, or networks, correspond to the roads most traveled (usage network), the semantic association between pages (content network), and the designer's connection of pages (structure network). For example, if a user is interested in quickly viewing what most of the other visitors at a site have seen, activation can be spread primarily through the usage network and the pages that accumulate the most activation can be returned to the user. Likewise, if a user is interested in a particular page and wants to be made aware of other pages on the same topic, activation can be spread through the content network and the associated pages can be returned. Spreading activation yields flexible, adaptive, and customizable aggregations of documents based upon a combination of different networks. This technique has the added advantage of being computable, under certain conditions, in $O(S(N-1))$ time, where S is the number of sources and N is the number of documents.

Applications

These methods are being investigated partly as potential enhancements to an Information Workspace [Card et al. 1991] that is connected to the WWW. The Web Forager [Card et al. 1996] 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 [Card et al. 1996] 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, like linear combinations of feature vectors and cocitation analysis, could be used to rapidly create structures such as Web Books. Since clusters and categorizes help create abstractions of information environments, it is expected that the implementation of these techniques will help reduce user complaints about finding, organizing, and revisiting information on the Web. 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. This ordering of documents in the workspace can assist users in finding relevant information faster than conventional Web interfaces. The current sluggishness of page retrieval on the World Wide Web could be greatly alleviated by pre-fetching the documents the users is likely to read next, helping further reduce current user speed related complaints. Chapter Five contains a more in depth examination of other possible applications of the techniques developed in this dissertation.

Chapter Outline

This dissertation addresses the top usability complaints of the Web by the empirical characterization of the desirability of items on the WWW. The primary contributions focus on the development of various methods to form meaningful collections of pages and to predict the likelihood of items receiving attention. These are to be considered weak methods, in that they are intended to work reasonably well across a variety of Web localities, as opposed to stronger methods that are intended to work very well but for known content or locality. This research draws heavily upon and expands upon

models and analysis performed by the library science and human memory research communities. It applies existing techniques like the Gamma Poisson Model and cocitation analysis to the new domain of the World Wide Web while discovering new *usage+content+structure* methods by the application of linear combinations and spreading activation. This research has already been presented in one form or another in the following peer reviewed publications:

- Margaret M. Recker and James E. Pitkow. Predicting Document Access in Large Multimedia Repositories. *ACM Transactions on Computer-Human Interaction*. 3(4):352-375, 1996.
- Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a Sow's Ear: Extracting Usable Structures from the Web. *Conference on Human Factors in Computing Systems (CHI 96)*, Vancouver, Canada. April 13–18, 1996.
- James E. Pitkow and Margaret M. Recker. A Simple Yet Robust Caching Algorithm Based on Dynamic Access Patterns. *The Second International World Wide Web Conference*, Chicago. IL. Oct. 20–25, 1994.
- James E. Pitkow and Margaret M. Recker. Integrating Bottom-Up and Top-Down Analysis for Intelligent Hypertext. In *Third International Conference on Intelligent Knowledge Management*, Maryland, MD. Nov. 29–Dec. 2, 1994.
- James E. Pitkow and Colleen M. Kehoe. Emerging Trends in the WWW User Population. *Communications of the ACM*. 39(6), 1996.

- James E. Pitkow. In Search of Reliable Usage Data on the WWW. In *Proceedings of the Sixth International World Wide Web Conference*, Santa Clara, CA, April 7-11, 1997.
- James E. Pitkow and Peter Pirolli. Life, Death, and Lawfulness on the WWW. In *Conference on Human Factors in Computing Systems (CHI 97)*, Atlanta, Georgia, March 22-27, 1997.

Some of the content contained herein is taken directly from some of the above publications, and where this is done it is duly noted. Chapter Two takes a library science/human memory approach to understanding the desirability of items on the Web. The Gamma Poisson Model is applied to several Web environments and recency is shown to be a strong predictor of future use. Chapter Three deals with attempts to create related collections of documents using 1) linear combinations of feature vectors seeded with a diverse set data and 2) cocitation analysis. The use of spreading activation to predict needed items and create groups of related pages is the topic of Chapter Four, which is followed by a discussion of applications and future research in Chapter Five.

CHAPTER II

DESIRABILITY

Perhaps the most suitable existing system with which to compare and understand the WWW is that of libraries. Using the definition of information ecologies in the introduction, libraries and the Web can clearly be viewed as information ecologies. Both are systems that contain dynamic information whose time-course of desirability is quantifiable. The information in both systems also continually undergoes birth and eventual death processes. For libraries, the birth of new information is primarily in the form of acquisitions. As time passes, the desirability of an acquired item fades, the shelf space consumed by the item may be better utilized by a newer, more desirable item. This process, of determining which items to remove and when to remove them is called relegation. For the Web, the act of publishing and conversion of existing material to the Web gives birth to information and although it is not currently widespread practice, staging, archiving and deletion of information can be viewed as progressive stages of the death process.

Just the same, libraries and WWW sites are not without difference. Digital information systems are not subject to the same physical limitations as buildings—disk space and memory continue to become smaller and faster, a property that construction materials like cement do not possess. This physical difference has a direct impact on the costs associated with enlarging a given collection. The availability of items is another

differentiating point. Libraries are primarily “frequency of borrowing” systems, since once an item is checked out from a library, it becomes unavailable for the duration of the loan period. The Web on the other hand, can be viewed as a “frequency of demand” system, it does not suffer from such usage limitations².

While the above discussion reveals some surface level comparisons between libraries and the Web, there exist some fundamental characteristics that both systems share as well, the foremost being exponential growth. Using data from fifty eight libraries over the course of forty years from the 1920s through 1960s [Dunn 1967], [Leimkuhler and Cooper 1971] point out the exponential growth rate of the total number of volumes held by the libraries can be modeled by the following equation:

$$N_t = N_0 e^{-at}$$

where N_t represents the size of the collection t years ago relative to the current size, N_0 , and a is the annual acquisition rate. Naturally, the acquisition rate a is library specific and changes in a will result in exponential shifts of the growth patterns.

Figure 4 shows the growth rate of Purdue libraries over this forty year period, revealing a nearly tenfold increase in circulation. The exponential growth rate was confirmed in other library systems by [Atkinson 1976] and the University of Pittsburgh study [Kent et al.

² This of course is in lieu of situations where items on the Web become temporarily unavailable, e.g., server maintenance, network failures, etc.

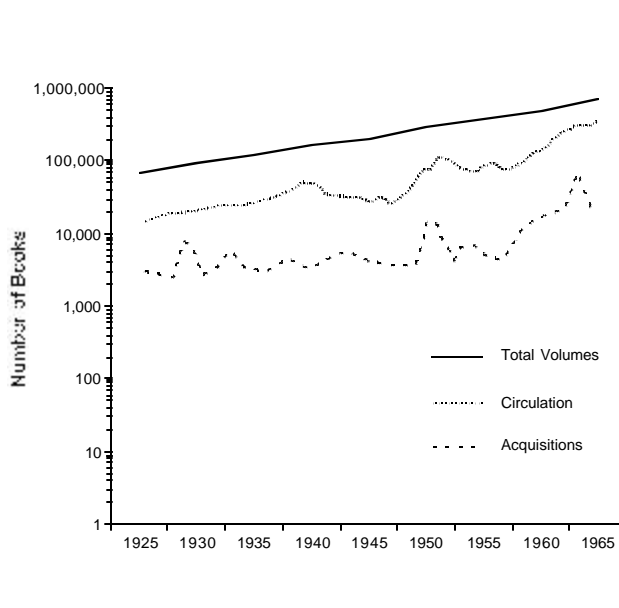


Figure 4. Growth rates for total volumes, circulation, and acquisitions for Purdue Library from 1925 through 1965 plotted on a semi-logarithmic scale.

1979]. Similar exponential growth has also been noted by [Leimkuhler and Cooper 1971] for the acquisition of items and the total circulation of items over time as seen in the semi-logarithmic plot in Figure 4.

Figure 5 shows the exponential growth rate of information servers on the WWW. In five years, the number of Web servers has grown from around fifty in August of 1992 to over 750,000 by January of 1997. If one uses the estimate of there being around 100 million documents on the Web [Internet Archive 1997], this equates to an average of 100 documents per Web server—a number that intuitively seems reasonable. The Web certainly demonstrates a faster rate of growth, but both systems exhibit exponential growth.

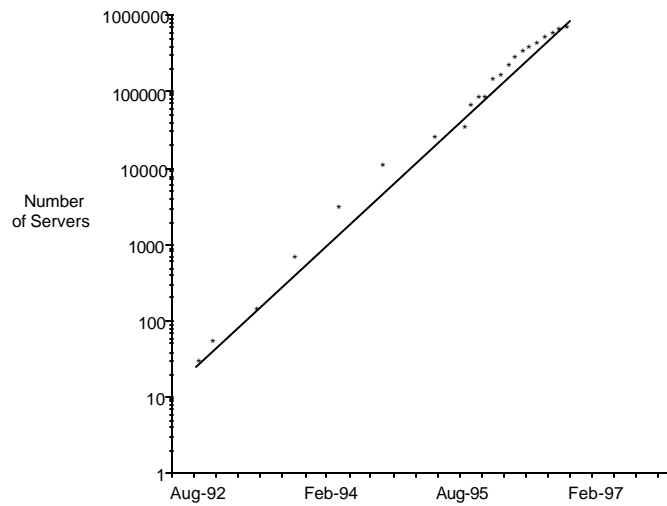


Figure 5. The exponential growth of WWW servers since August of 1992, when fifty servers were known to exist. An exponential curve fit is shown to account for most of the variance ($R^2 = 0.993$).

Another interesting property of libraries is frequency of circulation (FOC). When plotted on a logarithmic y-axis, the number of items that have circulated once, twice, three times, etc. during a fixed period of time form more or less a straight line. The semi-logarithmic frequency of circulation (FOC) property has been so widely observed that it has become known as *the straight-line phenomenon*.

Both libraries and the WWW have been motivated by goal of the Library of Alexandria—of becoming a repository for all human knowledge [Gore 1976]. During the first half of the twentieth century, academic and public libraries increased their holdings at exponential rates, fueled in large part by this notion of Alexandria. However, it became clear by the mid 1960's that this rate of was not sustainable due to the tremendous building costs associated with the need to accommodate additional acquisitions. It is not surprising

to find that the concept of self-renewing libraries became popular. Originally introduced by the University Grants committee in 1976 [University Grants Committee 1976], self-renewing libraries were defined as: a library of limited size in which beyond a certain point material should be reduced at a rate related to the rate of acquisition. The question was no longer how to raise enough money to build more physical space to hold more books, but rather what policies should be adopted to help facilitate the acquisition and relegation of materials. To this day, policies in this area of library management are widely debated.

In order to achieve no-growth, one needs to be able to make decisions as to which items to keep on the stacks and which to relegate. As pointed out in [Burrell 1980], several criteria exist, including:

1. subjective assessment by librarians,
2. recommendations by academic staff (peer review),
3. recorded usage of individual items,
4. last use of items,
5. publication date, and
6. accession date

with some modification of all the above being common practice. While all of the above techniques can facilitate decisions as to which books to remove, they all suffer from same problem—none are able to predict the costs of such decisions. As a result, certain books will be relegated that will be requested, while other books will remain on the shelves that will not be used. A certain degree of this inefficiency is bound to occur regardless of what method is used to determine relegation policies. Still, methods that are based upon

empirical regularities of actual circulation can estimate the costs involved with each decision. The success of the decisions will rely upon the model's ability to correctly match real circulation patterns. One of the most widely known model for library circulation and relegation decisions is the simple model set forth by [Burrell 1980], which will be discussed in the next section.

The Web is becoming a global repository for human knowledge. Some entities, like the Internet Archive [Internet Archive 1997], feel that everything digital ought to be forever preserved. One of the exciting, though possibly disturbing properties of the Web, is that the limiting physical costs that resulted in the introduction of self-renewing libraries are simply not present for the Web. Computer memory, processor speeds, and disk are readily accessible at ever decreasing costs. In fact, current figures by the [Internet Archive 1997] estimate that the entire Web consists of 300 gigabytes of textual content and an extra 3 terabytes of multimedia content. With disk costing roughly \$500/gigabyte (US) at the end of 1996, the entire Web can reside on local disk for under half a million dollars. Given the apparent absence of physical cost limitations, what will be the limiting factor to the exponential growth of information on the Web?

We believe that the limiting factor for the continued growth in the size of the Web is human attention. This attentional deficit is due in part to the effect of diminishing returns information creators will face as well as the fixed capacity of humans to process information. Since the total amount of human attention is a fixed quantity, as more and more information becomes available on the Web, the likelihood of newly produced content on the Web receiving attention decreases. This is expected to be the case even as

specialized communities of interest develop, i.e., communities only interested in a particular topic. This is also expected to be the case even as new technologies are developed that help filter/route information to users. Simply put, as the same amount of human attention is spread over more and more newly produced information, the ability to get the right information decreases in tandem with the attractiveness of producing new information. Furthermore, Web complain about the inability to find, relocate, and manage the amount of information on the Web today [Pitkow and Kehoe 1996]. If the Library of Alexandria becomes the model for the Web and all documents, editorials, images, personal home pages, list of links, etc. are stored for eternity, the problem now becomes not only finding the right information, but possibly finding the right version of the information. These are not simple problems with immediate and obvious answers. As the Web continues to grow these problems can only be expected to continue to get worse³.

One solution consists of the development of specialized tools and technologies that assist users in determining what information is worth attention. That is, users can benefit greatly from technologies that help predict what information they are likely to need or find useful. They can also benefit from information being ordered according to predicted interest as well. This essentially involves the development of techniques that predict the future desirability of items. Predictions of this sort can be done on an individual, group, or entire population basis. We call this process *promotion*, since it has the effect of promoting certain items based upon their desirabilities.

³ As with most environments, some form of steady-state dynamics will most likely emerge, balancing production with

It should be immediately clear that predicting what users will want (promotion) is the logical inverse of predicting what they will not want (relegation). If one assumes an ordering of items according to their desirabilities, then promotion and relegation occur at opposite sides of the continuum. The desirability of an item may be evaluated across multiple dimensions. For the research presented here though, a uni-dimensional ordering of the desirability of items is assumed in a step towards simplification.

This assumption does not however specify the underlying distribution of items according to their desirability. By borrowing a model of desirability initially designed for library circulation [Burrell 1980][Burrell 1985] and modified to account for items in human memory [Anderson and Milson 1989][Anderson and Schooler 1991], a model of the distribution of desirability for items on the WWW will be developed and evaluated. Promotion and relegation decisions that determine the cost associated with these decisions can be made against the model. Additionally, the model can be used to facilitate comparisons between different WWW ecologies.

The Gamma-Poisson Desirability Model

Clearly any model that helps practicing librarians manage their resources provides some utility. But in order for any model to be of real use to practicing librarians, [Sandison 1977] asserts that mathematical models of library circulation should: 1) be based on valid assumptions, 2) be explained in sufficiently simple terms for the ordinary librarian to carry

consumption yielding acceptable gains for both information producers and consumers.

out and 3) result in better advice than that obtainable by simpler techniques. Similarly, [Burrell 1980][Burrell and Cane 1982] state that the primary aim of library circulation models ought to be 1) to contain only a small number of parameters to characterize the library whose a) meaning is easy to understand and b) are not too difficult to estimate and 2) provide a qualitatively good fit for the data from various libraries over various time periods. With this in mind, [Burrell 1980] proposed a simple mathematical model in an attempt to explain the straight-line phenomenon of academic library loans.

For this model, every book in the library is assumed to possess a certain degree of desirability. As stated in the Introduction, Burrell defines desirability as the number of times that a book is borrowed⁴ during a fixed period of time T . In order to minimize cyclic variations in borrowing patterns that result from academic calendars while remaining sensitive to the effects of aging, Burrell uses a period of one year for T . When considering a fixed collection of books of size N , [Burrell 1980] shows mathematically that the straight-line phenomenon can be modeled reasonably well using a geometric distribution whose mean is T/a , where a is a collection specific constant. The constant a may be interpreted as the average time between borrowings for any randomly chosen book, or alternatively as the amount of time required for the average number of books borrowed to equal one.

For each book with desirability, λ , Burrell assumes that the number of times a book is borrowed occurs as a stochastic process, with different books having different

⁴ The choice of circulation rate for the desirability metric by Burrell is certainly appropriate given the primary aim of creating a simplistic model. Other forms of desirability that could have also been included, though at the cost of simplicity, include the number of copies per book, the number of unique borrowers, etc.

borrowing frequencies, or desirabilities. The probability then that a given book will be borrowed k times during a period T can be given by the following Poisson distribution:

$$P(k; \mathbf{I}) = \frac{e^{-IT} (IT)^k}{k!} \quad k = 0, 1, 2, 3, \dots$$

with the desirabilities for different items varying as a negative binomial distribution with parameter \mathbf{a} . That is, for a book with desirability λ , the time between uses is an exponential process with $1/\lambda$ representing the mean time between uses. The probability that a book will be loaned out r times during a time T is:

$$P_r(T) = \left[\left(1 - \frac{T}{T + \mathbf{a}} \right) \left(\frac{T}{T + \mathbf{a}} \right) \right]^r$$

which is a geometric distribution with parameter $T/(T + \mathbf{a})$. From this it can then be shown that the probability that a randomly chosen item will be borrowed a certain number of times during a fixed period is equal to the probability generating function for the geometric distribution of the straight-line phenomenon (the interested reader is referred to [Burrell 1980] for the mathematical proof).

In fitting the model to various data sets, [Burrell 1980] noticed that the model predicted far fewer zero circulation items. This lack of congruence between the observed data and the model can be explained by challenging the assumption that all items in a particular library are potential candidates for circulation. In any large collection, there will be items, e.g., items on permanent reserve, journals, special collections, missing items, etc., that will not be included into circulation statistics. This class of items is different than those items which could have been but were not circulated. The problem is that the exact number of these “dead” items in a collection is not determinable, thus it must be estimated. [Burrell 1980] shows that the number of dead items, β , can be estimated and incorporated into the model with only prior knowledge of the total number of items in the collection, the total number of items which have been borrowed during the period, and the total number of borrowings recorded during the period.

The subsequent decrease in the desirability of items, or aging, complicates the [Burrell 1980] model, since the original model assumes that the circulation rate for items remains constant over time. To handle aging, [Burrell 1985] assumes that items suffer exponential decay in their desirabilities, with equal probability of decay for each item in the collection. These assumptions are modeled as a non-homogenous Poisson process whose rate varies as $e^{at} \lambda$ at time t with constant a , where $t > 0$, $a > 0$, and λ is the initial desirability at $t=0$. The probability that an item will be circulated r times during the first n years is:

$$P(X_n=r) = (r+v-1)P_n^v q_n^r; r = 0, 1, 2, 3, \dots$$

where:

$$P_n^v = (1 + e/a (1-q^n))^{-1} \text{ and } q_n^r = 1 - P_n$$

The model parameters for the mixing gamma distribution function are ε and v and $\theta = e^{-a}$.

The aging factor is equal to $1 - \theta$ and represents the annual rate of decay. This model enables the future use of items to be predicted more accurately than the previous model.

In summary, Burrell proposes a circulation model, commonly referred to as the Gamma-Poisson Model (GPM), in which the number of borrowings for items occurs as a non-homogenous Poisson process with different items having different desirabilities that vary according to a negative exponential distribution with parameter a . The desirability of items over time decreases at an exponential rate and all items in the collection decay at the same rate. Dead items, which were found to falsely inflate the zero class of items, are incorporated into the analysis using the model parameter β . This model has been found to produce a reasonable fit to the circulation data from several academic libraries.

It is important to note that in keeping with the primary aim of a simplistic minimal parameter model, that this model utilizes only a few parameters to fit the data. Others have suggested different modeling distributions that provide more suitable fits to the data, the most notable being Morse's Markov model [Morse 1968]. [Tague and Ajiferuke 1987] go so far as to claim that neither the Morse's Markov model nor the GPM

provide statistically valid fits to the data. In rebuttal, [Burrell and Fenton 1994] show that if the GPM is modified to take into account loan period, the model produces more precise predictions. They further point out that the purpose of the GPM is to facilitate decisions, not necessarily to produce an exact fit to the data.

Anderson's Modification

With the intention of providing an explanation of the relationships between the recency, frequency, and spacing of events on human memory, [Anderson and Milson 1989] proposed using a Bayesian estimate of Burrell's mathematical notion of desirability. Using an information-retrieval perspective [Anderson and Milson 1989][Anderson and Schooler 1991] argue that one can view the desirability of an item in a memory system as an expression of the likelihood that the item will be needed now. The expected usage of such an item is called the *need probability*. The role of a rationally designed information system can be viewed as solving an optimization problem, where the costs involved with storing an item is balanced by the item's need probability. Such a system retrieves items ordered by their expected desirabilities, with the system halting the retrieval of items when the expected gain for retrieving items falls below the costs associated with retrieving the items. That is, if p denotes the need probability of an item with cost C associated with the retrieval of that item, then based on the expected gain G from the successful retrieval of the item, the system should stop when $C > pG$ [Anderson and Milson 1989].

The major contribution of [Anderson and Milson 1989] lies both in the formation of a mathematical model for the rational analysis of human memory and the discovery that

this form of analysis does a remarkable job at predicting the effects of recency, frequency, and spacing on human memory. It is important to note though that this explanation of human memory does not imply the engagement of human memory in statistical computations, only that the behavior of human memory parallels this form of statistical inference.

One of the shortcomings of [Anderson and Milson 1989] and other models of human memory, is that they do not provide a mechanistic explanation of the practice, retention, and spacing phenomenon. In an attempt to provide an explanation, [Anderson and Schooler 1991] argue that memory has adapted to the structure of the environment. Contrary to the popular belief that human memory functions follow an exponential form, [Anderson and Schooler 1991] show that a power function provides a better fit for both human memory functions and environmental sources. Specifically, they examine the relationship between the recency, frequency, and spacing of various environmental sources and show that the same power law relationships that are characteristic of human memory functions also exist for these environmental sources. These sources included the frequency, recency, and spacing of words in the New York Times headlines, words in parental speech, and electronic mail correspondence. That is, the more frequently and recently one is exposed to an item, the more likely one is to retain the memory of that item. The distribution of probabilities for frequency and recency across all items follows a power function. For environmental sources, if one analyses how often and recently words in the New York Times headlines appear, the likelihood that a word will appear the next day also follows a power function. This power function is found in words spoken by parents infants

during their first years of development as well as electronic correspondence between people.

A power function has the form:

$$P = At^{-b}$$

where A and b are constants and T represents a period of time. In the case of retention, b can be thought of as the forgetting rate; for practice, b can be viewed as the learning rate. Under logarithmic transform, the above equation expresses a linear relationship of the form:

$$\log P = \log A - b \log T$$

With respect to the power law model for human retention and the recency of headlines in the New York Times headlines, a strong linear relationship exists for both systems. If we assume the Web to be just another generalized information system then it should exhibit the same characteristics, i.e., power law relationships for frequency and recency, as these other information systems. The next section reviews the WWW data sets

Table 1. Summary of the data sets used in the desirability analysis. For each data set, the dates used in the analysis is given along with the total number of requests, range of requests, the average number of requests per day and the total number of requests. The attentional growth rate is the percentage of documents that were accessed for the first time during the collection period.

Data set	Collection period	Total number of requests	Range of requests /file	Mean requests /day	Total number of items	Attentional growth rate
Georgia Tech One	1/1/1994 - 3/31/1994	305,000	300 - 12,000	3,379	2,000	2.32%
Georgia Tech Two	3/1/1996 - 5/31/1996	4,175,582	1 - 86,728	46,395	42,697	1.10%
Xerox PARC One	4/1/1996 - 6/30/1996	696,126	1 - 39,523	7,649	4,303	1.08%

and method by which frequency and recency calculations were performed followed by a presentation of the results.

Application to WWW Information Ecologies⁵

The analysis used the log files from several WWW repositories across several years. The initial data set, Georgia Tech One (from the site: <http://www.gatech.edu>), consists of accesses during a three month period, January 1 through March 31, 1994. From the log file, we removed all accesses made by Georgia Tech machines. We felt that these accesses added noise to the data because they often represent users testing the presentation of new documents during the authoring cycle⁶, or were the result of default document accesses made upon initial execution of client Web browsers. The second data set, Georgia

⁵ Portions of the following material first appeared in [Recker and Pitkow 1996].

⁶ Recall that in early 1994 there were no integrated, direct manipulation, presentation-based HTML authoring tools, so authors typically created HTML documents by iterating between making edits and viewing the changes to the document in a Web browser. A very inefficient process to say the least.

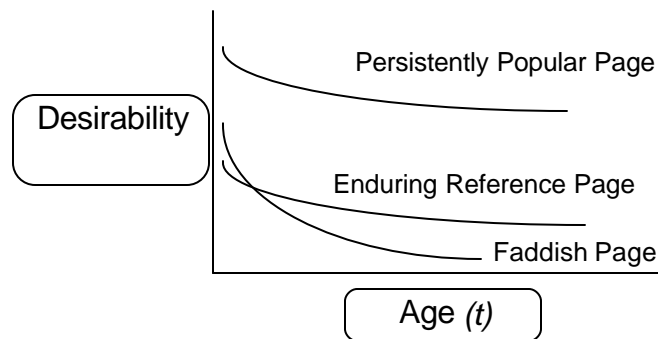


Figure 6. An example of the time-course desirability of items on the Web. Certain items on the Web are popular for only a short period of time (Faddish Pages), while others are used regularly for reference purposes (Enduring Reference Pages) or are continually accessed (Persistently Popular Pages).

Tech Two (from the site: <http://www.cc.gatech.edu>), consists of accesses during a three month period, March 1 through May 31, 1996. All internal accesses were removed as well. The final data set, Xerox PARC One (from <http://www.parc.xerox.com>), contains accesses to the external Xerox PARC OneWeb site during the period April 1, 1996 to June 30, 1996. Xerox PARC One is a hybrid research and corporate Web site with controlled content.

The trimmed log file from Georgia Tech One comprised 35 megabytes of data, with a mean record length of 100 bytes and totaling roughly 305,000 requests. The number of requests ranged from 300 to 12,000 documents per day, with a mean of 3,379 accesses per day over the three month period. Some individual documents were accessed up to 40,000 times per week (which represented a high volume Web server at that time). At the time of our initial analysis, Georgia Tech One contained over 2,000 multimedia documents. Documents on the server embedded many different forms of media, including text, postscript, GIF, jpeg, mpeg, audio, and CGI script requests. Table 1 summarizes the

characteristics for Georgia Tech One as well as the other data sets. These other data sets are quite different from Georgia Tech One in the range of document accesses, number of documents, and their structure.

These repositories exhibit certain irregularities. For example, accesses to repositories are subject to fluctuations, which result, in part, from reduced weekend activity. However, Web ecologies also contains many temporally-dependent documents. In these cases, the content of documents may change, while the document names do not. The desirability of items within ecologies also changes over time. Conceptually, the desirability of faddish WWW pages, long-lasting popular pages, and occasionally used but enduring reference material is presented in Figure 6.

There are many other types of irregularities in the way documents are accessed.

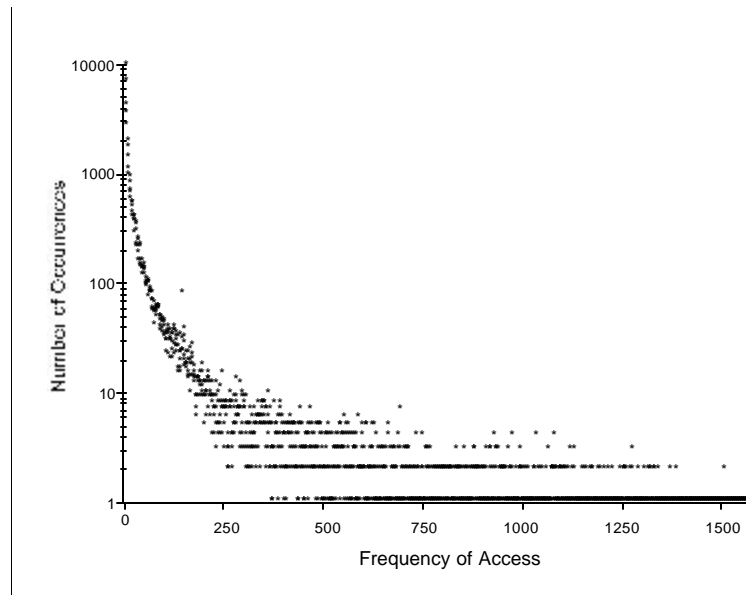


Figure 7. Frequency of Access curve (FOA) for the Georgia Tech Two data set showing that a large number of files are accessed with low frequency.

These include, for example, client-side caching of documents and images, proxy-caches, CGI script requests (where each request may result in a unique document request), non-graphical clients (which do not request images and movies embedded in documents), document creation, deletion, and renaming. Finally, the number of documents and their links continues to change and grow. In Georgia Tech One, 2.3% new documents were requested per day over the course of the collection period. While this is not a precise measure of the number of new documents authored per day—some authored documents are not requested—it does give some indication into the growth rate of each locality.

It is our belief that these irregularities, which exist both in the dynamic nature of document structure and in their requests, are fundamental characteristics of dynamic information ecologies. Furthermore, it seems likely that repositories with these characteristics will become increasingly common. It therefore becomes important to develop robust techniques and methods for predicting and modeling information seeking patterns in such ecologies.

Given the success of the GPM and the power law function at modeling the characteristics of other information systems, we hypothesized that WWW sites would exhibit similar characteristics. The applicability of the GPM 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. Figure 7 shows that the FOA curve for Georgia Tech Two follows a negative binomial distribution. The FOA curves for the other data sets follow similar distributions.

Table 2. The table shows the accesses of documents A, B, C, and D to an HTTP server across a nine day period. Cells with two letters indicates that the document was requested twice. In this example, the window size is seven days and the pane size is one day, resulting in two windows and two panes for the nine day period.

Document	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8	Day 9
A	—	—	AA	—	AA	A	A	A	A
B	B	—	B	BB	B	—	B	—	—
C	C	C	—	C	C	C	—	CC	—
D	—	D	—	D	DD	D	—	D	—
window 1								Pane 1	
window 2									Pane 2

Frequency Analysis

In our analysis, we applied the algorithm developed by [Anderson and Schooler 1991]. The algorithm involves calculating the relationship between the frequency of term use during a period of time (the *practice period*) and whether or not that term occurs during a subsequent period of time (the *test*). Thus, each appearance of a term in the practice period can be viewed as placing a demand upon memory to retrieve information about the desired memory item. The need probability of the item is the likelihood of that item appearing during the test period.

For our data set, we were interested in determining the relationship between the number of document requests during a period of days (called the *window*) and the probability of access on a subsequent day (called the *pane*). That is, given the frequency of past document requests during a specified window, we were interested in determining the

probability of a new request during the pane. Specifically, we computed the frequency of accesses for all requested documents during each seven-day window in the data set, aggregated documents with the same frequency into the same bin, and determined whether or not the documents in each bin were accessed during the pane. By aggregating over all documents with the same frequency of access, we estimated the probability that a document accessed with a given frequency will be needed on the next day. This measure is the need probability. We selected a window of seven days and a pane of one day to encompass the typical fluctuations and access patterns inherent in the calendar week, though we did not confirm this selection empirically.

We illustrate the algorithm for computing the frequencies of document requests and their need probabilities using an example of a set of document requests to a server. Table 2 shows a series of document requests for documents A, B, C, and D during a nine day period (cell entries with multiple letters denote multiple document requests). The bottom of the table shows the time periods for the first two windows and panes. From Table 2 we see that on Day 4, document B was accessed twice, documents C and D were accessed once, and document A was not accessed. While this is a small example, one can imagine constructing a table with all accesses to all documents over all days in the sample period. Note that for any given period and window size, there will be $(\text{number of days} - \text{window size})$ windows and panes in the analysis. In the above example, there are $(9 - 7) = 2$ windows and panes.

The algorithm for tabulating frequencies and calculating need probabilities is as follows:

- For Window 1 (day 1 through 7), group documents with a the same frequency of access values into bins. For example, we find that documents A and B are accessed 6 times and documents C and D are accessed 5 times. Documents A and B are therefore grouped into the bin called “Frequency 6” and documents C and D are grouped in the “Frequency 5” bin (see the second column of Table 2). Theoretically, if each document were requested a unique number of times during the window, the number of non-empty frequency bins would equal the number of documents, though this rarely occurs.
- For every document in each frequency bin, determine if the document is accessed at least once in Pane 1 (day 8). If the document is accessed, denote this as a one (Boolean for true), otherwise, denote this as zero (Boolean for false). Sum the Boolean values for all documents in the frequency bin. For example, for “Frequency 6” we find that A is accessed at least once but B is not accessed (the third column Table 2). Therefore, the number of documents with at least one access in Pane 1 for “Frequency 6” is 1 (1+0) (the “Needed” column of Table 2). For the “Frequency 5” bin, we find that documents C and D were both accessed at least once, resulting in a value of 2 (1+1)
- For each frequency bin, compute the probability of access by dividing the number of documents with at least one access in Pane 1 by the total number of documents in the bin. In the above example, the probability of access for “Frequency 6” is the number of documents with at least one access in Pane 1 (1+0) divided by the total number of documents in the “Frequency 6” bin (1+1), or .50. For the “Frequency 5”

Table 3. For Window 1 and Window 2, the table shows how documents with different access frequencies are grouped into different frequency bins, whether or not the document was requested at least once during the pane, and the summations of each frequency bin. The need probability is computed by dividing the summation of whether the documents in each bin were needed during the panes by the total number of documents in each bin.

Window 1	Document Name	Access Pane 1?	Needed?	Probability
Frequency 6	Doc A	Yes	1	
	Doc B	No	0	
Sum	2		1	0.50
Frequency 5	Doc C	Yes	1	
	Doc D	Yes	1	
Sum	2		2	1.00

Window 2	Document Name	Access Pane 2?	Needed?	Probability
Frequency 7	Doc A	Yes	1	
Sum Window 1 & Window 2	1		1	1.00
Frequency 6	Doc C	No	0	
	Doc D	No	0	
Sum Window 1 & Window 2	4		1	0.25
Frequency 5	Doc B	No	0	
Sum Window 1 & Window 2	3		2	0.66

bin, the number of documents accessed at least once during Pane 1 (1+1) divided by the total number of documents in the bin (1+1) is 1.0. Thus, the probability that a document accessed five times during the first seven-day window will be accessed during Pane 1 is 1.0 (the last column of Table 2).

- Iterate, using Window 2 (day 2 through day 8) and Pane 2 (day 9), accumulating the number of documents and number of accesses per bin. Continue iterating through Windows and Panes over the entire data set by incrementing the start date of each successive window by one.

We continue our example using Window 2 and Pane 2. In Window 2 (day 2 through day 8) we find that documents C and D are accessed 6 times, though neither document is accessed on Pane 2 (day 9). Thus, for two iterations, our new probability of access for “Frequency 6” is the total number of documents with at least one access in their panes (1+0+0+0) divided by the total number of documents in the two windows with an access frequency of 6 (1+1+1+1), or .25. That is, documents that were accessed six times within the 2 seven-day windows have a 25% chance of being accessed the next day. For the “Frequency 5” bin, we see that document B was requested five times during the window but was not requested during Pane 2. Thus, the new probability of accesses for “Frequency 5” bin is 0.66: 2 requests (during Pane 1 and Pane 2) divided by 3 documents in the “Frequency 5” bin (during Window 1 and Window 2). Note also that a new bin is created for 7 accesses because document A was requested 7 times during Window 2. Since Document A was indeed requested during Pane 2, the probability of access for the “Frequency 7” bin is 1.0.

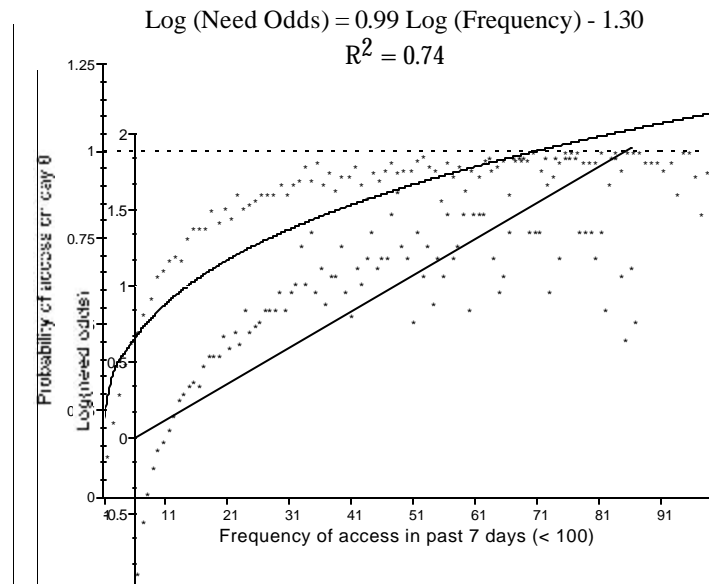


Figure 8. Calculated probability of a document being accessed on Day 8 as a function of the number of times it was accessed in the previous 7 days (for Frequencies < 100).

Figure 9. Transformation of plotting log need odds of a document access on Day 8 as a function of the log number of times it was accessed in the previous 7 days (and linear regression).

Results

Using the algorithm described above, a window size of seven days, and a pane size of one day, we computed the probabilities of access for all documents based on frequency of access during each previous window in the data set.

Figure 8 shows the relationship between the frequency of document accesses during the past seven days (for frequencies < 100) and the probability that it will be accessed on the current day. Following the approach of [Anderson and Schooler 1991], the

data are fit with a power function. As the frequency of document accesses increases, the probability of access on the current day approximates a power function.

For frequency of accesses greater than 100 per window, we observed increased variability in predicting the probability of access. A similar systematic deviation in the correlation between high frequencies and probabilities was noted in [Anderson and Schooler 1991] for the New York Times Headlines and electronic mail addresses. This effect was dismissed by [Anderson and Schooler 1991] as the number of items affected was small and such items represent extremes not reproducible in traditional memory experiments. In our case, the number of items is also small, but decreases the strength of this model since it does not take into account all data. Later, we present an approach that accounts for high frequency rates of access.

As proposed by [Anderson and Schooler 1991], a more interesting comparison is to plot the log of frequencies of access against “log(Need Odds).” Recall that if p is the probability then

$$\text{Need Odds} = p/(1-p)$$

Figure 9 plots the relationship between need probability and frequency, with a logarithmic transform. In this figure, the linear relationship suggested by the desirability

model provides an approximate fit between “Log(Need Odds)” and the log of the frequencies. The regression equation for the linear fit is:

$$\text{Log}(\text{Need Odds}) = 0.99 \text{Log}(\text{Frequency}) - 1.30$$

which accounts for 72% of the variance. Other mathematical models may provide a better fit.

Recency Analysis

We also analyzed the recency of accesses for documents during a 7-day window and measured their probability of access in a 1-day pane. This analysis parallels the retention function in human memory research. Recency probabilities are computed like the frequency probabilities, using the approach of [Anderson and Schooler 1991]. Instead of forming bins with similar frequency values, documents are aggregated into bins of similar recency values for each window. Thus, we compute how many days have elapsed since each document was last requested in a window, and determine whether or not that document was accessed during the next pane. The latter computation is called the document's need probability.

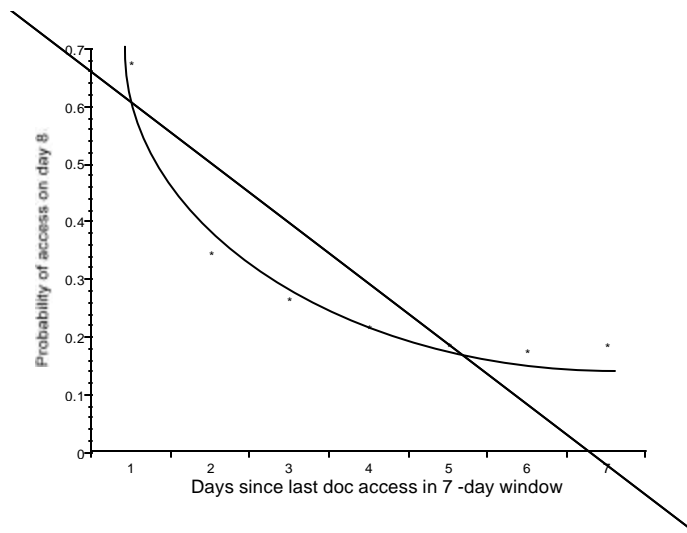


Figure 10. Probability of a document access on Day 8 as a function of how long it has been since the document was accessed in the previous 7 days. The data are fit by a power function.

We illustrate how the recency values and probabilities are computed with the example used for frequency calculations in Table 2. The algorithm used for recency is the same algorithm used for frequency computations, but with different criteria for bin packing. For documents last accessed one day ago during Window 1 (day 1 through 7), we observe that both documents A and B were last accessed during day 7, yielding a recency value of 1. On Pane 1 (day 8), document A is requested at least once but document B is not. The need probability for the “Recency 1” bin is thus the sum of at least one access during Pane 1 (1+0) divided by the total number of documents in the “Recency 1” bin (1+1), or 0.50 $((1+0)/2)$. For a recency of value 2, during Window 1, we find that documents C and D are most recently accessed two days in the past from Pane 1 (on Day 6). We then look in Pane 1 to determine if documents C and D are accessed at least once

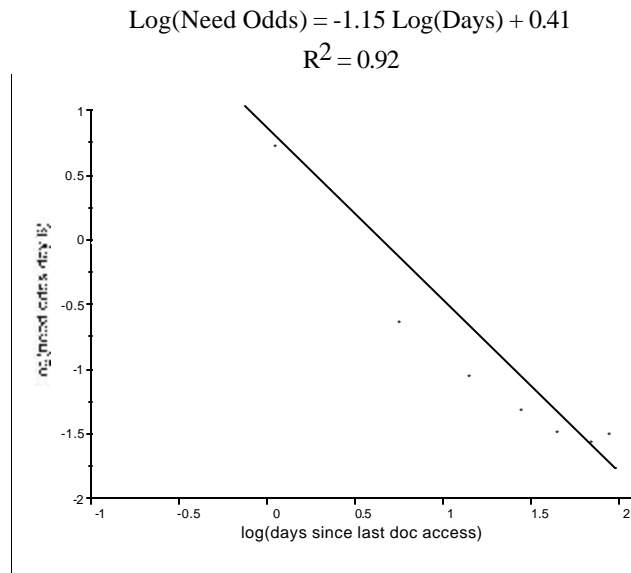


Figure 11. Transformation of plotting log need odds of a document access on Day 8 as a function of the log of how long it has been since the document was accessed in the previous 7 days (and linear regression).

and find that both C and D are accessed at least once. Thus, for the “Recency 2” bin in Window 1, the probability of access equals the sum of occurrences of accesses in the pane (1+1) divided by the number of documents in Window 1 with an access recency of 2 (1+1), or 1.00 ((1+1)/2). As with the frequency computation, probabilities are computed for all recency values, for all windows and panes in the data set by incrementing the start date of each successive window by one. Note that the number of recency bins for each window has as an upper bound the size of the window.

Results

Figure 10 plots the probability of access on day 8 against how many days have passed since the document was last accessed in the previous 7-day window⁷. The plot shows the steep negative slope typically found in retention plots in memory research. As days elapse since the last access in the window, the document is much less likely to be accessed in the next pane. Figure 11 plots the relationship between need probability on day 8 and recency of document access, with a logarithmic transform. As can be seen, there is a strong relationship between “log(Need Odds)” and “log(days)”. The regression equation is

$$\text{Log}(\text{Need Odds}) = -1.15 \text{Log}(\text{Days}) + 0.41$$

and accounts for 92% of the variance.

In summary, the recency analysis shows the logarithmic relationship typically found in the retention memory literature. In addition, recency shows a much better fit than frequency.

⁷ We have defined Day 1 to mean the document was accessed one day previous to the pane (or the last day of the window), in contrast to [Anderson and Schooler 1991] who used Day 0.

Table 4. The results of a power law model to frequency and recency data from three data sets. In all cases, recency provides a better predictor of future access than frequency.

Data Set	Frequency Regression Equation	R²	Recency Regression Equation	R²
Georgia Tech One	Log(Need Odds) = 0.99 Log(Frequency) - 1.30	0.76	Log(Need Odds) = -1.15 Log(Days) + 0.41	0.92
Georgia Tech Two	Log(Need Odds) = 0.91 Log(Frequency) - 2.39	0.85	Log(Need Odds) = -0.57 Log(Days) + 0.14	0.91
Xerox PARC One	Log(Need Odds) = 0.88 Log(Frequency) - 2.47	0.78	Log(Need Odds) = -0.83 Log(Days) + 0.08	0.97

Results from Other Data Sets

In order to show that these findings are specific to a particular Web locality at a given point in time, the technique was applied to the other data sets, Georgia Tech Two and Xerox PARC One. Table 4 shows that for all data sets, recency is a better predictor of future access than frequency. Figure 12 shows the logarithmic transform of need odds versus frequency for the Georgia Tech Two data set. For this analysis, items that were accessed up to 1000 times per window were included in the analysis, since Georgia Tech Two had a significantly higher average access rate than Georgia Tech One.

Summary

In summary, the recency analysis shows the logarithmic relationship typically found in the retention memory literature and certain environmental sources. Additionally, recency provides a much better fit than frequency. This is a substantial shift in perspective from other attempts to characterize WWW access logs which tend to focus solely on frequency metrics.

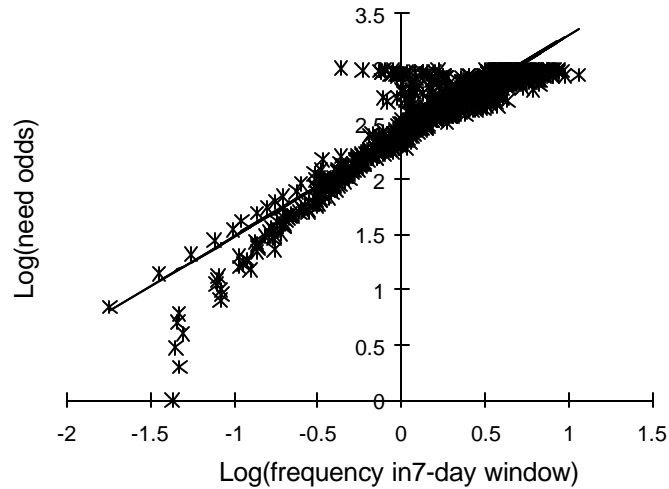


Figure 12. Results of plotting log need odds against log frequency for the Georgia Tech Two data set. Items that were accessed less than 1000 times per window were included in the analysis.

There is also an interesting curiosity—the human memory findings involve the behavior of individuals whereas the library and WWW circulation findings involve the collective behavior of thousands of individuals. With the former information system, the brain is the central element of study, enabled with a complex organic substrate that communicate and processes information. For example, it is easy to imagine that within a brain, the history and context factors are formed and managed via biological signaling between neurons. With the latter system, no element exists that can coordinate and communicate the intended behaviors of the other individuals. No inter-network of neurons exists, only a collection of individuals. How then is it that both classes of information systems exhibit the same properties?

[Simon 1957] contends that while a variety of processes exhibit heavily skewed distributions, there need not be common stochastic process underlying and acting uniformly throughout them. The initial processes studied in [Simon 1957] were the distribution of words in prose by frequency, the distribution of scientists by the number of published papers, the distribution of cities by populations, and the distribution of incomes by size. Intuitively, being skeptical of a common stochastic process makes sense, as it is difficult indeed to imagine what could be acting upon each of these systems.

[Anderson and Schooler 1991]'s argument that human memory has adapted to environmental sources seems more reasonable. At one level, there are events occurring around us with a certain probability that effects the manner in which we process and store this and other information. Yet all of the environmental sources studied in [Anderson and Schooler 1991] are of human origin. Humans write the headlines to the New York Times. Humans utter words to infants. Humans determine with whom and when they write email. Thus, these environmental sources are not purely natural events. It is maybe not as surprising then to find that human memory behaves in a similar manner to the events generated by humans. Yet, this 'human-in-the-loop' argument does not completely explain this phenomenon either, as we eventually end up once again trying to explain how it is that one person knows to email another with this regularity and another person knows to email another with a different regularity.

A different explanation lies in [Price 1976]. [Price 1976] makes the observation that the same class of skewed distributions studied in [Simon 1957] can be explained by what he calls the Cumulative Advantage Distribution, or 'Success Breeds Success'

phenomenon. Imagine a system where items are initially chosen with some base probability. Once an item is chosen, the chances of it being chosen in the future increases. Items that are not chosen do not incur any penalty for not being chosen. For example, suppose we have an urn with an equal number of red and black balls, with red balls signifying success and black balls signifying failure. Initially, there is an equal chance of selecting either color. Upon selection of the first ball, if a red ball is chosen, c more red balls are added to the urn and selection of another ball continues. If a black ball is chosen, no black balls are added and selection stops since failure has been achieved. The chances of selecting a red ball (success) once an initial red ball is selected increases whereas there is no punishment for selecting a black ball (failure). In this manner, success breeds success (SBS). [Price 1976] shows the mathematical equivalence of this type of system to those outlined in [Simon 1957].

It should be immediately clear that SBS speaks directly to the desirability of items on the WWW. Once an item on the Web receives attention, the likelihood that it will receive more attention is increased. If an item is not accessed, no penalty is incurred. Furthermore, as we've just seen, the time of last access is more predictive than the frequency of access. While this gives us a plausible story to explain why items are accessed, it does not provide us with the *mechanism* behind why items are accessed. Although the research is still in the formative stage, initial analysis of data collected by monitoring people's experiences with the Web [Catledge and Pitkow 1995][Tausher 1997], indicates the presence of SBS distributions. On an individual level, there is an initial probability that a person will visit a page followed by a certain probability that the person will revisit the

page (and they do roughly 58% of the time [Tausher 1997]). Pages that are not revisited are not penalized, they are simply not revisited. If indeed SBS distributions occur on an individual, micro level, it is not surprising to find SBS distribution at the group, macro level. More research in this area is needed before any final conclusions can be made.

Application

As originally motivated in [Pitkow and Recker 1994a] and subsequently in [Pitkow and Recker 1994b] an obvious application of the modeling of desirability based upon recency is caching. These findings strongly suggest some form of a least recently used (LRU) replacement policy for server-side caches. Figure 13 shows (see [Pitkow and Recker 1994a][Recker and Pitkow 1996] for more details) that a LRU server-side caching policy results in a very respectable hit-rate (80%) and low miss rates (8% to 30% depending upon which storage strategy is implemented). These findings helped pave the way for the development of and comparison against more complicated algorithms for proxy-caches⁸ [Abrams et al. 1995]. It should be noted that while these more complicated algorithms perform better than a simple LRU policy, the improvements tend to be small, and as a result, may not be worth the added effort of implementation and computational complexity. Furthermore, as pointed out in [Wooster and Abrams 1997], latency may be a

⁸ Proxy-caches are a special form of caches that act as a gateway for a number of users to access the Web, as is the case of users behind a corporate firewall. They handled the requests for the users and keep a cache of retrieved documents, thus decreasing the retrieval time experienced by users for cached pages.

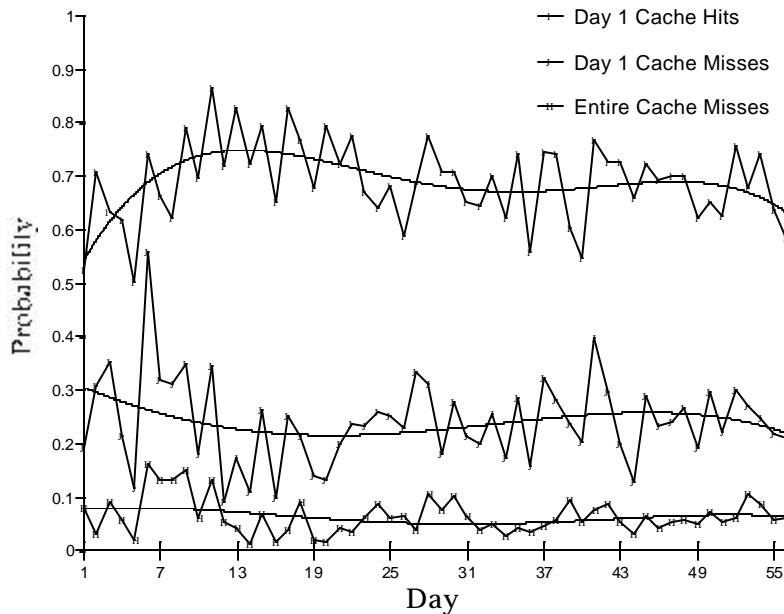


Figure 13. The graph displays the cache hit and miss ratios for a replacement policy of one day and the cache miss ratio for a caching policy of that uses a seven day window. The latter policy assumes that all of the past seven days of accesses are kept in the cache. The x-axis represents day to day performance of the caching policy for all the days in the data set excluding weekends.

better measure of the effectiveness of proxy-cache algorithms given their proximity to end users.

The effectiveness of a LRU replacement policy for caching data and the strength of using recency as a primary modeling characteristic may very well depend upon the ratio of the number of users of the cache versus the number of documents being retrieved. In the case of server-side caching, the number of users (N) typically far outweighs the number of documents being requested (D), that is, $N/D > X$, where X is some parameter, possibly one. This leads to a fair degree of commonality among requests, with many users requesting the same items. It would not be surprising to find a phase shift occur when N/D

$< X$ since there is low probability of requests for the same information from a small set of users against a much larger universe of documents. Another way of expressing this that there is not enough user input to reliably establish the desirability of the information. One might conjecture that desirability requires a certain degree of consensus building, which when faced with an overwhelming amount of information, small populations may not be able to generate. More investigation into the nature of these parameters is necessary.

CHAPTER III

STRUCTURING

Introduction

One aim of this dissertation is to develop methods that automatically categorize and aggregate hypertext information. Once information is categorized, users can increase their rate of information gain per unit time by quickly determining the utility of information, without having to determine the categories themselves. Unlike other approaches to automatic categorization [Cutting et al. 1992] or search engines [Lycos 1994], 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 their associated 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 hypertext and its use that can also be exploited. We refer to methods that utilize this combination of user behavior, the content of the Web, and the Web's topology as *usage+content+structure* techniques.

As mentioned in the Introduction, enabling users to rapidly make decisions about the value of information can greatly increase the user's overall efficiency. In this chapter, we first present a *usage+content+structure* technique that enables pages in Web environments

to be automatically assigned types. Example types include: head pages (the first page to visit in a set of related pages), index pages (pages whose primary purpose is navigation), reference pages (pages that are used to repeatedly explain concepts of content), and content pages (pages that primarily deliver information). In the second half of this chapter, semantic clusters of documents are created using a structural approach called *cocitation analysis*. This technique is taken from the study of citations in the field of Library Science and is shown to produce interesting clusters of related documents.

Document Typing⁹

Many kinds of information can be used to classify or organize collections of WWW pages including the textual content, the hyperlink structure, and various characteristics of the pages including file-system attributes, and usage statistics. Most of this information can be gathered in a straight forward manner for fixed collections of WWW pages, particularly with privileged access to the server's file-system. Over time, the Web infrastructure can be adapted to provide this information directly through standard protocols. Current attempts at providing such functionality include Uniform Resource Characteristics (URC) [Daniel and Mealling 1994] and Apple Computer's Meta Content Framework (MCF) [Guha 1996].

For the purposes of this chapter, we will focus on all the pages served by the Xerox Corporate WWW server (<http://www.xerox.com>) as it existed during the summer of

⁹ Portions of the following material first appeared in [Pirolli et al. 1996].

1996 (this site is the Xerox One site analyzed in the previous chapter). Our goal was to analyze this data and to design algorithms for various extractors which would annotate and aggregate WWW pages at this Web locality. In particular, we have designed methods for classifying nodes into a number of functional categories.

Categorization techniques typically attempt to assign individual elements (e.g., WWW pages) into categories based on the features they exhibit. Based on category membership, we may quickly predict the functionality of an element. For instance, in the everyday world, identifying something as a “chair” enables the quick prediction that an object can be sat on. With respect to the Web, identifying a page as an “head page” informs the user that this page is the logical first page of the collection and is likely to contain an overview of the information contained in the collection as well as pointers to this information. Our techniques will thus rely on the particular features that we can extract about WWW pages at a Web locality.

One may conceive of a Web locality as a complex abstract space in which are arranged WWW pages of different functional categories or types. These functional categories might be defined by a user’s specific set of interests, or the categories might be extracted from the collection itself through inductive technologies [Cutting et al. 1992]. An example category might be organizational home page. Typical members of the category would describe an organization and have links to many other Web pages, providing relevant information about the organization, its divisions or departments, summaries of its purpose, and so on.

We specified a set of functional categories for the purposes of this study. Each functional category was defined in a manner that has a graded membership, with some pages being more typical of a category than others, and Web pages may belong to many categories. That is, for every category, each page has a certain probability of belonging to that category. Membership is determined by selecting the highest value members of each category. Thus, this technique is “fuzzy” in that each page can belong to more than one category. We defined the following types of Web pages for this study:

- *head*: Typically a related set of pages will have one page that would best serve as the first one to visit. Head pages have two subclasses:
 - *organizational home page*: These are pages that represent the entry point for organizations and institutions, usually found as the default home page for servers, e.g., <http://www.org>
 - *personal home page*: Usually, individuals have only one page within an organization that they place personal information and other tidbits on.
- *index*: These are pages that serve to navigate users to a number of other pages that may or may not be related. Typical pages in this category have the words “Index” or “Table of Contents” or “toc” as part of their URL.
 - *source index*: These pages are that are also head nodes, those that are used as entry points and indices into a related information space.

- *reference*: A page that is used to repeatedly explain a concept or contains actual references. References also have a special subclasses:
 - *destination*: In graph theory these are best thought of as “sinks”, pages that do not point elsewhere but that a number of other pages point to. Examples include pages of expanded acronyms, copyright notices, and bibliographic references.
 - *content*: These are pages whose purpose is not to facilitate navigation, but to deliver information.

Data Sources and Collation

The data used for subsequent analyses was derived from two sources: a traversal of the Xerox’s external Web site, whose Uniform Resource Locator (URL) is <http://www.xerox.com>, and the logs of requested items maintained by the Xerox Web server. For this analysis, we choose the access logs from March through May of 1995, in which 1.4 million items were requested. Three basic kinds of data were extracted:

- *Topology* and *meta-information*, which are the hyperlink structure among WWW pages at a Web locality and various features of the pages, such as file size and URL.
- *Usage frequency* and *usage paths*, which indicate how many times a WWW pages has been accessed and how many times a traversal was made from one WWW page to another.

- *Text similarity* among all text WWW pages at a Web locality

Topology and Meta-information

The topology of the Xerox site was ascertained via “the walker”, an autonomous agent that, given a starting point, performs an exhaustive breadth-first traversal of pages within the locality. The walker used the Hypertext Transfer Protocol (HTTP) [Fielding et al. 1997] to request and retrieve items, parsing the returned object to extract hyperlinks. Only links that pointed to objects within the site were added to a list of items to explore. Thus, the walker produced a graph representation of the Web locality’s hyperlink structure, with each node having at least the following meta-information properties: name, title, list of children (pages associated by hyperlinks), file size, and the time the node was last modified. It is important to note that the walker may not have reached all nodes that are accessible via a particular server—only those nodes that were reachable from the starting point were included. This analysis produces an adjacency matrix for the particular locality that we call the *topology matrix*, which represents the node to node hypertext relations, and a set of meta-information called the *meta-document vectors*, which represents the meta-information for each WWW page.

Usage Frequency and Usage Paths

Most servers have the ability to record transactional information about requested items. This information usually consists of at least the time and the name of the URL being requested as well as the machine name making the request. The latter field may represent

only one user making requests from their local machine or it could represent a number of users whose requests are being issued through one machine, as is the case with firewalls and proxies. This makes differentiating the paths traversed by individual users from these access logs non-trivial, since numerous requests from proxied and firewalled domains can occur simultaneously. That is, if 200 users from behind an America Online proxy are simultaneously navigating the pages within a site, how does one determine which users took which paths? This problem is further complicated by local caches maintained by each browser and intentional reloading of pages by the user [Pitkow 1997].

The algorithm we implemented to determine user's paths, a.k.a. "the whittler", utilized the Web locality's topology along with several heuristics. The topology was consulted to determine legitimate traversals while the heuristics were used to disambiguate user paths when multiple users from the same machine name were suspected. The latter scenario relies upon a least recently used bin packing strategy and session length time-outs as determined empirically from end-user navigation patterns [Catledge and Pitkow 1995]. Essentially, new paths were created for a machine name when the time between the last request and the current request was greater than the session boundary limit, i.e., the session timed out. New paths were also created when the requested page was not connected to the last page in the currently maintained path. These tests were performed on all paths being maintained for that machine name, with the ordering of tests being the paths least recently extended. This produced a set of paths requested by each machine and the times for each request. From this, a vector that contained each node's frequency of requests and a matrix containing the number of traversals from one page to another were computed using

software that identified the frequency of all sub-strings for all paths [Pitkow and Kehoe 1995]. These are referred to hereafter as the *frequency vectors* and the *path matrix* respectively.

Additionally, the difference between the total number of requests for a page and the sum of the paths to the page was computed. Intuitively this generates a set of *entry point* candidates. These are the WWW pages at a Web locality that seem to be the starting points for many users. Entry points are defined as the set of pages that are pointed to by sources outside the locality, e.g., an organization's home page, a popular news article, etc. Table 5 shows the results of this analysis. The Xerox Home page and the 1995 Xerox Fact Book are the top pages identified are among the pages most visited by users, as might be expected of people seeking information about Xerox. Also among the top WWW pages are Xerox PARC's Digital Library Home Page, PARC's Map Viewer, the Bookwise Home Page, all of which have received substantial outside press or awards that would draw the attention of users. Entry points might provide useful insight to Web designers based on actual use, which may differ from their intended use on a Web locality. Entry points also may be used in providing a set of nodes from which to spread activation.

Inter-document Text Similarity

Techniques from information retrieval [vanRijsbergen 1979] can be applied to calculate a *text similarity matrix* which represents the inter-document text similarities among WWW pages. In particular, for each WWW page, we tokenized and indexed the text using the TDB [Cutting et al. 1991] full-text retrieval engine. Like many schemes, each document is represented by a vector, where each component of the vector represents a word. Entries

Table 5. The most popular starting points for people visiting the Xerox Corporate Web site (<http://www.xerox.com>).

% Visits Outside	Number Visits	Pages
99.96	2,662	/95FactBook/Title.html
96.18	12,377	/PARC/docs/mapviewer-legend-world.html
99.58	16,004	/Products/XIS/BookWise.html
99.99	19,130	/PARC/dlhx/library.html
94.29	24,107	/ (the default Xerox Home Page)

in the vector for a document indicate the presence or frequency of a word in the document. For each pair of pages, we computed the dot product of these vectors, which produces a similarity measure, which was entered into the text similarity matrix for the Web locality.

It should be noted that the exact specifics of what the TDB retrieval engine does is not as important, as *any* text retrieval engine or set of strategies can be employed, so long as it produces a document by document similarity matrix.

Web Categorization

Previous hypertext research extols the value added by strongly typed node and link systems [Trigg 1983], yet most of the information available on the Web is poorly typed. Even so, a quick tour of WWW pages across Web localities reveals that certain classes of documents do indeed exist. This next section presents an approach to categorization and discussion of the results obtained from categorization of the Xerox Web locality.

Web Page Feature Vectors

In order to perform categorizations we represented each WWW page at the Xerox Web locality by a vector of features constructed from the above topology, meta-information, usage statistics and paths, and text similarities. These WWW page vectors were collected into a matrix. Specifically, a new matrix was created with each row representing a WWW page, and the columns representing the page's:

- *size*, in bytes, of the item
- *inlinks*, the number of hyperlinks that point to the item from the Xerox Web space
- *outlinks*, the number of hyperlinks the item contains that point to other items in the Xerox Web space
- *frequency*, the number of times the item was requested in the sample period
- *sources*, number of times the item was identified as the start of a path traversal
- *csim*, the textual similarity of the item to it's children based upon previous TDB calculation
- *cdepth*, the average depth of the item's children as measured by the number of '/' in the URL.

A logarithmic transform and a z-score normalization was applied to the size, inlinks, outlinks, frequency, and sources values. Two additional matrices were derived from the original data set, one with zero size items removed and the other with only item

whose sizes were between 1000 and 3000 bytes. The removal of zero size items was done to eliminate files that contain no content from interfering with the analysis. These were files that generally served some role to specific applications, e.g., locks, versions, etc. The latter matrix was generated to handle a specific class of pages, pages that contained a medium amount of information. The range was determined by examination of the distribution of files sizes.

Given the above properties and shapes of the distributions, linear separable categories were assumed. This enabled categories to be identified by solving a set of linear equations of the form:

$$c_i = w_1 v_{1i} + w_2 v_{2i} + \dots + w_n v_{ni}$$

for all nodes i in Xerox Web space, where the v_j are the measured features of each Web page, and the w_j are weights. In what follows, these weights were set *a priori* by us to be either -1, 0, or +1. One of the benefits of using linear equations is that it is an unsupervised algorithm. That is, the algorithm does not require a set of pre-categorized pages as input. This property is especially useful with respect to the Web, since the number of pages that would need to be classified is on the order of thousands of pages, which would require significant amounts of human energy and expertise to categorize correctly. Another benefit is that the categories output from this analysis can be used as the training

set for supervised learning algorithms. The two approaches to categorizing document collections are not mutually exclusive.

Table 6 shows the weights used to order of Web pages for each of the categories outlined earlier. For example, we hypothesized that Content Nodes would have few inlinks and few outlinks, but have relatively larger file sizes. The equation used to determine this category of nodes had a positive weight, +1, and negative weight, -1, on the inlink and outlink features. For Head Nodes, being the first pages of a collection of documents with like content, we expected such pages to have high text similarity between itself and its children, would have a high average depth of its children, and that it would be more likely to be an entry point based upon actual user navigation patterns.

Table 6. Node type definitions and precision for linear combination category assignment.

Node Type	Size	Number Inlinks	Number Outlinks	Depth of Children	Similarity to Children	Freq.	Entry Point	Precision
Index	+1 * (outlinks /size)		+1					0.67
Source Index	+1 * (outlinks /size)		+1				+1	0.53
Reference	+1	-1	-1	-1				0.64
Destination Reference	+1	-1	-1	-1			-1	0.53
Head			+1	+1	+1		+1	0.70
Organization Home Page		+1	+1		+1		+1	0.30
Personal Home Page	(> 1000 < 3000 k)					-1	-1	0.51
Content	+1	-1	-1					0.99

Evaluation

Once the set of WWW pages for each category was identified, the top twenty-five members with the highest scores as a result of solving the linear equations specific to each category were extracted, and the first page of the corresponding Web pages printed for off-line evaluation. Each printed page was read and then rated by the three judges as either “belonging to” or “not belonging to” the category it had been associated with. Precision scores were computed for each retrieved set of twenty-five WWW pages, where precision was the proportion of pages rated as “belonging to” the category. Table 6 shows the average precision (geometric mean) for each node category.

As one would expect due the large number of content nodes in a Web locality, the precision at which content nodes can be identified was quite high (precision = 0.99). Equally encouraging was the identification of Head and Index Nodes (precision of 0.70 and 0.67 respectively). Table 7 shows the list of top five Head Nodes. The ability of the categorization scheme to correctly identify Index, Source Index, Reference, Head, and Personal Home pages all had moderate precision scores. It is believed that these scores can be increased by fine tuning the weights for each linear combination. The incorporation of additional factors may also increase precision. Not surprisingly, the lowest precision in Table 6 was associated with the correct identification of Organizational Home Pages, of which there are only about ten such pages at the Xerox Web locality.

Table 7. The top five head nodes as determined by linear combination.

URLs of Page	Titles of Page
/PARC/DigiTrad/DigiTradKeywords.html	Digital Tradition Keywords
/RXRC/Cambridge/trs/html/index.html	RXRC Cambridge Technical Report Series
/PARC/istl/gir/fishkin.html	Ken Fishkin's Public Home Page
/PARC/spl/eca/oi/gregor-invite/transcript.html	Why are Black Boxes so Hard to Reuse?
/Investor/10K-94-Part-IV-g.html	Xerox Corporation 1994 Form 10-K

Summary

The application of a set of linear combinations to feature vectors that contained usage, content, and topology values shows significant promise in being able to aggregate WWW pages into meaningful categories. This approach is a what we call weak approach, in that it sacrifices high precision for the ability to generalize to a large number of sites. That is, it is expected that this technique will work reasonably well across a wide variety of sites rather than working very well at only a few sites. This is not to discredit the attempts of stronger methods. Both methods are useful and deserve further research.

One of the weaknesses of this method though is the reliance upon usage and meta data. While this is not a issue if the analysis is performed by the owners of the locality, it becomes a problem when an independent entity attempts to structure the information on someone else's site. This scenario is typical of both search engines like Lycos [Lycos 1994] and directories like Yahoo [Yahoo 1994]. The next structuring technique being presented deals exclusively with the topology of documents, a characteristic that is readily available to any Web robot.

Clustering¹⁰

One way to approach the automatic clustering of hypertext documents is to adapt the existing approaches of clustering standard text documents [Cutting et al. 1992]. However, there are several impracticalities with such existing text-clustering techniques. Text-based clustering [Cutting et al. 1992] 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 [Botafogo et al. 1992]. 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 [Internet Archive 1997].

In the previous section, each WWW document was represented as a feature vector, with features extracted from information about text-content similarity, hypertext connections, and usage patterns. Categorization 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

¹⁰ Portions of the following material first appeared in [Pitkow and Pirolli 1997].

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 600,000 WWW sites that currently exist [Netcraft 1996].

As a potential way of circumventing these difficulties, we decided to try out *cocitation analysis* [Garfield 1979]. Our adaptation of this clustering technique is based solely on the analysis of hypertext link topology. Unlike other link-topology techniques, cocitation 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 *cocited*. 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 cocitation analysis might be useful in telling us something about the semantic structure of a collection and the thinking of the authoring community.

Cocitation 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 management 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 judgments with author's citations, thus avoiding many of the semantic problems found in term and title based analyses [Garfield 1979].

It was not until the mid-1970s however that [Small and Griffith 1974] developed cocitation analysis as a method for measuring the common intellectual interest between a pair of documents. The principal component of cocitation analysis measures the number of documents that have cited a given pair of documents together. This metric is referred to as *cocitation strength*. Unlike other forms of citation analysis, cocitation 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 cocitation analysis yields insight into the implicit semantic structures of the WWW.

Algorithm

The original algorithm developed by [Small and Griffith 1974] 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 pre-filtering retains the most important (or at least the most popular) documents. Next, the extracted documents are sorted by cited frequency and all pairs of documents that have been cited together by the same source document are

formed. The resulting list contains unique cocitation pairs and their associated frequency of co-occurrence.

The final step in cocitation analysis creates a set of clusters whose elements are indirectly or directly related by cocitation. This is accomplished by clustering all documents that have at least one document of the cocitation 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 process 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 process repeats until all pairs belong to a cluster. We refer to this process as the all-pairs method of cocitation analysis.

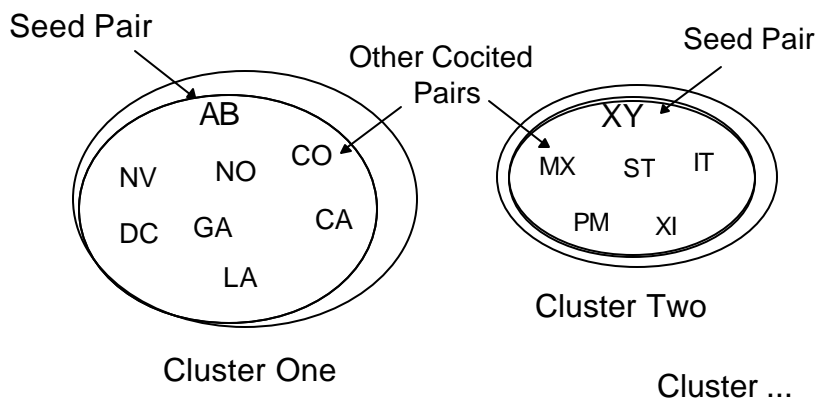


Figure 14. Diagram of the process of the all-pairs cocitation algorithm. Initially a seed pair is chosen at random, and all other pairs that contain any of the cluster are added element (initially A or B in the case of Cluster One). This process repeats, until all pairs have been added. Another random seed pair is then chosen and the process repeats 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 cocitation 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 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 cocitation 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 (<http://www.gvu.gatech.edu>) which contained 5,582 HTML files, 15,139 non-HTML files and 24,768 hyperlinks¹¹. 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 cocitation clustering analysis mentioned above was applied

¹¹ Objects embedded into HTML pages, e.g., images, were not considered hyperlinks for this analysis.

Table 8. For each range of cluster sizes, the total number of clusters formed are given for various citation frequency thresholds.

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

using several different citation frequency thresholds (one, three, five, and ten). Table 9 shows the distribution of the size of clusters using different citation frequency thresholds. For example, using a citation frequency threshold of three, six clusters were formed where each cluster contained between 101 and 500 pages. Table 9 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 cocitation 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.

Table 9. The total number of pages included in the range of cluster sizes using various citation frequency thresholds.

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

As one would expect, lowering the number of times a document is cited results in more documents being included into the cocitation 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.

Table 10. Examples of the types and sizes of clusters formed by the all-pairs method of cocitation analysis.

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 humor publication

The trend for a significant proportion of the documents to belong to a few large clusters is an effect typically found in traditional cocitation analysis of publications [Small and Griffith 1974]. These 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 10 shows examples of the types of clusters formed using this form of cocitation analysis.

Other Cocitation Techniques

Other techniques used in cocitation analysis include hierarchical clustering, multi-dimensional scaling, and factor analysis [McCain 1990], though these techniques typically use authors as the unit of analysis instead of documents. In the case of the WWW, the

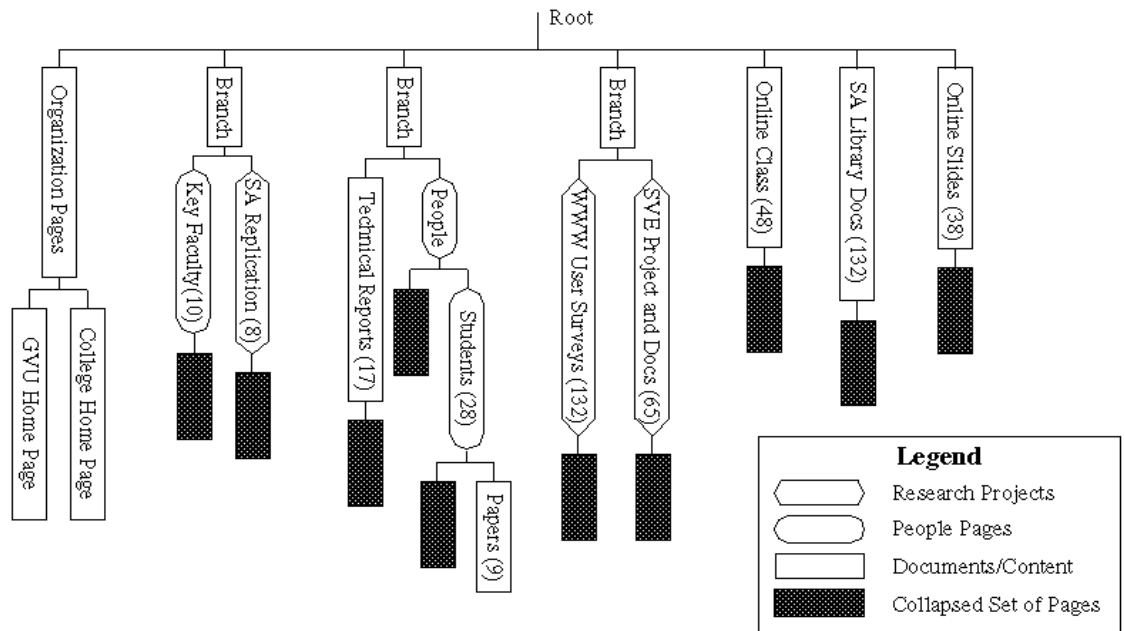


Figure 15. Manually drawn and labeled representation of hierarchical clustering of cocitation matrix. This algorithm was able to identify interesting sets of pages. For example, it was able to identify the two organizational head pages in the collection and distinguish students as a subclass of people.

author of a document can not be reliably determined across sites¹². In our version of these techniques, the cocitation 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 cocitation patterns for each document [White 1990].

For our analysis, we took the three, five, and ten cocited frequency threshold cocitation matrices computed in the above analysis and calculated the Euclidean distance

¹²One could however treat individual WWW sites as authors and perform cocitation analysis at that level.

matrix on the log transformed cocitation frequencies. The resulting distance matrix was then run through a complete linkage clustering algorithm.

Figure 15 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 cocited 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, sub-clusters 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. The next section reviews the results of a quick empirical evaluation of the "goodness" and the precision of the clusters.

Evaluation

In order to determine the effectiveness of the cocitation algorithm, a mini experiment was performed. In this experiment, five randomly selected cocitation clusters were compared to five randomly generated clusters. The cocitation clusters were formed using the hierarchical method from the GVU WWW site. The random clusters were formed by selecting items with equal probability across the entire set of Web pages at GVU. The testing environment consisted of a WWW page that contained the ten clusters randomly

Table 11. Results of the experiment to determine the effectiveness of the cocitation algorithm. The cocitation algorithm performed significantly better than randomly formed clusters with respect to the “goodness” of the clusters and well as precision.

Grouping	Mean	Variance	Standard Deviation	Average Precision	Variance	Standard Deviation
Cocitation	1.067	0.067	0.258	0.973	0.003	0.055
Random	4.067	0.781	0.884	0.242	0.070	0.264

ordered on the page. For each cluster, the URL of each page was presented. The URL was also linked to the corresponding page in the Gvu site. This permitted the inspection of content of all pages in the cluster as well as the corresponding URL of the pages.

Following each cluster, a series of three questions followed. The first question asked for the evaluators to rate on a scale of 1 (good) to 5 (poor) how well formed the group of pages was. The second question asked how many pages did not belong in the group. The final question asked the evaluators to write a few words that described the group of pages. From these three set of questions, the overall “goodness” of the group can be computed along with the precision (how many pages were correctly in the group). Recall (the number of pages that should have been in the group) could not be computed as there was no a priori determination of what the correct clusters should be. Three subjects participated in the experiment. The results are presented in Table 11.

The cocitation algorithm produced significantly better formed clusters, $t(16) = 1.75, p < 0.5$, using a two-tailed T-test for unequal variances. As evident from the mean and variance, the cocitation clusters received the best scores in all but one case by one reviewer. Not surprisingly, the randomly formed clusters did not perform well. With respect

to precision, the average precision score for each cluster was determined and a comparison of means for the cocitation and random clusters was performed. The cocitation algorithm performed significantly better than the randomly formed clusters, $t(4) = 2.78, p < 0.05$. The average precision for the cocitation clusters was near perfect, 0.97. Evaluator agreement ($r = 0.83$) was high for both measures. Furthermore, the evaluators were able to write concise descriptions about the meaning of the cocitation clusters but were often unable to do so for the random clusters.

Summary

Categorization facilitates the user's ability to classify and organize materials, resulting in more efficient foraging in information environments. This chapter presented two new techniques that enable WWW documents to be categorized into specific types of pages and the formation of semantic groupings of related documents. These methods automatically form abstractions and structures that the WWW currently lacks. Since the data sources used to assign document types requires access to usage statistics, it may not be possible to categorize localities that one does not have access to or operate. This does not forego the development of server-side software that embodies the set of heuristics and algorithms. Cocitation analysis, however, only requires structural information, which is readily available. It is thus expected that cocitation analysis could be performed over the entire Web, though sampling may be necessary to form the initial cluster in a computationally efficient manner.

While it is not the purpose of this dissertation to embody these techniques into an application or system, it does seem appropriate to sketch out some possible uses. It should not be difficult to imagine that instead of the document view users of the Web are currently faced with, they would be instead presented with a visualization of the document space similar to those developed in [Card et al. 1995] and [Mukherjea and Foley 1995]. Both of these interfaces could be enhanced by the automatic formation of groups of Web pages to display to users. Users can thus view the major clusters and node types when visiting a particular Web site rather than view individual pages. Users would experience an increase in the amount of information gained from interaction as both the amount of information accessed would increase as well as a decrease in the time necessary to access the information. These users could be information consumers or information providers. Both sets of users can profit from the reduction in the number of documents that require attention and the time necessary to attenuate to them.

The results of node typing and cocitation analysis are higher level abstractions of Web environments, abstraction which can form the basic unit of analysis instead of page level analysis. One can then examine the desirability of categories and clusters as well as the desirability of individual pages. This will hopefully result in more sophisticated analyses and models of information ecologies that explain and model more complexity.

C H A P T E R I V

SPREADING ACTIVATION

Spreading activation can be characterized as a process that identifies knowledge predicted to be relevant to some focus of attention. Figure 16 presents a conceptual diagram of the mechanisms of spreading activation. Initially, one has a network of weighted connections, or associations that determined the relatedness of one node to another. Activation is pumped in a source node or a set of nodes and spread through all connection in the network. Nodes that do not receive any activation during this process lose a small amount of whatever activation they already contain. Once this cycle is completed, the process is repeated, pumping new activation into the source node(s) and spreading the activation amongst the set of connected nodes. After several iterations of this, the network settles into an asymptotic pattern, i.e., the activation levels stabilize in a particular configuration. The nodes with the largest amount of activation at the end of this process are those that are strongly connected to the initial source node(s). In this manner, spreading activation provides a mechanism for determining the set of related entities given a particular source node.

With respect to the Web, if we assume that the nodes represent WWW pages, then the starting node could represent the location of a user in the Web and the network could be the topology, or hyperlink structure of the Web. The set of source nodes could

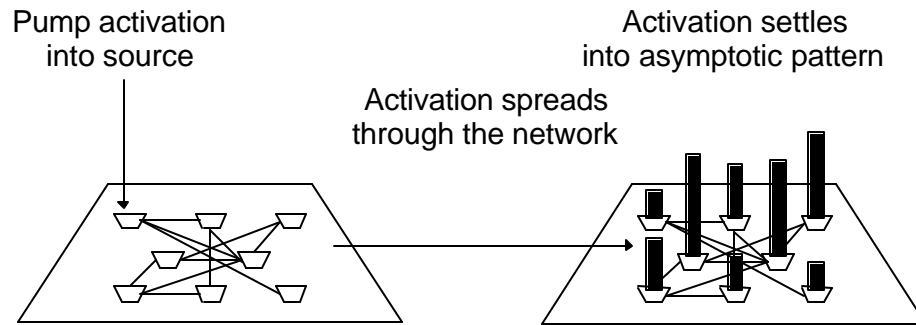


Figure 16. Conceptual diagram of the mechanism of spreading activation. Initially, activation is directed into a source node. Via the set of associations between nodes, this initial activation is spread through other nodes. The process is then repeated, pumping more activation into the source node and activation flowing between associated nodes until an asymptotic pattern emerges. The nodes with the highest activation represent the nodes with the strongest association to the source node. We use a combination of usage, content, and topology networks for spreading activation through Web pages.

also represent other interesting sets of pages, like the person's navigation history in the Web locality. Likewise, the networks need not be limited to just the topology of the Web, it could contain the paths people take through the Web or the textual similarity of nodes.

Suppose a user is interested in a set of one or more Web pages and wants to find related pages to form a small Web aggregate, such as a WebBook [Card et al. 1996]. The identification of the initial set of interesting Web pages could be determined by the users or it could also be automatically determined using the functional categorization techniques described in Chapter Three. Regardless of how the initial set of inputs is identified, spreading activation can be used to automatically create these book-like aggregations. Or, instead of creating books, spreading activation could order the pages based upon their expected need.

This chapter first discusses in more detail the types of networks used for our application, followed by a review of the mathematical foundation for spreading activation. The application of spreading activation to the Web is presented along with results and examples and compared to other structural techniques. Finally, an evaluation of the goodness of the set of predicted pages is presented.

Networks for Spreading Activation¹³

In our research into the utility of employing the spreading activation algorithm, we used three kind of graphs, or networks, to represent strength of associations among WWW pages: (1) the hypertext link *topology* of a Web locality, (2) inter-page *text similarity*, and (3) the *usage paths*, or flow of users through the locality. The collection methods used to form each representation was discussed in Chapter Three. Each of these networks or graphs is represented by matrices in our spreading activation algorithm. That is, each row corresponds to a network node representing a WWW page, and similarly each column corresponds to a network node representing a WWW page. If we index the $1, 2, \dots, N$ WWW pages, there would be $i = 1, 2, \dots, N$ columns and $j = 1, 2, \dots, N$ rows for each matrix representing a graph network.

Each entry in the i^{th} column and j^{th} row of a matrix represents the strength of connection between page i and page j (or similarly, the amount of potential activation flow

¹³ Portions of the following material first appeared in [Pirolli et al. 1996].

or capacity). The meaning of these entries varies depending on the type of network through which activation is being spread:

- in topology networks, an entry of 0 in column i , row j , indicates no hypertext link between page i and page j , whereas an entry of 1 indicates a hypertext link.
- in text similarity networks, an entry of a real number between 0 and 1 in column i , row j indicates the inter-document similarity of page i to page j .
- in usage path networks an entry of an integer strength in column i row j , indicates the number of users that traversed from page i to page j .

Conceptually, activation is pumped into one or more of the graph networks at nodes representing some starting set of Web pages and it flows through the arcs of the graph structure, with the amount of flow modulated by the arc strengths (which might also be thought of as arc flow capacities). The asymptotic pattern of activation over nodes will define the degree of predicted relevance of Web pages to the source pages. By selecting the topmost active nodes or those above some set criterion value, we may extract and rank WWW pages based on their predicted relevance.

Activation Algorithm

The particular version of spreading activation we use is the leaky capacitor model developed in the ACT* theory [Anderson and Pirolli 1984] and studied parametrically by [Huberman and Hogg 1987]. An activation network can be represented as a graph defined by matrix R , where each off-diagonal element $R_{i,j}$ contains the strength of association

between nodes i and j , and the diagonal contains zeros. The strengths determine how much activation flows from node to node. The set of source nodes of activation being pumped into the network is represented by a vector C , where C_j represents the activation pumped in by node i . The dynamics of activation can be modeled over discrete steps $t = 1, 2, \dots, N$, with activation at step t represented by a vector $A(t)$, with element $A(t, i)$ representing the activation at node i at step t . The time evolution of the flow of activation is determined by:

$$A(t) = C + M A(t - 1)$$

where M is a matrix that determines the flow and decay of activation among nodes. It is specified by:

$$M = (1 - g) I + a R$$

where $g < 1$ is a parameter determining the relaxation of node activity back to zero when it receives no additional activation input, and a is a parameter denoting the amount of activation spread from a node to its neighbors. I is the identity matrix.

[Huberman and Hogg 1987] showed that the characteristic dynamic behavior of spreading activation depends on the relation among g , a , and the mean number of arcs per

node, m In the general case, there is a phase transition when $a = g$. When a/g is small, the total activation in the net rapidly rises to an asymptotic pattern and is localized in the network. When $a > g$, there is another phase transition at $m = 1$. With $a > g$ when the network contains sparsely connected nodes with $m < 1$, the total activation rises indefinitely but the pattern remains localized. Our usage path graph structures are such sparse networks. With $a > g$ with richly connected nodes with $m > 1$, the total activation rises indefinitely and all parts of the network affect all others, so that inputs of activation at any node tend to create the same pattern of activation. Our text similarity graphs are richly connected graphs. Given this characterization of the phase space of spreading activation regimes, we chose parameters such that $a/g < 1$ to identify Web structure aggregates.

Example 1: Predicting the Interests of Home Page Visitors

To illustrate, consider the situation in which we identify the most frequently visited organization home page using our categorization information, and wish to construct a Web aggregate that contains the pages most visited from that page. The most popular organization page can be identified as in Chapter Three Table 6 and the corresponding component of C given a positive value, and the remaining elements set to zero. Setting the association matrix R to be the usage path matrix, we then iterate the above equation for N time steps (in our simulations we used $N = 10$). Selecting the twenty-five most active

Table 12. Examples of Web pages selected using spreading activation.

Activation Source	Network	Most Active Web Pages Found (No. found)
Xerox Home Page	Usage paths	Xerox product descriptions (10) Financial reports (6) Business Division home pages (5) General info (2) Search form (1)
Highest rated member of <i>Personal Home Page</i> category	Text similarity	Group project overviews (5) Other people hotlists (4) Company info (4) Personal interests (4) Other similar people (3) Informal groups (1) Workshop attendee list (1) Wildlife award report (1) Someone else's talk (1)

pages constructs the collection described in Table 12. We could have used an activation threshold rather than a fixed set size to circumscribe a Web aggregate.

In Table 12 it is evident that a user who is focused on the Xerox home page is predicted to then shift their attention (by traversing Web links) to WWW pages describe mainly Xerox products, businesses, and financial reports. From this, we might infer that users interested in the Xerox home page are also interested in what Xerox sells or how they are doing financially.

Example 2: Assessing the Typical Web Author at a Locality

Consider another situation in which we are interested in the Web pages having the highest text similarity to the most typical person page in a Web locality. In other words, we might be interested in understanding something about what a typical person publishing in a Web locality says about themselves. In this case, the most typical person page is identified as in Table 6, the corresponding C element set to positive activation input (zeros elsewhere), and R is set to the text similarity matrix. Iteration of this spread of activation for $N = 10$ time steps selects the collection described in Table 12. By reading the group project overviews, the home pages of related people, personal interest pages, and formal and informal groups to which the person belongs, we should get some sense of what people are like in the organization.

Combining Activation Nets

Because of the simple properties of our activation networks, it is easy to combine the spread of activation through any weighted combination of activation pumped from different sources and through different kinds of arc--that is, simultaneously through the topology, usage, and text similarity connections. Consequently, the Web locality can be lit up from different directions and using different colors of predicted relevancy. For instance one might be interested in the identifying the pages most similar in content to the pages most popularly traversed.

Table 13. Results of the experiment to determine the effectiveness of the spreading algorithm. The spreading algorithm performed significantly better than randomly formed clusters with respect to the “goodness” of the clusters and well as precision.

Grouping	Mean	Variance	Standard Deviation	Average Precision	Variance	Standard Deviation
Spreading Activation	1.533	1.124	1.060	0.884	0.038	0.195
Random	4.133	0.838	0.915	0.326	0.012	0.136

Evaluation

In an attempt to measure the overall effectiveness of the spreading activation algorithm, a mini experiment similar to that performed to test the effectiveness of the cocitation algorithm was performed¹⁴. Five groups of pages formed by spreading activation were intermixed with five randomly generated clusters. Unlike the cocitation experiment, the data set used was the Xerox Corporate Web site. Since this site suffered a redesign during the course of this research, the content of many of the pages in the clusters had disappeared. To overcome this shortcoming, the evaluation page contained only the URLs of the clusters. These URLs were not linked to corresponding Web pages for inspection. The experiment used three evaluators.

Table 13 shows the results of the experiment. Spreading activation produced better formed clusters, $t(27) = 2.05, p < 0.05$, using a two-tailed T-test for unequal variances, and achieved significantly higher precision scores, $t(7) = 2.36, p < 0.05$.

¹⁴ In fact, the cocitation experiment was conducted at the same time as the spreading activation experiment.

Evaluator agreement was strong, $r = 0.82$, across both measures. As with the cocitation algorithm, spreading activation produced well formed clusters with high precision.

Although the cocitation experiment differed in the ability of the evaluators to inspect the content of the pages in their determination of cluster goodness, there was not significant difference in the evaluator's rating of each method for goodness of cluster formation, $t(16) = 2.11$, $p = 0.12$, as well as precision, $t(5) = 2.57$, $p = 0.37$. Both tests assumed unequal variance and used a two-tailed T-test. These results are encouraging, but more experimentation is necessary to fully evaluate the effectiveness of the methods given the limited number of number of clusters and evaluators.

Comparison to Link Topology-based Approaches

[Botafogo and Shneiderman 1991] have reported on purely graph theoretic techniques for splitting a hypertext into aggregates. These techniques are based on identifying articulation points in the undirected graph and removing them to create a set of sub-graphs. A node is an articulation point if removing it and its edges would disconnect the graph. [Botafogo and Shneiderman 1991] describe two algorithms which repeat this procedure iteratively. These algorithms removes indices (nodes with relatively high number of out-links) and references (nodes with relatively lots of in-links) on each iteration. The logic of this is to prevent these functional nodes from over-connecting the graph. However, in our case, many of the nodes identified were in fact table-of-contents-like nodes which are very important elements of a Web group.

Table 14. Web aggregation using link topology as outlined in [Botafogo et al 1991].

Cluster	No.	Node ID	First URL
Botafogo Method			
1	32	24	/liveworks/lwi_web/about.html 6
2	23	60	/XSoft/vrpeform.HTML 2
3	11	75	/Products/MajestiK/MajestiKSeries.html 3
4	20	97	/XPS/prodpage/4050.htm 0
5	14	333	/show/PressReleases/Overview.html 4
6	40	456	/PARC/spl/eca/oi-project.html 6
7	50	497	/digitrad/short 0
8	13	508	/PARC/people/jyu/html_training/index.html 0
9	22	615	/RXRC/Cambridge/trs/ps/1995/EPC-1995-101.ps 4
10	14	632	/RXRC/Cambridge/trs/ps/1994/EPC-1994-106.ps 2
Simplified Botafogo Method			
1	22	1	/printsolutions.html 1
2	35	9	/ic1.html 2
3	20	24	/liveworks/lwi_web/about.html 0
4	59	215	/digitrad 0
5	14	333	/show/PressReleases/Overview.html 4
6	32	456	/PARC/spl/eca/oi-project.html 2
7	13	508	/PARC/people/jyu/html_training/index.html 0
8	47	513	/RXRC/Cambridge/people/thumbnails.html 4
9	63	750	/PARC/spl/eca/oi/gregor-invite/P000.html 0

Applying their first algorithm to the graph structure of the Xerox Web produces ten Web groups with at least ten nodes, shown in Table 14. In addition, we tried a simpler algorithm which iteratively removes articulation points until all groups are below twenty-five nodes in size or contain no articulation points. In particular, we didn't remove indices or references during iteration. This leads to nine clusters (again of at least 10 nodes),

shown in Table 14. The two algorithms produced eight Web groups in common, though often not including the same nodes. In addition, the simplified algorithm produced one extra Web group, while the two extra Web groups produced by the [Botafogo and Shneiderman 1991] algorithm were caused by splitting a Web group and by including a spurious Web group.

Unsurprisingly, these algorithms were quite effective at pulling out very typically highly-connected book structures. For example, the thirteen node "html_training/index.html" book was a TOC with 12 nodes for sections which pointed back and forth. These are essentially highly-authored sections of the Web and cluster together in a number of ways. For example, again as would be expected, there was a very high correlation between the URLs of the nodes within these Web group. Most of the nodes typically sharing a prefix of two or three path name parts, though Web groups that were less book-like tended to also bring in a few nodes from other locations on the server.

Comparison to Other Web Specific Approaches

Given the newness of the Web, there has not been a tremendous amount of other research on trying to create meaningful clusters of pages and/or predict relevant pages for users given a particular content. Two efforts though are worth consideration. The first is the work on the Navigational View Builder by [Mukherjea and Foley 1995], which uses a modification of [Botafogo and Shneiderman 1991] to create structural clusterings and a set of manually created meta-information to create content clusterings. Although a formal comparison between the structural methods used by the Navigation View Builder and

those presented in this dissertation has not been conducted, it is expected that they would follow the findings of the comparison to the [Botafogo and Shneiderman 1991] methods.

The other system is Letizia [Lieberman 1997]. Letizia manages a frequency based keyword profile for each user based upon visited Web pages. When a user is visiting pages, Letizia searches the pages connected to the current page and uses the keyword profile to determine relevancy. Recommendations are then made to the user about which pages may be of interest. This system is more of an artful integration of ideas than a methodological contribution.

Summary

The chapter presented a new method for predicting the interests of users in a given context. The context can be defined by the current location of the user, by categorizing a set of pages a user is interested in, or in response to a query. Depending on the informational goals of the user, weighted combinations of each matrix can be used. This enables foraging to be customized for many users in many situations. If usage data is not available for all sites, this can either be approximated from other ascertainable data, e.g., using the number of times a page is linked to as a measure of popularity, or left out of the computations. It is expected that this technique will be useful if embodied into client software.

C H A P T E R V

CONCLUSION

Summary

This dissertation has presented a variety of techniques that help characterize the World Wide Web. Linear combinations of feature vectors utilize a novel fusion of data, integrating usage, content, and structural data to determine node types. Cocitation analysis creates semantic clusters by using only the topology of the Web, a property that leads itself well towards scaling to the entire Web. By using the Gamma-Poisson Model of desirability, recency of access is shown to be a better predictor of future access than frequency, revealing the dynamic temporal nature of the Web is not easily, nor necessarily accurately, captured by frequency based metrics. Finally, spreading activation through usage, content, and structural networks provides a computationally affordable manner to predict document access given a current context. These techniques help create abstractions, categories, and predicted sets that can be used to facilitate user navigation, information management, and structuring of the Web.

In the larger context, the research contained in this dissertation begins to characterize the structure, use, and content of the Web, in new and novel ways. These characterizations are important as they help build the foundations to direct future research.

The encapsulation of the Web as an information system whose main relational property is desirability establishes a clean framework to design empirical measurements. Empirical findings, e.g., recency is a stronger predictor of future use than frequency, provide deeper insight into the structural properties and the regularities of the Web, creating opportunities for well grounded hypothesis testing and theory generation. Without theories and characterization that deal with the use, content, and structure of the Web, the Web might remain a vast repository for information, but of marginal utility to users.

Future Directions

There are numerous applications and future research directions of the characterizations and techniques contained in this dissertation.

Several things need to be said about the generalizability of linear combinations for automatic generation of node types. First, this technique is highly dependent upon the goodness of the features in the feature vectors. For certain Web localities, it may very well be that the collection of documents is structured such that decomposition of the topology into a hierarchy does not produce very meaningful parent-child relationships. This would skew the weights applied to the content similarity to children and the average depth of children features in an indeterminable manner. It is thus expected that the heuristics applied to each locality as well as the features themselves may need to be tuned. Second, there exists a tension between developing strong methods that result in increased precision for certain localities and developing weaker methods that perform reasonably well across a wider range of localities. Clearly, the linear combination work is a weak method and the

contribution of this technique should be assessed accordingly. Third, linear combinations is an unsupervised learning technique, and as such, is well suited to scale the entire Web. Given the success of the feature vectors in determining categories, these features may provide useful input data for supervised learning techniques, which may turn out to outperform linear combinations. Indeed, it is hoped that both weak and strong methods will be developed that surpass the ability of linear combinations as research into automatic typing progresses.

Several interesting opportunities exist with respect to cocitation analysis and scaling to handle the entire Web. A first order attempt may wish to consider only information at the site level, where there are two orders of magnitude fewer data to process. This would provide the user, especially novices, with the top sites to visit, the major arteries of the Web. This could also be of interest as another metric beyond the quite fallible yet current practice of “hit counts” [Pitkow 1997] with which to measure the popularity of sites for marketing and advertising purposes.

Content analysis performed on resulting clusters of sites can be used to automatically assign category labels and site summarization. With this, users can get a high level view not only of the popular content on the Web but also the relationships between content. In a way, this parallels the efforts of other structuring techniques like self-organizing maps [Kohonen 1990] but with a cleaner interpretation of the meaning of the results. However, the success of cocitation analysis at this level may depend upon amount of interconnectedness of sites (diameter), which may be quite different from the interconnectedness of pages within Web localities. Research that parameterizes the

cocitation effectiveness as a function of diameter would be a useful area of exploration, since the results should apply equally well to analysis at the site level as well as individual localities.

Cocitation analysis can also be applied to the entire Web at the document level. This might provide more meaningful results than analysis at the site level as large sites, especially in academia, often contain information on a wide variety of topics, which could inaccurately bias cocitation strengths. As with site level analysis, marketers and advertisers could use this metric to augment other indicators of document popularity. An interesting application would be to combine the two techniques based upon the desired granularity of users. When users are exploring the Web with a specific goal, the site level analysis might provide a useful abstraction to navigate relevant content quickly, followed by document level analysis to find other related documents on the Web. This naturally leads to a variety of novel interfaces and visualizations to support such functionality as well as task specific analysis to determine the amount of information gained, e.g., as per the COKCF. Regardless of the level of analysis, there may also be other measures besides cocitation strength to drive the computation.

The mathematics behind the spreading activation algorithm support embedding into real-time information interfaces. Given that all possible combinations of source activation and different networks can be computed a priori, the interface only needs to be able to sum the desired networks, possibly applying different weights to each network. Control over the weights could reside either within the interface, which could learn the appropriate weights by relevance feedback, or it could reside with the user, who could

manipulate various interface components to adjust the weighting. A dual model of control, with the interface making weighting decisions in the absence of user control may be the more promising technique. Another alternative would be to provide predefined weightings, e.g., “the road most traveled” button would weight the usage network heavily and form a sort of guided tour, or book of the most traversed pages and paths. It is also suspected that other interesting and useful networks exist.

Information Ecologies

With the ever increasing penetration of personal computers into the fabric of society, the study of information ecologies becomes ever more important. The motivation for this field of studying the relationship between humans and information is clear: once information becomes digital, we have the ability to facilitate interaction with information based upon a variety of data including the life-cycle of the information (creation, modification, archiving, and deletion, etc.) as well as the usage history of the information (who used it when and for what purposes, etc.). The exploitation of usage histories has previously been examined in the research on Read and Edit Wear [Hill et al. 1992]. Just the same, it seems that now the conditions are finally right—networked computing is becoming the dominant computing paradigm, technologies like the WWW provide access to resources despite physical location, native protocols, and computing platform, and storage and computation costs have decreased to the point where it is now affordable to collect and process life-cycle and usage data on a wide scale. So, what opportunities exist in the field of information ecologies?

First and foremost is the need to collect data about the interactions of users in heterogeneous, distributed information environments. Barring the costs of instrumentation, there is no reason for user modeling to remain within the boundaries of specific applications, let alone physical boundaries. The first efforts need to be expended on creating an infrastructure and process for distributed interaction data collection and user modeling. That is, an architecture that supports the collection of usage, content and structure data from any application, whether it be email, word processing, spreadsheets, Web browsing, etc., or from any filesystem is necessary to facilitate the formation and maintenance of a complete user information profile (for more on this idea, see [Pitkow 1996]). The actual location of the user, whether at the office, in the car, or at home should be handled gracefully by the system. This architecture enables the creation of an individual, group, and aggregate user profiles that models all interactions of users with information. It also facilitates the further development of Information Foraging and other theories of information consumption and production in dynamic information ecologies.

This is not to posit that analysis of usage in information environments has not been conducted. Several studies on command usage, mostly with the UNIX operating system [Hanson et al. 1984][Desmarais and Pavel 1987][Siochi and Ehrich 1991][Kay and Thomas 1995], and usage of specific applications, mostly text editors and hypertext systems [Hammer and Rouse 1979][Good 1985][Egan et al. 1989][Vaubel and Gettys 1990][Guzdial 1993][Santos and Badre 1995] have been conducted. The relative lack of copious research may very well be a consequence of the high costs associated with the instrumentation of existing systems for data collection. While these studies employ a wide

range of techniques including multivariate analysis, Markov Chaining, sequence analysis, and descriptive statistics, none explore the fundamental nature of information accessed. That is, while the command studies typically look at the frequency of the UNIX command 'ls' (list the contents of the current directory), they did not investigate *what* was being listed, and therefore do not explain *why* the command was issued. Without an underlying theory behind information behaviors, these forms of analyses can at best model usage in an ad-hoc manner, forgoing predictive and explanative abilities the would accompany a theory of information ecologies.

With the development of smarter appliances and the blurring of the boundaries between physical and virtual workspaces, one of the more promising initial areas of exploration exists within corporate Intranets. These information environments typically provide very well structured information coupled with typically very well structured tasks. Since a clearer understanding of the tasks exists within well structured environments than the tasks of typical WWW users, the development of robust weak and strong analytical technique that model information usage on an individual, group, and aggregate level is simplified to a certain extent. Additionally, well structured environments enable crisper evaluations of the overall effectiveness of the system.

In general, architectures can be built that constantly provide relevance—if the system makes a prediction about what information the user is likely to find relevant and the user does not utilize the information, negative reinforcement is provided. The same capability exists for positive reinforcement. This permits the exploration into certain classes of learning algorithms that currently suffer from problems of generating appropriate

training sets and gathering relevance feedback in a low impact manner. Architectures built explicitly to capture user interactions also enable various forms of usability analysis. As instrumented applications and interfaces emerge with the co-evolution of distributed information collected/management architectures, the challenges of developing new analytical techniques will be significant, especially those that incorporate the temporal nature of information and desirability.

The ever increasing amount of information necessitates the development of theories and techniques to facilitate the management human attention efficiently. As with any new emerging field, the study of information ecologies holds much promise and potential. Major frontiers exist on the development of supporting architectures for the collection and management of information along with the development of new analytical techniques to process to extract meaningful representations. But as with any frontier, half the fun is not knowing exactly what the future will hold.

REFERENCES

- [Abrams et al. 1995] Marc Abrams, Charles Standridge, Ghaleb Abdulla, Stephen Williams, and Edward Fox. Caching Proxies: Limitations and Potentials. In *Proceedings of the Fourth International World Wide Web Conference*. Boston, Massachusetts, December 1995.
- [Anderson and Pirolli 1984] John R. Anderson and Peter. L. Pirolli. Spread of Activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10:791–798, 1984.
- [Anderson and Milson 1989] John R. Anderson and Robert Milson, Human Memory: An Adaptive Perspective. *Psychological Review*, 96(4):703-719, 1989.
- [Anderson 1990] John R. Anderson. *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1990.
- [Anderson and Schooler 1991] John R. Anderson and L. J. Schooler. Reflections of the Environment in Memory. *Psychological Science*, 2(6):396–408, 1991.
- [Anderson 1993] John R. Anderson. *Rules of the Mind*. Lawrence Erlbaum Associates, Hillsdale, New Jersey, 1993.
- [Atkinson 1976] R. J. C. Atkinson. *Capital provisions for University Libraries (“The Atkinson Report”)*. London, H.M.S.O., 1976.
- [Berners-Lee et al. 1992] Tim Berners-Lee, Robert Cailliau, Jean-François Groff, and Bernd Pollermann. World-Wide Web: The Information Universe. *Electronic Networking: Research, Applications, and Policy*, 1(2):52-55, 1992.
- [Berners-Lee et al. 1994] Tim Berners-Lee, Robert Cailliau, Ari Luotonen, Henrik Nielsen, and Arthur Secret. The World-Wide Web. *Communications of the ACM*, 37(8):76-82, 1994.
- [Botafogo and Schneiderman 1991] R. A. Botafogo and Ben Shneiderman. Identifying Aggregates in Hypertext Structures. In *Proceedings of the Hypertext 91 Conference*. Pages 63-74, ACM, New York, 1991.
- [Botafogo et al. 1992] R. A. Botafogo, E. Rivlin, and Ben Shneiderman. Structural Analysis of Hypertexts: Identifying Hierarchies and Useful Metrics. *ACM Transactions on Information Systems*, 10(2):142–180, 1992.

- [Burrell 1980] Quentin. L. Burrell. A Simple Stochastic Model for Library Loans. *Journal of Documentation*, 36(2):115–132, 1980.
- [Burrell and Cane 1982] Quentin L. Burrell and Violet R. Cane. The Analysis of Library Data. *Journal of the Royal Statistical Society (Series A)*, 145:439-471.
- [Burrell 1985] Quentin. L. Burrell. A Note on the Aging in a Library Circulation Model. *Journal of Documentation*, 41(2):100-115, 1985.
- [Burrell and Fenton 1994] Quentin L. Burrell and Michael R. Fenton. A Model for Library Book Circulations Incorporating Loan Periods. *Journal of the American Society for Information Science*, 45(2):101-116, 1994.
- [Card et al. 1991] Stuart Card, Jock Mackinlay, and George Robertson. The Information Visualizer: An Information Workspace. In *Conference on Human Factors in Computing Systems (CHI 91)*, New Orleans, Louisiana, April 28 - May 2, 1991.
- [Card et al. 1994] Stuart Card, Peter Pirolli, and Jock Mackinlay. The Cost-of-Knowledge Characteristic Function: Display Evaluation for Direct-Walk Dynamic Information Visualization. In *Conference on Human Factors in Computing Systems (CHI 94)*, pages 238-244, Boston, Massachusetts, April 24-28, 1994.
- [Card et al. 1996] Stuart Card, George Robertson, and William 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, Canada, April 14-18, 1996.
- [Catledge and Pitkow 1995] Lara Catledge and James E. Pitkow. Characterizing Browsing Behaviors on the World Wide Web. *Computer Networks and ISDN Systems*, 27(6), 1995.
- [Cutting et al. 1991] Douglas R. Cutting, Jan O. Penderson and Per-Kristian Halvorsen. An Object-Oriented Architecture for Text Retrieval. *Proceedings of RIAO 91 Intelligent Text and Image handling*, pages 285-298, Barcelona, Spain, 1991.
- [Cutting et al. 1992] Douglas. R. Cutting, D. R. Karger, Jan. O. Penderson, and John. 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.
- [Daniel and Mealling 1994] Ron Daniel Jr. and Michael Mealling. URC Scenarios and Requirements. Draft-ietf-uir-urc-req-00.txt. Internet Engineering Task Force Working Draft, 1994.

- [Desmarais and Pavel 1987] M. Desmarais, and M. Pavel, M. User Knowledge Evaluation: An Experiment with UNIX. In Bullinger, H. and Shackel, B., editors, *INTERACT'87*, pages 151–156. Elsevier Science Publishers, B.V. (North-Holland), 1987.
- [Dunn 1967] O. C. Dunn, W. F. Serbert, and J. A. Seheuneman. *The Past and Likely Future of 58 Research Libraries 1951-1980: A Statistical Study of Growth and Change*. Purdue University Libraries, 1967.
- [Egan et al. 1989] D. E. Egan, J. R. Remde, L. M. Gomez, T. K. Landauer, J. Eberhardt, and C. C. Lochbaum. Formative Design-evaluation of Superbook. *ACM Transactions on Office Information Systems*, 7(1):30–57, 1989.
- [Fielding et al. 1997] Roy Fielding, Jim Gettys, Jeff C. Mogul, Henrik Frystyk Nielsen, and Tim Berners-Lee. RFC 2068—Hypertext Transfer Protocol—HTTP/1.1. University of California Irvine, Digital Equipment Corporation, M.I.T., 1997.
- [Garfield 1979] Eugene Garfield. *Citation Indexing*. ISI Press, Philadelphia, Pennsylvania, 1979.
- [Good 1985] M. Good, M. The Use of Logging Data in the Design of a New Text Editor. In *Conference on Human Factors in Computing Systems (CHI 85)*, San Francisco, California, April 14-18, 1985.
- [Gore 1976] Daniel Gore. The Theory of the Non-Growth, High Performance Library. In *Farewell to Alexandria: Solutions to Space, Growth, and Performance Problems of Libraries*. Pages 164-180, Greenwood Press, Connecticut and London, 1976.
- [Gray 1994] Mark Gray. Growth and Usage of the Web and the Internet. Online Publication. <http://www.mit.edu:8001/people/mkgray/net>, 1996.
- [Guha 1996] R. V. Guha. Meta Content Framework. Online Publication. <http://mcf.research.apple.com>, 1996.
- [Guzdial, 1993] Mark Guzdial. Characterizing Process Change Using Log File Data. Technical Report 93-41, Gvu Center, Georgia Institute of Technology, 1993.
- [Hammer and Rouse 1979] J. Hammer and W Rouse. Analysis and Modeling in Free Form Test Editing Behavior. In *Conference on Cybernetics and Society*, Denver, Colorado, 1979.
- [Hanson et al. 1984] S. J. Hanson, R. E. Kraut, J. M and Farber. Interface Design and Multivariate Analysis of UNIX Command Use. *ACM Transactions of Information Systems*, 2(1):42–57, 1984.

[Hill et al. 1992] William C. Hill, James D. Holland, Dave Wroblewski, and Tim McCandless. Edit Wear and Read Wear. In *Conference on Human Factors in Computing Systems (CHI 92)*, Monterey, California, May 3-7, 1992.

[Huberman and Hogg 1987] Bernardo Huberman and Tadd Hogg. Phase Transitions in Artificial Intelligence Systems. *Artificial Intelligences*. 33:155-171, 1987.

[Internet Archive 1997] Internet Archive. Online Publication. <http://www.archive.org>, 1997.

[Kay and Thomas 1995] J. Kay and R. C. Thomas. Studying Long-term System Use. *Communications of the ACM*, 38(7):61-69, 1995.

[Kent et al. 1979] A. Kent, K. L. Montgomery, J. Cohen, S. Bulick, W. N. Sabor, R. Flynn, and D. L. Shirey. *Use of Library Materials: The University of Pittsburgh Study*. Marcel Dekker, New York, 1979.

[Kohonen 1990] Teuvo Kohonen. The Self-Organizing Map. *Proceedings of IEEE*. 78(9):1464-1479, September, 1990.

[Leimkuhler and Cooper 1971] Ferdinand F. Leimkuhler and Michael D. Cooper. Analytical Models for Library Planning. *Journal of the American Society for Information Sciences*. 22:390-398, 1971.

[Lieberman 1997] Henry Lieberman. Autonomous Interface Agents. In *Conference on Human Factors in Computing Systems (CHI 97)*, Atlanta, Georgia, March 22-27, 1997.

[Lycos 1994] Lycos. Online Publication. <http://www.lycos.com>, 1994.

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

[Morse 1968] P. M. Morse. *Library Effectiveness: A Systems Approach*. M.I.T. Press, Boston, Massachusetts, 1968.

[Netcraft 1996] Netcraft Inc. Netcraft Survey of HTTP servers. Online Publication. <http://www.netcraft.co.uk/Survey/Reports/>, 1996.

[Mukherjea and Foley 1995] Sougata Mukherjea and James Foley. Visualizing the World Wide Web with the Navigational View Builder. *Computer Networks and ISDN Systems*. 27(6), 1995.

[Recker and Pitkow 1996] Margaret M. Recker and James E. Pitkow. Predicting Document Access in Large Multimedia Repositories. *ACM Transactions on Computer-Human Interaction*. 3(4):352-375, 1996.

[Pirolli 1991] Private Communication.

[Pirolli et al. 1996] Peter Pirolli, James Pitkow, and Ramana Rao. Silk from a Sow's Ear: Extracting Usable Structures from the Web. *Conference on Human Factors in Computing Systems (CHI 96)*, Vancouver, Canada, April 13–18, 1996.

[Pirolli and Card 1995] Peter Pirolli and Stuart Card. Information Foraging in Information Access Environments. In *Conference on Human Factors in Computing Systems (CHI 95)*, Denver, Colorado, May 7-11, 1995.

[Pirolli et al. 1996] Peter Pirolli, P. Shank, Marti Hearst, and C. Diehl. Scatter/Gather Browsing Communicates the Topic Structure of a Very Large Text Collection. In *Conference on Human Factors in Computing Systems (CHI 96)*, Vancouver, Canada, April 13-18, 1996.

[Pitkow and Recker 1994a] James E. Pitkow and Margaret M. Recker. A Simple Yet Robust Caching Algorithm Based on Dynamic Access Patterns. *The Second International World Wide Web Conference*, Chicago, Illinois, Oct. 20–25, 1994.

[Pitkow and Recker 1994b] James E. Pitkow and Margaret M. Recker. Integrating Bottom-Up and Top-Down Analysis for Intelligent Hypertext. In *Third International Conference on Intelligent Knowledge Management*, Maryland, Maryland, Nov. 29–Dec. 2, 1994.

[Pitkow 1996] James E. Pitkow. Information Ecologies. Proposal of Dissertation Research. Unpublished. Georgia Institute of Technology, May 17, 1996.

[Pitkow and Kehoe 1995] James E. Pitkow and Colleen M. Kehoe. Results from the Third World Wide Web User Survey. *The World Wide Web Journal*, 1(1), 1995.

[Pitkow and Kehoe 1996] James E. Pitkow and Colleen M. Kehoe. GVU's Sixth WWW User Survey Results. Online Publication. http://www.gvu.gatech.edu/user_surveys, 1996.

[Pitkow 1997] James E. Pitkow. In Search of Reliable Usage Data on the WWW. In *Proceedings of the Sixth International World Wide Web Conference*, Santa Clara, California, April 7-11, 1997.

[Pitkow and Pirolli 1997] James E. Pitkow and Peter Pirolli. Life, Death, and Lawfulness on the Electronic Frontier. In *Conference on Human Factors in Computing Systems (CHI 97)*, Atlanta, Georgia, March 22-27, 1997.

- [Price 1976] Derek de Solla Price. A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science*. September-October 1976.
- [Santos and Badre 1995] P. Santos and A. Badre. Automatic Chunk Detection in Human-Computer Interaction. Technical Report 94-4, GVU Center, Georgia Institute of Technology, 1995.
- [Simon 1957] Herbert A. Simon. *Models of Man Social and Rational*. Chapter Nine: On a Class of Skew Distribution Functions. New York, New York, John Wiley and Sons, 1957. Also in *Biometrika*, 52:425-440, 1957.
- [Siochi and Ehrich 1991] A. C. Siochi and R. W. Ehrich. Computer Analysis of User Interfaces Based on Repetition in Transcripts of User Sessions. *ACM Transactions of Information Systems*, 9(4):309-335, 1991.
- [Small and Griffith 1974] H. Small and B. Griffith. The Structure of Scientific Literatures I: Identifying and Graphing Specialties. *Science Studies*, 4(17):17-40, 1974.
- [Smith and Winterhalder 1992] E. A. Smith and B. Winterhalder, Editors. *Evolutionary Ecology an Human Behavior*. de Gruyter, New York, New York, 1992.
- [Sandison 1977] A. Sandison. Models for Librarians. *Journal of the American Society for Information Science*. 28:300-301, 1977.
- [Stephens and Krebs 1986] D. W. Stephens and J. R. Krebs. *Foraging Theory*. Princeton University Press, Princeton, New Jersey, 1986.
- [Tague and Ajiferuke 1987] Jean Tague and Isola Ajiferuke, The Markov and the Mixed-Poisson Models of Library Circulation Compared. *Journal of Documentation*. 43(3):313-231, 1987.
- [Trigg 1983] Randall H. Trigg. A Network-based Approach to Text Handling for the Online Scientific Community. Ph.D. Thesis, University of Maryland, 1983.
- [University Grants Committee 1976] University Grants Committee. *Capital provisions for University Libraries: report of a Working Party*. London. H.M.S.O., 1976.
- [vanRijsbergen 1979] C. J. vanRijsbergen. *Information Retrieval*. Butterworth & Company, Boston, Massachusetts, 1979.
- [Vaubel and Gettys. 1990] K. P. Vaubel, and C. F. Gettys. Inferring User Expertise for Adaptive Interfaces. *Human Computer Interaction*, 5:95-117, 1990.

[Webster 1988] Webster New World Dictionary Third College Edition. Simon and Schuster, Inc., New York, New York, 1988.

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

[Wooster and Abrams 1997] Roland P. Wooster and Marc Abrams. Proxy Caching that Estimates Page Load Delays. In *Proceedings of the Sixth International World Wide Web Conference*. Santa Clara, CA, April 1997.

[Yahoo 1994] Yahoo. Online Publication. <http://www.yahoo.com>, 1994.

V I T A

James E. Pitkow was born on April 27, 1970 in Philadelphia, Pennsylvania. He graduated with honors in Psychology from the University of Colorado in 1993.