Silk from a Sow's Ear: Extracting Usable Structures from the Web

*Peter Pirolli, James Pitkow, Ramana Rao**

Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304, USA
E-mail: {pirolli, rao}@parc.xerox.com, pitkow@cc.gatech.edu

**authors are ordered alphabetically*

ABSTRACT

In its current implementation, the World-Wide Web lacks much of the explicit structure and strong typing found in many closed hypertext systems. While this property probably relates to the explosive acceptance of the Web, it further complicates the already difficult problem of identifying usable structures and aggregates in large hypertext collections. These reduced structures, or localities, form the basis for simplifying visualizations of and navigation through complex hypertext systems. Much of the previous research into identifying aggregates utilize graph theoretic algorithms based upon structural topology, i.e., the linkages between items. Other research has focused on content analysis to form document collections. This paper presents our exploration into techniques that utilize both the topology and textual similarity between items as well as usage data collected by servers and page meta-information lke title and size. Linear equations and spreading activation models are employed to arrange Web pages based upon functional categories, node types, and relevancy.

Keywords

Information Visualization, World Wide Web, Hypertext.

INTRODUCTION

The apparent ease with which users can click from page to page on the World-Wide Web (WWW) belies the real difficulty of understanding the what and where of available information. The primary approaches widely provided for finding information are search systems like Lycos and Harvest and human-organized directory browsers like Yahoo and Internet Yellow Pages [18]. Though these approaches are quite powerful, they don't exhaust the potential space of approaches.

We suggest that the ecological approach of Information Foraging Theory [12] motivates the use of techniques for automatic categorization and a particular kind of associative retrieval of WWW pages involving *spreading activation* [4]. Categorization techniques are used to identify and rank particular kinds of WWW pages, such as "organization home pages" or "index pages." Associative retrieval techniques identify and rank pages predicted to be related and relevant to currently viewed pages. Both techniques work on the basis of a variety of features of WWW pages, including the text content, usage patterns, inter-page hypertext link topology, and meta-document data such as file size and file name. In combination, these techniques may be used in support of novel information visualization techniques, such as the WebBook [6], to form and present larger aggregates of related WWW pages. We will argue that these techniques can be used to optimize user information-seeking and sensemaking.

To be published in the *Proceedings of the ACM Conference on Human Factors in Computing Systems*, *CHI-96*, Vancouver, Canada, 1995.

INFORMATION FORAGING ON THE WWW

In an information-rich world, the limiting resource is the time and attention of users. We are developing Information Foraging Theory [12] as an attempt to understand how people adaptively allocate their time and attention in the pursuit and consumption of valuable information. Foraging theory in anthropology [16] and biology [17] attempts to explain how humans and animals optimize their gain of food energy from flux of the physical and organic environment. Similarly, Information Foraging Theory attempts to explain how human *informavores* optimize their information gain from the flux of the cultural and technological environment. We can think of the WWW as an information environment in which information-seeking users will want optimized foraging technologies. We will argue that categorization and associative retrieval techniques improve the optimality of WWW interactions.

Categorizing to Optimize Foraging Decisions

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 [12]. 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 their time more usefully.

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 their ability to rapidly categorize the prey types (WWW page types), assess their prevalences on the plain (web locality), assess their profitabilities (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 1 provides a graphical illustration of the improvements provided by categorization and ranking (see Pirolli and Card [12] for detailed technical discussion).

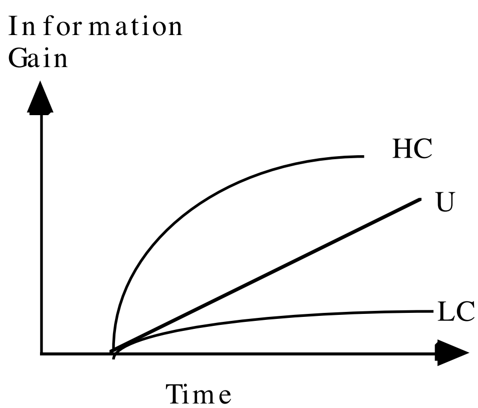

Figure 1. Uninformed search through an uncategorized web locality produces a linear information gain function as in U. Categorizing and ranking WWW pages allows an information forager to rapidly identify high value, ranked, categories (HC) and low value categories (LC) and concentrate on exploiting the HC gain curve.

Spreading Activation to Predict Needed Information

Anderson and Milson [1, 3] have recently argued that human memory has adapted through evolution to optimize the retrieval of needed information (memories) based on the current 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. More recently, Anderson [2] 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). We will use a kind of spreading activation mechanism [4] to predict the needed, relevant information, computed using past usage patterns, degree of shared content, and WWW hyperlink structure.

Monitoring the Evolution of Web Ecologies

The WWW itself is an evolving ecology of information-bearing items, each of which is competing for human attention. One can think of the knowledge content of the WWW as a collection of *memes* [9], or cultural knowledge units, analogous to genes, that are transmitted and replicated. People who manage WWW sites (e.g., Webmasters) may want to assess general properties of the information evolving at their sites, as well as general properties about which users are attending to which kinds of information. We will illustrate how categorization and associative retrieval techniques provide a means for monitoring the interaction of users and WWW pages in the information habitat of a WWW site.

RELATION TO WEB VISUALIZATION

Most current WWW browsers provide very little support for helping people gain an overall assessment of the structure and content of large collections of WWW pages. Information Visualization could be used to provide an interactive overview of web localities that facilitates navigation and general assessment. Visualizations have been developed that provide new interactive mechanisms for making sense of information sets with thousands of objects [15]. The general approach is to map properties and relations of large collections of objects onto visual, interactive structures. Such visualizations provide value to the extent that the mapped properties help users navigate around the space and remember locations or support the unit tasks of the user's work.

Visualizations can be applied to the Web by treating the pages of the Web as objects with properties. Mukerjea [11] has applied a number of visualizations to the Web. Each of these provides an overview of a Web locality in terms of some simple property of the pages. For example, a Cone Tree shows connectivity structure between pages and a Perspective Wall shows time-indexed accesses of the pages. Thus, these visualizations are based on one or a few characteristics of the pages.

As argued above, our approach to extracting structure from the Web using categorization and spreading activation, can be used to form higher level abstractions that reduce the complexity and increase the richness of an overview. We have developed methods for annotating pages with their functional types and relevancy/importance assessments as well as aggregating the Web into collections which can be treated as collections. These collections could be visualized as WebBooks in the WebForager [6].

OVERVIEW OF THE APPROACH

Many kinds of information can be used to classify or organize collections of WWW pages including the textual content, the connectivity (hyperlink) structure, and various characteristics of the pages including file-system attributes and access statistics, as well as usage statistics. Most of this information can be straight-forwardly gathered for fixed collections of pages, particularly with privileged access to the server's storage system. Over time, the Web infrastructure can be adapted to provide this information directly through standard protocols.

We have gathered information for several Web localities, but in this paper, we focus on the pages served by the Xerox WWW server. 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, and spreading activation based on selecting one or more source nodes (WWW pages) and dimensions of interest, and aggregating nodes into higher level collections.

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. Our techniques rely on particular features we can extract about WWW pages at a Web locality.

One may conceive of a Web locality as a complex abstract space in which WWW pages of different functional categories or types are arranged. 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 [13]. 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 this study that correspond to the various roles that Web pages typically play. 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. Specifically, we defined the following types of Web pages for our 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 server 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 entry points and indices into a related information space. Thus they are also head nodes.
- *reference*: A page that is used to repeatedly explain a concept or contains actual references. References also have a special subclass:
    - *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.

We will use spreading activation mechanisms to predict the relevance of Web pages at a locality to WWW pages that are in the a current focus of attention of the user. The degree of relevance of Web pages to one another can be conceived as similarities, or strength of associations, among Web pages located in an abstract space. These strength-of-association relations can be represented formally using a composite of several graph structures. Each graph structure contains nodes representing Web pages, and directed arcs among nodes are labeled with values representing strength of association among pages. One type of graph structure represents the hypertext link topology of a Web locality by using arcs labeled with unit strengths to connect one graph node to another when there exists a hypertext link between the corresponding Web pages. This is perhaps the most common intuition that people hold when thinking about a locality. A second type of graph structure represents the inter-page text content similarity by labeling arcs connecting nodes with the computed text similarities between corresponding Web pages. This is a common way of conceptualizing documents in search-based information retrieval. A third type of graph structure represents the flow of users through the locality by labeling the arcs between two nodes with the number of users that go from one page to another. This essentially reflects the mutual desirability of information among Web pages based on observed usage patterns. Predicted relevance of WWW pages can be based on one or all of the similarity structures represented by these graphs of hypertext links, text overlap, or usage patterns.

In summary, given this conception of a Web locality, we use three component processes to identify and extract Web structure. First is the categorization of Web pages into types and the users' selection of pages according to their degree of category membership in different types. Second is the spread of activation from the identified sources through some combination of the link connectivity, text similarity, and usage spaces. Third is the selection and aggregation of Web pages based on their pattern of relevancy as measured by activation.

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. Four basic kinds of data were extracted:

- *Topology* , which is the hyperlink structure among WWW pages at a Web locality.
- *Page meta-information*, which includes 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

These data were used to construct two types of representation, used in the categorization and spreading activation computations respectively:

- Feature-vector representations of each WWW page that represent values of each page on each dimension.
- Graph representations (in matrix form) of the strength of association of WWW pages to one another.

Topology

The site's topology 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) 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 hyperlink structure of the Web locality. 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*.

Page Meta-Information

For each node visited in a site at least the following meta-information properties was collected: name, file size, and the time the node was last modified. These and other page characteristics can be collected by the walker from the HTTP reply header or the HTML document header. In addition, another process, called “the ripper”, was used to extract characteristics directly from the Web server’s file system. The meta-information for each page was collected into *meta-document vectors*.

Usage Statistics and Paths, and Entry Points

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.

Table 1. Node type definitions and results.

Node Type	Size	Number Inlinks	Number Outlinks	Depth of Children	Similarity to Children	Freq.	Entry Point	Precision
Index	- (outlinks/size)		+					0.67
Source Index	- (outlinks/size)		+				+	0.53
Reference	+	-	-	-				0.64
Destination Reference	+	-	-	-			-	0.53
Head			+	+	+		+	0.70
Org. Home Page		+	+		+		+	0.30
Personal Home Page	> 1 k & < 3 k					-	-	0.51
Content	+	-	-					0.99

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 [7]. 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 k-substrings for any n-string [14]. 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 2 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.

Table 2. The most popular starting points

% Visits Outside	Number Visits	Pages
99.96	2662	/95FactBook/Title.html
96.18	12377	/PARC/docs/mapviewer-legend-world.html
99.58	16004	/Products/XIS/BookWise.html
99.99	19130	/PARC/dlbx/library.html
94.29	24107	/ (the default Xerox Home Page)

Inter-document Text Similarity

Techniques from information retrieval [19] 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 full-text retrieval engine [8]. Like many schemes, each document is represented by a vector, where each component of the vector represents a word. Entries 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 the these vectors, which produces a similarity measure, which was entered into the text similarity matrix for the Web locality.

WEB CATEGORIZATION

Previous hypertext research extols the value added of strongly typed node and link systems, 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 dataset, one with zero size items removed and the other with only item whose sizes were between 1000 and 3000 bytes.

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:

$$c_i = w_1 v_1 + w_2 v_2 + ... + w_n v_n \quad (1)$$

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. Future works may investigate classification algorithms for optimizing the setting of these weights, for instance, by using Linear Discriminant Analysis, or more sophisticated machine learning techniques.

Table 1 shows the weights used to order Web pages for each of the categories. 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 has a positive weight, +1 on the size feature, 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.

Evaluation

Once the set of WWW pages for each category were identified, the top 25 members with the highest scores on Linear Equation 1 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 authors as either "belonging to" or "not belonging to" the category it had been associated with. Precision scores were computed for each retrieved set of 25 WWW pages, where precision was the proportion of pages rated as "belonging to" the category. Table 1 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 = .99). Equally encouraging was the identification of Head and Index Nodes (precisions of .70 and .67 respectively). Table 3 shows the list of top five Head Nodes. Not surprisingly, the lowest precision in Table 1 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 3. The top 5 head nodes.

Titles of Page
Digital Tradition Keywords
RXRC Cambridge Technical Report Series
Ken Fishkin's Public Home Page
Why are Black Boxes so Hard to Reuse?
Xerox Corporation 1994 Form 10-K

SPREADING ACTIVATION

Spreading activation can be characterized as a process that identifies knowledge predicted to be relevant to some focus of attention. The particular version we use is a leaky capacitor model developed in the ACT* theory [4] and studied parametrically by Huberman and Hogg [10].

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 [6]. We assume that the identification of such "interesting seed sets" of Web pages could also be automatically determined based on functional category using methods described above.

Networks for Spreading Activation

As outlined above, 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. 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 page, and similarly each column corresponds to a network node representing a page. If we index the 1, 2, ...,N pages, there would be i = 1, 2, ...N columns and j = 1, 2, ...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 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, $s \geq 0$, 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, $s \geq 0$, 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 a starting set of Web pages and it flows through the arcs of the graph structure, with amount of flow modulated by arc strengths (which can 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

An activation network can be represented as a graph defined by matrix **R**, where each off-diagonal element $\mathbf{R}_{i,j}$ 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 $\mathbf{C}_i$ represents the activation pumped in by node i. The dynamics of activation can be modeled over discrete steps $t = 1, 2, \ldots N$, 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

$$\mathbf{A}(t) = \mathbf{C} + \mathbf{M}\,\mathbf{A}(t - 1), \tag{2}$$

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

$$\mathbf{M} = (1 - \gamma)\,\mathbf{I} + \alpha\,\mathbf{R}, \tag{3}$$

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

Huberman and Hogg showed that the characteristic dynamical behavior of spreading activation depends on the relation among γ, α, and the mean number of arcs per node, μ. In the general case, there is a phase transition when $\alpha = \gamma$. When α/γ is small, the total activation in the net rapidly rises to an asymptotic pattern and is localized in the network. When $\alpha > \gamma$, there is another phase transition at $\mu = 1$. With $\alpha > \gamma$, when the network contains sparsely connected nodes with $\mu < 1$, the total activation rises indefinitely but the pattern remains localized. Our usage path graph structures are such sparse networks. With $\alpha > \gamma$, with richly connected nodes with $\mu > 1$, 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 $\alpha/\gamma << 1$ 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 we did in Table 1 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 Equation 2 for *N* time steps (in our simulations we used *N* = 10). Selecting the 25 most active pages constructs the collection described in Table 4. We could have used an activation threshold rather than a fixed set size to circumscribe a Web aggregate.

In Table 4, 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

Table 4. 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 *Personal Home Page* 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)

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 1, 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 4. 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 though 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.

PURE LINK TOPOLOGY-BASED APPROACHES

Botafoga et al [5] 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 subgraphs. A node is an articulation point if removing it and its edges would disconnect the graph. Botafoga et al 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 in order to prevent these functional nodes from overconnecting 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.

Applying their first algorithm to the graph structure of the Xerox Web produces 10 web groups with at least 10 nodes. In addition, we tried an algorithm which iteratively removes articulation points until all groups are below 25 nodes in size or contain no articulation points. We didn't remove indices or references during iteration. This leads to 9 clusters (again of at least 10 nodes. The two algorithm produced 8 web groups in common, though often not including the same nodes. In addition, the simplified algorithm produced one extra web group, while the 2 extras web groups produced by the Botafoga algorithm were caused by splitting a web group and by including a spurious web group.

These algorithms were quite effective at pulling out highly-connected book structures. For example, one 13 node book was a TOC with 12 nodes for sections which pointed back and forth. However such structures are highly-authored sections of the web and cluster together in a number of ways. For example, there was a high correlation between the URLs of the nodes within these web groups. Most of the nodes typically shared a prefix of two or three pathname parts, though web groups that were less book-like tended to also bring in a few nodes from other locations on the server. More systematic comparison of pure topology-based methods and those based on one or more of the sources outlined here is an open area for further work.

SUMMARY

Higher level abstractions over hypertext can be used to improve navigation and assimilation of hypertext spaces. Previous work on structure extraction has typically used topological or textual relationships to drive analysis. In this paper, we have integrated these data sources as well as new sources including usage statistics and page meta-information to develop new techniques for node typing, group extraction, and relevancy determination. These methods can provide leverage in designing more usable overviews and visualizations of Web spaces.

REFERENCES

1. Anderson, J.R., *The adaptive character of thought*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.
2. Anderson, J.R., *Rules of the mind*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1993.
3. Anderson, J.R. and R. Milson, Human memory: An adaptive perspective. *Psychological Review*, *96* (1989). 703-719.
4. Anderson, J.R. and P.L. Pirolli, Spread of activation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10* (1984). 791-798.
5. Botafogo, R.A. and B. Schneiderman. Indentifying aggregates in hypertext structures. in *Third ACM Conference on Hypertext*. (1991, San Antonio, TX). ACM. pp. 63-74.
6. Card, S.K., G.G. Robertson, and W.M. York. The WebBook and the Web Forager: An Information Workspace for the World-Wide Web. in *Conference on Human Factors in Computing Systems, CHI-95*. (1996, ACM. pp.
7. Catledge, L. and J. Pitkow, Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, *27* (1995). .
8. Cutting, D.R., J.O. Pedersen, and P.-K. Halvorsen. An Object-Oriented Architecture for Text Retrieval. in *RIAO '91 Intelligent Text and Image Handling*. (1991, Barcelona, Spain). pp. 285-298.
9. Dawkins, R., *The selfish gene*. Oxford University Press, Oxford, 1976.
10. Huberman, B.A. and T. Hogg, Phase transitions in artificial intelligence systems. *Artificial Intelligences*, *33* (1987). 155-171.
11. Mukherjea,S., J.D.Foley, and S.Hudson. Visualizing complex hypermedia networks through multiple hierarchical views. in *Human Factors in Computing Systems CHI-95*.(1995, Denver, CO). ACM. pp. 331-337.
12. Pirolli, P. and S. Card. Information foraging in information access environments. in *Conference on Human*