Using Information Scent to Model User Information Needs and Actions on the Web

Ed H. Chi, Peter Pirolli, Kim Chen*, James Pitkow
Xerox Palo Alto Research Center
3333 Coyote Hill Road
Palo Alto, CA 94304
{echi,pirolli,kchen,pitkow}@parc.xerox.com

ABSTRACT
On the Web, users typically forage for information by navigating from page to page along Web links. Their surfing patterns or actions are guided by their information needs. Researchers need tools to explore the complex interactions between user needs, user actions, and the structures and contents of the Web. In this paper, we describe two computational methods for understanding the relationship between user needs and user actions. First, for a particular pattern of surfing, we seek to infer the associated information need. Second, given an information need, and some pages as starting points, we attempt to predict the expected surfing patterns. The algorithms use a concept called "information scent", which is the subjective sense of value and cost of accessing a page based on perceptual cues. We present an empirical evaluation of these two algorithms, and show their effectiveness.

Keywords
Information foraging, information scent, World Wide Web, usability, data mining, information retrieval.

INTRODUCTION
"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."[1] ~Herb Simon (Nobel Prize Winner)

The Web has exploded into an information ecology with several hundred million users and over a billion Web pages. Understanding the complexity of the interactions of information seekers with the Web ecology is a difficult scientific endeavor that has great practical value [12,18]. For example, understanding the goals and behavior of information seekers could aid in more effective Web designs and services. In an effort to understand user goals and actions, most Web sites record user interaction data in some form. There are tools for understanding complex user interactions from these logs. Data mining approaches attempt to identify strength of associations among user clicks and actions [10]. Other approaches, such as Accrue Insight [1] and Astra SiteManager [4] produce descriptive statistics and analyses, such as number of hits, unique users, reading times, session lengths, and download times.

While effective in some ways, these solutions are unsatisfactory because we currently lack an underlying model for understanding the goals of these user actions. Consequently, it is difficult to go from such summary data to actual insights about the users' needs. In particular, little attention has been paid to extracting the *intentions* of users, or to predicting goal-based Web surfing behavior.

This paper aims to develop predictive models capable of modeling and extracting user needs and simulating usage with respect to these needs. Such an approach should answer questions concerning the interests of the visitors to a Web site, and predict their paths through the Web site given their interests. We describe in detail two algorithms that have been developed to answer these questions by using a concept called *information scent*. This underlying assumption, derived from Information Foraging theory [15,17] is that user behavior in the information environment is guided by information scent, which is determined from the perception of the value and cost of the information with respect to the goal of the user. Using this concept, the methods presented here (a) analyze the needs of Web site visitors based on their surfing patterns and (b) simulate usage of the Web based on user goals. We have implemented this concept in a working analytical system intended for researchers, practicing Web site designers, and content providers.

While a rough description of these methods appeared previously in the context of a visualization system [9], we were not able to include the details of the implementation due to space constraints, nor did we present strong evaluations of the algorithms. In this paper, we will present details of how these methods are implemented, as well as present two new

* Kim was supported by a summer undergraduate internship program.
[1] As quoted by Hal Varian in *Scientific American* Sept. 1995.

SIGCHI'01, March 31-April 4, 2001, Seattle, WA, USA.

evaluations of our approach. We also describe some new refinements to the algorithms.

This paper is organized as follows. First, we discuss some related work. We then describe our scent algorithms with some examples. Next, we evaluate the algorithms to show their consistency and validity. Last, we mention some potential applications and give concluding remarks.

RELATED WORK

In a previous paper [9], we described a visualization system that uses the two scent algorithms described here in conjunction with the Longest Repeated Subsequences (LRS) path data-mining algorithm [21]. Because of the breadth of that paper, we were not able to include detailed descriptions of the two algorithms, nor strong empirical evaluations of the algorithms. In this paper, we focus on the details and the evaluations of these algorithms.

There are few systems that try to predict Web site usage patterns or usability based on Web site designs. WebCriteria SiteProfile [26] employs software agents as surrogate users to traverse a Web site and derives various usability metrics from simulated surfing. The simulated browsing agents navigate a Web site and record download times and other information. The data are integrated into metrics that assess: (a) the load times associated with pages on the site and (b) the accessibility of content (ease of finding content). However, the simulation agent performs a random walk of the Web site. The accessibility metric is based upon the hyperlink structure of the site and the size of content; whereas an analysis of the actual content is not performed. It neither simulates users with specific information needs, nor users who can perceive navigational choices and make navigational decisions. Moreover, WebCriteria's own research shows that its software has no correlation with user behavior [20].

The scent usage prediction algorithm (WUFIS, Web User Flow by Information Scent), while firmly based in Information Foraging theory, is similar to several information retrieval algorithms based on network inferences. Turtle and Croft proposed the use of Bayesian networks to model information retrieval problems [25]. They represent queries and documents on an inference network, which is similar to our approach. More recently, a number of efforts in the Web research community have concentrated on combining linkage information with user queries in order to rank search results [6,5,7,14,24]. Most similar to our approach, Chakrabarti et. al. [7] and Silva et. al. [24] proposed combining link-based with keyword-based pieces of evidence in a single information retrieval model. Chakrabarti's system uses the text surrounding a link as keyword-based evidences to determine a weight for each link analyzed. This weighting is then used to compute rankings of the retrieval results using a modified version of the Kleinberg algorithm [14].

While these algorithms are similar in nature to our work here, there are several important differences. First, we use the algorithms for the purpose of accurately simulating users flowing down various hyperlink document paths, which is markedly different from using these algorithms to rank or retrieve search results. Several steps in our algorithm only make sense in the context of computing probabilities of link traversal. For example, since all probabilities leaving from one page source should add up to 1, we have an added normalization step to ensure this. Moreover, in detailed sections below, we formulate an approach that combines Chakrabarti's approach of using surrounding text around a link with previous approaches of analyzing the whole page content.

Second, in this paper we also describe a novel companion algorithm (IUNIS, Inferring User Need by Information Scent) for extracting information need out of traversal paths utilizing the idea of information scent. There are a number of reasons why we would want to infer information needs. For example, there also has been some research on suggesting information pieces to users based on a user's traversal history through a hypertext document collection. Most notably, Alexa.com provides the "What's Related" button on the Netscape Browser by predicting related pages from the traversal histories of past users who had been at the given page [2]. However, they do not use traversal history to extract user information needs, but instead uses the history to simply suggest another page like it.

Understanding the information needs of users is important to analysts who are looking at the usage log data of a Web site. By understanding the information need of a Web site's users, the Web site can be redesigned to better suit the user tasks. For example, the method described here could also be used to determine the information goal of a user currently traversing a Web site, thus tailoring the Web site on-the-fly to better suit the task.

OUR APPROACH

Information Goals and Information Scent

Information foraging theory [17] has been developed as a way of explaining human information-seeking and sense-making behavior. Here we use the theoretical notion of *information scent* developed in this theory [15, 17] as the basis for our analysis techniques for predictive modeling.

Figure 1: Users forage for information by surfing along links. Snippets provide *proximal* cues to access *distal* content.

On the Web, users typically forage for information by navigating from page to page along Web links. The content of pages associated with these links is usually presented to the user by some snippets of text or graphic. Foragers use these *proximal* cues (snippets; graphics) to assess the *distal* content (page at the other end of the link).[2] Information scent is the imperfect, subjective perception of the value and cost of information sources obtained from proximal cues, such as Web links, or icons representing the content sources.

[2] Furnas referred to such intermediate information as "residue" [11].

Our assumption is that, for the purposes of many analyses, users have some information need, and their surfing patterns through the site are guided by information scent. That is, we assume that users make navigational choices not randomly, but based on some rationale. Given this framing assumption we have developed techniques for these two questions:

- *(The Prediction Question) Simulate user actions based on user needs.* Given an information need and some set of pages as starting points, we use information scent to predict the expected surfing patterns, and thereby simulate Web site usage.
- *(The Inferring Question) Infer user needs based on user actions.* For a particular pattern of surfing, we seek to infer the associated information need or goal.

The two algorithms (WUFIS and IUNIS) are designed to answer these two questions. These techniques are based on psychological models, and are closely related to Web data mining techniques based on the analysis of Content, Usage, and hyperlink Topology (CUT) [8,21]. In the following subsections, we will describe the algorithms.

Web User Flow by Information Scent (WUFIS)

For the Prediction Question, we are interested in simulating users arriving at Web sites with particular information needs. Given their intent and their entry point, we want to model their surfing behavior within the Web site. *Web User Flow by Information Scent* (WUFIS) is a new predictive modeling technique based on a combination of information retrieval techniques and spreading activation [3]. We analyze the quality of Web links in providing good scent that leads users to the content that they seek. Conceptually, our model works by calculating the probability that a user will flow down a particular Web link, given a specific information need.

Figure 2 shows the flowchart of the WUFIS process. First, we extract the content and linkages of a Web site. We obtain the hyperlink topology as an adjacency matrix $\boldsymbol{T}$. We also obtain the (word x document) $\boldsymbol{W}$ matrix. An entry in the $\boldsymbol{W}$ matrix specifies how often a word occurs in that document. We then calculate the TF.IDF weighting [23] of the words in the $\boldsymbol{W}$ matrix. TF.IDF (Term Frequency by Inverse Document Frequency) is a common information retrieval technique for calculating term importance of a word by weighting it against how frequently it occurs in the document collection [23, p.543]. This new weighting gives us a $\boldsymbol{W_{TF.IDF}}$ matrix. An entry $\boldsymbol{(i,j)}$ in the $\boldsymbol{W_{TF.IDF}}$ matrix tells us how important word $\boldsymbol{i}$ is in document $\boldsymbol{j}$.

The proximal scent of a link is calculated as a degree of similarity between the proximal cues and the information need. A user's information need is expressed as a weighted keyword[3] vector $\boldsymbol{Q}$. For each link, we obtain the proximal cues that are associated with that link, and put this information into a matrix $\boldsymbol{K}$. There are a variety of ways to obtain proximal cues. For example, we may look at (1) the words in the link itself, (2) the text surrounding a link [7], (2) the graphics related to a link, (3) the position of the link on the page, etc.

Using the user information need $\boldsymbol{Q}$, and $\boldsymbol{K}$, $\boldsymbol{T}$ and $\boldsymbol{W_{TF.IDF}}$, we can now calculate the scent of each link. We multiply the link cues matrix $\boldsymbol{K}$ against $\boldsymbol{Q}$ to obtain the proximal scent matrix $\boldsymbol{PS}$, which specifies the probabilities of users following each particular link.

Specifically, to do this, for each non-zero entry $\boldsymbol{(i,j)}$ in the $\boldsymbol{T}$ topology matrix, we look up the corresponding proximal cues in the $\boldsymbol{K}$ matrix. These proximal cues form a vector $\boldsymbol{K(i,j)}$, and we look up the weighting of each keyword in $\boldsymbol{K(i,j)}$ in the matrix $\boldsymbol{W_{TF.IDF}}$, thus forming a new keyword weighted vector $\boldsymbol{K'(i,j)}$. We then multiply $\boldsymbol{K'(i,j)}$ against the information need query vector $\boldsymbol{Q}$. This will give us a value $\boldsymbol{PS'(i,j)}$, which specifies the degree of similarity between a link's proximal cues and the information need $\boldsymbol{Q}$.

Once we have all the $\boldsymbol{PS'(i,j)}$'s, we have an un-normalized $\boldsymbol{PS'}$ matrix. To obtain the Proximal Scent matrix $\boldsymbol{PS}$, we normalize the $\boldsymbol{PS'}$ matrix so that each of the columns sums to 1. Each entry in the normalized Proximal Scent matrix $\boldsymbol{PS}$ specifies the probability of a user flowing down that link.

The above algorithm works reasonably well in practice. However, we know that we currently are not able to capture all of the proximal cues for a given link. In many cases, this is because image graphics that are linked to other pages are extremely difficult to analyze for their proximal cues. To fix this problem, whenever we are not able to extract proximal cues, we propose to substitute a "distal scent" in place of the proximal scent values.

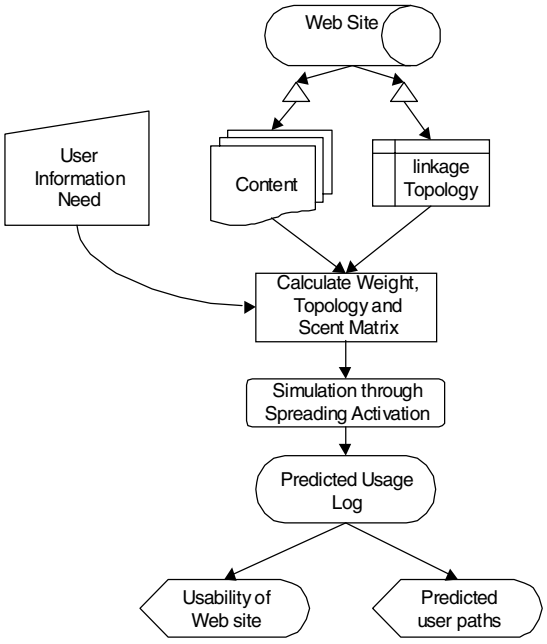

Figure 2: Flow chart of the Web User Flow by Information Scent (WUFIS) Algorithm

[3] We can also use a multi-modal vector that utilizing more than just keywords [12], but for the sake of simplicity here, we will restrict ourselves to just the keywords.

To calculate the distal scent of a link, we describe that link using words on the distal page (the page at the end of the link). For images, this works well because the words on the distal page perfectly describe the semantic of a hyperlinked image. Given that we have an imperfect capability to describe an image, this is a nice placeholder for these cases. To do this, we combine the Proximal scent ***PS*** matrix with the Distal scent matrix ***DS*** into a new scent matrix ***S*** by replacing the entries that are empty in the ***SP*** matrix with the corresponding entry in the ***DS*** matrix:

S = ***PS***;

For each entry ***S(i,j)*** in the ***S*** matrix that is zero:

Replace the entry ***S(i,j)*** with ***DS(i,j)***.

Renormalize ***S***.

To summarize, at the end of this process, we will have obtained a scent matrix ***S***, where each entry in ***S*** describes the probability a user will traverse down that link given a information need ***Q***.

Spreading Activation

At this point, we can then use the scent matrix to simulate users flowing down various links of a Web site, giving each link a different proportion of the users relative to the strength of the scent. The probability associated with each link essentially specifies the proportion of users that will flow down various link choices. We use a network flow algorithm called *spreading activation*. We take an entry page, and construct an entry vector E. We set the initial activation vector, $A(1) = E = [0\ \ 1\ \ 0\ \ 0\ \ 0\ \ 0\ \ 0]^T$

We can pump it through the scent matrix:

$$A(2) = S\,A(1) = [0\ \ 0\ \ 0.25\ \ 0.5\ \ 0\ \ 0.25\ \ 0]^T$$

To simulate multiple steps of flow, we just iterate the activation back through the ***S*** matrix. The algorithm goes through $t=1...n$ number of iterations:

$$\boldsymbol{A(t) = \alpha S\,A(t-1) + E.}$$

The final activation vector ***A(n)*** gives the result of the simulation.

The parameter α simulates the proportion of users that do not go from step *t-1* to step *t*. In the simulations reported here, we employed constant scalars $0 \le \alpha \le 1$. We can also use a step-dependent function $\alpha(t)$ that varies according to the "Law of Surfing" discovered in [13].

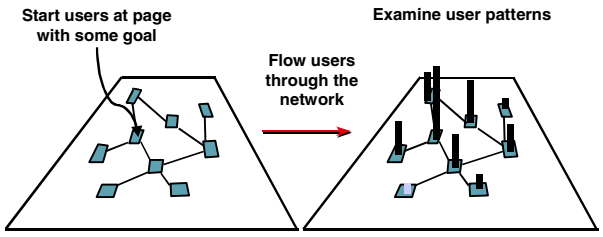

Figure 3: WUFIS and Spreading Activation

This generates a predicted usage log, which can be used to extract simulated user paths and infer the usability of a Web site.

Inferring User Need by Information Scent (IUNIS)

We will now propose the companion algorithm for the Inferring Question. We describe a method for inferring the information need of a user based on the user's traversal path through a hypertext collection. A user typically forages for information by making traversal decisions based on the user's task. For example, at any point in a traversal through the Web site, the user has expressed her interest in various pieces of information by the decision to traverse certain links. This user's traversal history is a list of documents that approximates the information need. The question is, therefore, given a traversal path through a hypertext collection, what can we say about the information need expressed by that traversal path?

In order to compute the information need of a traversal path, we make the following observation. The input to the model should be a list of documents and the order in which they were visited. The output is weighted keyword vector that describes the information need. Notice that this is the direct reverse of the simulation given by the WUFIS model formulated above. The input to the WUFIS is a weighted vector of keywords that describes the information needs, and the output is a list of documents that are visited by the simulated users. Therefore, intuitively, we need to only reverse the computation for WUFIS to obtain a list of user need keywords. So it seems we need to reverse the flow of activation to obtain our result. However, remember that the scent matrix ***S*** in the WUFIS model is computed with respect to a given information need. The scent matrix is already biased with a given information need!

Therefore, we cannot simply pump spreading activation through ***S***, because the associated strengths are already biased with a set of keywords describing a particular information need. Therefore, to obtain our IUNIS algorithm, instead of pumping activation through ***S***, we need to pump activation through the topology matrix ***T***. The resulting activation vector can then be multiplied with the weight matrix $\boldsymbol{W_{TF.IDF}}$ to obtain a set of keywords (along with a set of associated keyword weights).

Furthermore, we can be smarter about the way we use the path information. We can weight certain documents more heavily than others in the initial spreading activation vector. We can do this in several ways. First, we want to give the most weight to the most recently visited document, with the intuition that more recent documents tend to better describe the user's information need.

Second, we can weight the documents by page access TF.IDF weighting, in analogy to the TF.IDF computation of word weights in the WUFIS algorithm. In this case the term frequency corresponds to the access frequency of the page by the given user and the inverse document frequency corresponds to the ratio of total users to the number of distinct users who have accessed the given page. This helps to weight

down pages that are accessed by many users and may not be very relevant to the user's information need (e.g. a site's splash page).

Third, if a list of content pages (as opposed to index pages, for instance) is known ahead of time, we may weight the content pages more heavily than the index pages.

Figure 4 is the flow chart that describes this process. As shown in the Figure, we can obtain traversal paths by extracting user sessions out of the Web server usage logs. We note that the IUNIS method can also be applied to any weighted list of documents, such as using a predicted future path of a user to infer potential future information needs. Or the flow can be reversed to predict paths based on an ending position.

In the IUNIS algorithm especially, we should use the more sophisticated spreading activation algorithm with continuous pumping of source activation vector $\boldsymbol{E}$ and an α dampening factor: $\boldsymbol{A(t) = \alpha S\, A(t-1) + E.}$

Detailed Example

We now give a detailed example of the IUNIS inferring algorithm using the example Web collection in Figure 5. We index the full corpus of words in the following way:

0: Java 1: API 2: Sun 3: Home 4:Coffee

5: Support 6: Petes 7: Tea

Given the path *P0→P1→P3*, we can express the path as a vector with most recent documents weighted most heavily: $\boldsymbol{P} = [1\ \ 2\ \ 0\ \ 3\ \ 0\ \ 0\ \ 0\ ]$. We weight the elements of $\boldsymbol{P}$ by page access TF.IDF weighting:

$\boldsymbol{Q} = weighted(\boldsymbol{P})$

$= [2.72\ \ 7.39\ \ 0.0\ \ 20.09\ \ 0.0\ \ 0.0\ \ 0.0\ ]$

Without looking into the future, (that is, with 0 iterations on the spreading activation), we have a list of keywords as:

$A(1) = Q^T$

$R1 = \boldsymbol{W_{TF.IDF}} * A(1)$

$=[27.48\ \ 20.09\ \ 2.72\ \ 10.11\ \ 0.0\ \ 0.0\ \ 0.0\ \ 0.0]$

which gives us "Java" at 27.48 and "API" at 20.09 as the two top keywords, followed by "Home" at 10.11.

With one iteration of spreading activation:

$A(2) = 0.5\ T\ A(1) + E$

$= [6.4\ \ 17.4\ \ 10.0\ \ 20.1\ \ 0.0\ \ 3.7\ \ 0.0]$

$R2 = \boldsymbol{W_{TF.IDF}} * A(2)$

$=[51.3\ \ 20.1\ \ 6.4\ \ 23.9\ \ 3.7\ \ 10.0\ \ 0.0\ \ 0.0]$

which gives the top three as "Java" at 51.3 and "Home" at 23.9 and "API" at 20.1.

With two iterations:

$A(3) = 0.5\ T\ A(2) + E$

$R3 = WA(3)=[63.2\ \ 20.1\ \ 11.4\ \ 35.7\ \ 8.7\ \ 10.0\ \ 1.8\ \ 1.8]$

which gives "Java" and "Home" as top two, and "Sun" and "API" as the next two.

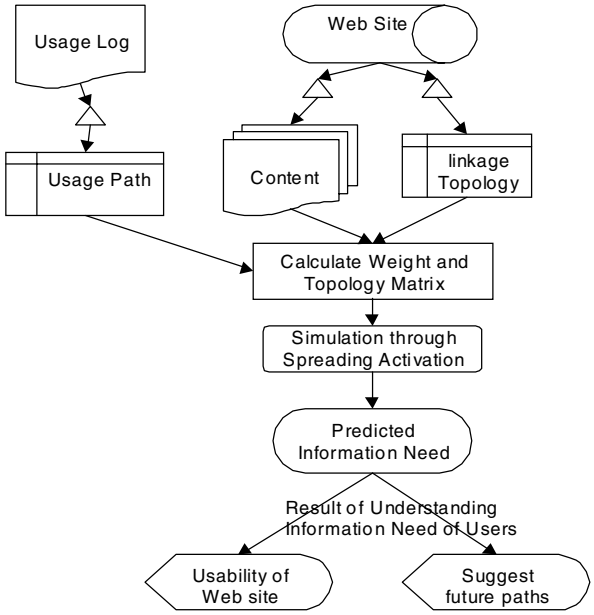

Figure 4: Flow chart for the Inferring User Need by Information Scent (IUNIS) algorithm

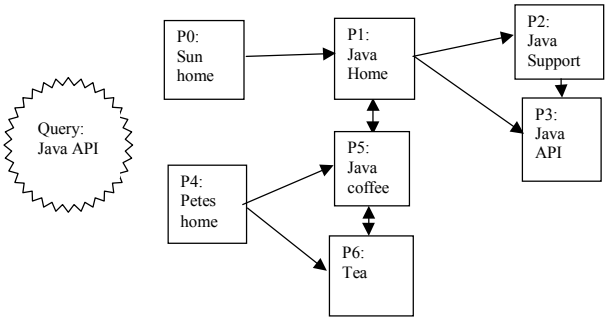

Figure 5: Example Web Collection

EVALUATION

In this section, we describe the empirical studies that we have carried out in order to validate our model and our algorithms. First, we present results of running the WUFIS algorithm on 19 different Web sites, and validate each of the simulations by human judges who view and blindly rate the outputs of the simulations. We also present an evaluation by human judges of the IUNIS algorithm outputs to show that it generates good keyword summaries.

Evaluation of WUFIS Simulated Paths

Method

Data: We chose 19 different Web sites of varying size and type. The sizes range from 27-12,000 pages, while site types range from Informational to E-commerce and Large Corporate sites. We've also varied the information need query vector from very general (looking for product information) to very specific (migraine headaches). Three different versions of the WUFIS algorithm are used on these Web sites: the algorithm using just the ***PS*** *proximal* scent matrix, just the ***DS*** *distal* scent matrix, and the fully *combined* ***S*** scent matrix. The ending position of spreading activation paths are generated, and the top ten URL positions are

extracted, as they represent the most relevant information to the query according to the algorithm.

Procedure: A human judge rates each [query, site, and URL]-tuple from a scale of 0-10, with 10 meaning that the URL is "most relevant" to the query, and 0 meaning "not relevant at all". These ratings are done blindly, so that the judge does not know which algorithm produced which URL. After all 570 (=10 URLs x 3 algorithms x 19 sites) rating are collected, the results are tabulated, and averaged to obtain the ratings for each algorithm for each site.

Results and Discussion

Table 1 presents a summary of these ratings of WUFIS simulation runs.

There are eleven cases when Proximal scent is extracted and available (marked by the existence of a rating in the Proximal column). In these ten cases, the Proximal scent matrix consistently outperformed Distal scent matrix. Eight out of these eleven cases, the Proximal scent matrix performed 9 or higher. This shows that the Proximal works well whenever we are able to extract it using our method. This gives us much confidence in the Proximal approach, and shows the need to continue developing good proximal cue extraction techniques.

With one exception, in all cases where both Proximal and Distal ratings are available, the Combined algorithm's performance is effectively the average of Proximal and Distal. We hypothesize that a good Proximal scent matrix is indicative of a site that is well designed with the proximal browsing cues, while a good Distal scent matrix indicate search engines would work well because the site is content-rich. These measurements may be indicative of a Web site's intrinsic usability.

Most importantly, we see that the average rating for Combined (7.54) is higher than Distal (6.69) and Proximal (5.20). In eight cases, we were not able to completely extract the Proximal scent because of problems such as links with only images as the proximal cue (marked by n/a). In those cases, we were able to fix these problem areas to achieve a Combined matrix that worked better than the Distal matrix alone. These facts show that the Combined scent matrix is able to use the Proximal scent when it available after extraction, but can use the Distal scent to fix problem areas when it is needed.

Evaluation of the IUNIS Path Summaries

We now turn our attention to an evaluation of inferring algorithm. IUNIS produces a set of keywords that are

Web Site	Category	# Docs	# Words	Query	Proximal	Distal	Combined
ArtShow.com	Museum/Gallery	1,063	13,655	Landscapes, Painting, Paintings, Landscape	9.1	5.1	7.3
Achoo.com	Informational	770	11,568	Migraines, Headache, Headaches, Treatment, Drugs	n/a	4.3	4.5
Healthgate.com	Informational	2,921	25,743	Migraine, Headache, Headaches, Treatment, Durgs	n/a	6.3	7.3
Healthfinder.gov	Informational	4,684	68,742	Migraine, Headache, Headaches, Treatment, Drugs	n/a	4.8	5.1
medweb.emory.edu/MedWeb	Informational	11,284	15,488	Migraine, Headache, Headaches, Treatment, Drugs	n/a	7	8.2
Inxight.com	Small Company	207	6,827	Product, Products, Reproduct, Productivity	10	6.4	7.4
Loudcloud.com	Small Company	103	5,624	Product, Products, Productivity, Reproduct	10	5.9	8.3
Aircourier.com	Small Company	27	933	September, Mexico, Cheap, Airfare	7	6.9	8.1
Snapple.com	Small Company	64	7,957	Twisted, Cap, Promotion, Promotions	10	8.1	8.7
Target.com	E-commerce	2,836	13,001	Clothes, Clothing, Accessory, Accessories, Apparel	10	7	9
Kmart.com	E-commerce	860	9,152	Clothes, Clothing, Accessory, Accessories, Apparel, Shirt, Shorts	9.2	8.2	9
PaloAltoDailyNews.com	Online news	412	14,985	Housing, Homes, Shortage, Homelessness, Homeless	n/a	8.4	8.4
Newsweek.com	Online news	1,969	61,414	Democratic, Issues, Gore, Lieberman	n/a	8.6	9
Trailplace.com	Message Board	232	2,786	Trail, Trails, Trailheads	7	6.5	7
VolunteerMatch.org	Nonprofit/Search	12,054	42,697	Children, Youth, Palo, Alto, Ca	10	4.4	4.4
Julliard.edu	Education	114	12,725	Admission, Admission, Applications, Applications, Procedure	9	8.7	8.5
Cs.umn.edu	Education	4,928	54,070	Collaborative Filtering	n/a	8.6	9.4
Canon.com	Large Corporation	1,254 (depth 8)	14,652	Printer, Printers	n/a	6	6.4
Xerox.com	Large Corporation	2,218	21,819	Product, Products	8.6	6.4	7.3
Average					**5.20**	**6.69**	**7.54**

Table 1: Analysis of WUFIS algorithm results on 18 different Web sites of varying sizes and types.

intended to indicate the core concepts involved in users' information needs, as expressed by their surfing paths. We performed an evaluation of the IUNIS keyword summaries with respect to their ability to communicate the content of user paths.

Method

Participants: Eight participants from Xerox PARC volunteered for this evaluation. All were members of the research staff or consultant researchers.

Materials: Ten paths identified by the LRS Significant Surfing Path algorithm [21] from the May 18, 1998 www.xerox.com data set were randomly selected from all LRS paths of length = 6 Web pages. Ten booklets were constructed, with each booklet containing the six Web pages associated with a path, in the order that the pages occurred on the path.

In addition, the top 20 keywords identified by the IUNIS algorithm for each of the ten paths were used to form ten path summaries. All ten of these 20-word path summaries were placed on a single rating sheet. Beside each path summary was a five-point Likert scale. A copy of this rating sheet was attached to each of the ten path booklets.

Procedure: The N = 8 participants were asked to read through each path booklet and to rate each of the path summaries with respect to the relevance of the summary as a description of the Web pages on the path. The ratings ranged from "1 = Not relevant" to "5 = Highly Relevant" with "3 = Neutral" as the midpoint. In addition, participants were asked to identify which of the 10 summaries was the best match to each of the path booklets.

Results and Discussion

The keyword summaries generated by IUNIS for a particular path (*matching summaries*) should be rated more relevant than the summaries generated for other paths (*non-matching summaries*). On the scale "1 = Not relevant" to "5 = Highly Relevant", the matching summary mean was 4.58 (median=5) and the non-matching summary mean was 1.97 (median=1). This difference was highly significant, $F(1,781) = 283.08$, $MSE = 1.73$, $p < .001$. This indicates that the path keyword summaries generated by the IUNIS algorithm were judged to be very good representations of the path content.

An even stronger test is provided by the participants' selection of "the best summary" for each of the paths. On average, participants chose the IUNIS summary as the best match 5.6 times out of 10 (S.D. = 1.3; chance selection = 1 time out of 10). A measure of the degree of match is provided by Cohen's kappa statistic, which ranges like a correlation coefficient (in this case, kappa = 0 indicates random association and kappa = 1 equals perfect association). Cohen's kappa = 0.51 for the degree of match between participants selections of "best summary" and the IUNIS summary. This indicates a good match between participants' selection and IUNIS.

Overall, this evaluation yielded strong evidence that the IUNIS algorithm generates sets of keywords that judges evaluate as good summaries of WWW paths.

Conclusion

In this paper, we have outlined two computational methods for modeling user needs. Our algorithms are based on the concept of information scent. First, given a user information need, we described an algorithm that simulates usage of the Web. Second, given a user path through the Web, we described an algorithm that infers the information needs expressed by that path.

In the WUFIS user flow simulation, our assumption is that users will make navigational decisions guided by scent. This assumption is derived from the fact that users make navigational decision rationally rather than randomly. Our initial experiments with Web sites present evidences that the algorithm works well. Moreover, detailed case studies have been examined in [9], which shows several real-world scenarios where WUFIS matched actual usage.

In the IUNIS inferring user need algorithm, our central assumption is that the pages that the user has already seen are an approximation of the kinds of information that the user is interested in. In our experiments, we show that the IUNIS algorithm extracts keywords that are representative of the content on the user path. Furthermore, it is worth noting that the IUNIS algorithm does not depend on the fact that the user path actually contains the information that the user is seeking, but merely that it contains the *kind* of information that the user is seeking.

There is a huge predicted utility to being able to simulate Web site usage and infer user information goals. Here we briefly mention a few applications:

- *Personalize Web environments.* We have used this research in a surf-along tool that suggests pages to Web users based on their interest profile. Imagine a system watching a series of visits that are made by a user: if the system can infer the information goal of the user, then it can use this information to tailor the information environment to the task [15].
- *Help design Web site.* We've used this research to understand the usability of a Web site [9]. By helping Web site analysts understand the information goals of users who visit the Web site, we help them understand how well the Web site is supporting the users in achieving their goals.
- *Help identify parts of a Web site as bad designs.* Given real usage paths, we envision that using our methods could automatically flag parts of a Web site where users have a hard time finding information based on their information needs.

Clearly, this is not an exhaustive list. We believe this is a useful tool as part of a Web analysis toolkit that will enable researchers to better understand the usage of the Web, designers to better design their Web sites, and end-users to seek information more efficiently.

Acknowledgement
Kim Chen was supported under a summer 2000 internship at Xerox PARC. Allison Woodruff, Chris Olston, Pam Schraedley helped proof-reading this paper, and providing valuable feedback. This research was supported in part by an Office of Naval Research grant No. N00014-96-C-0097 to Peter Pirolli and Stuart Card.

REFERENCES

1. Accrue Insight. (1999) http://www.accrue.com
2. Alexa Internet. (1999) http://www.alexa.com
3. Anderson, J. R., Pirolli, P. L. (1984) Spread of Activation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 10, 791-798.
4. Astra SiteManager. (1999) http://www.merc-int.com
5. Bharat, K. and Henzinger, M. R. (1998) Improved algorithms for topic distillation in a hyperlinked environment. In *Proc. of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 104-111).
6. Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual web search engine. In *Proc. Of the 7th International World Wide Web Conference (WWW7)* (pp. 107-117), Brisbane, Australia.
7. Chakrabarti, S., B. Dom, P. Raghavan, S. Rajagopalan, D. Gibson, and J. Kleinberg. (1998) Automatic resource compilation by analyzing hyperlink structure and associated text. In *Proc. Of the 7th International World Wide Web Conference (WWW7)* (pp. 65-74), Brisbane, Australia.
8. Chi, E.H., Pitkow, J., Mackinlay, J., Pirolli, P., Gossweiler, R., and Card, S. (1998). Visualizing the Evolution of Web Ecologies. *Proceedings of the Human Factors in Computing Systems, CHI '98.* (pp. 400-407). Los Angles, CA.
9. Chi, E. H., Pirolli, P., Pitkow, J. (2000) The Scent of a Site: A System for Analyzing and Predicting Information Scent, Usage, and Usability of a Web Site. Proceedings of Human Factors in Computing Systems, CHI 2000. (pp. 400-407). Hague, Netherlands.
10. Cooley, R., Mobasher, B., Srivastava, J. (1997) Web Mining: Information and Pattern Discovery on the World Wide Web (A Survey Paper), in *Proc. of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*. Nov. 1997.
11. Furnas, G.W. (1997) Effective view navigation. *Proceedings of the Human Factors in Computing Systems, CHI '97* (pp. 367-374), Atlanta, GA.
12. Heer, J., and Chi, E.H. Identifying Web user types using multi-modal clustering. (submitted for publication)
13. Huberman, B. A., Pirolli, P., Pitkow, J., Lukose, R. (1998) Strong Regularities in World Wide Web Surfing. *Science*, 280, 95-97.
14. Kleinberg, J. M. (1998) Authoritative sources in a hyperlinked environment. In *Proc. Of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms*, (pp. 668-677), San Francisco, CA.
15. Olston, C., Chi, E. H. (2000) ScentTrails: Integrating Browsing and Searching on the World Wide Web. (Submitted for publication)
16. Pirolli, P. (1997) Computational models of information scent-following in a very large browsable text collection. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '97* (pp. 3-10), Atlanta, GA.
17. Pirolli, P. and Card, S.K. (1999) Information foraging. *Psychological Review 106(4)* (pp. 643-675).
18. Pirolli, P., Pitkow, J., and Rao, R. (1996) Silk from a sow's ear: Extracting usable structures from the web. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '96* Vancouver, Canada.
19. Pirolli, P. and Pitkow, J.E. (1999) Distributions of surfers' paths through the World Wide Web: Empirical characterization. *World Wide Web, 1*, 1-17.
20. Pirolli, P. (2000) A Web site user model should at least predict something about users. Internetworking, 3:1. http://www.sandia.gov/itg/newsletter/mar00/critique_max.html
21. Pitkow, J. and Piroll, P. (1999) Mining longest repeated subsequences to predict World Wide Web surfing. *Proceedings of the USENIX Conference on Internet.*
22. Pitkow, J. and Pirolli, P. (1997) Life, death, and lawfulness on the electronic frontier. *Proceedings of the Conference on Human Factors in Computing Systems, CHI '97* (pp. 383-390).
23. Schuetze, H., Manning, C. (1999) Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.
24. Silva, I., B. Ribeiro-Neto, P. Calado, E. Moura, N. Ziviani. (2000) Link-based and Content-based Evidential Information in a Belief Network Model. In *Proc. of the 21st ACM SIGIR Conference on Research and Development in Information Retrieval* (pp.96-103). Athens, Greece.
25. Turtle, H., Croft, W. (1991) Evaluation of an inference network-based retrieval model. *ACM Transactions on Information Systems*, 9(3):187-222
26. WebCriteria SiteProfile. (1999) http://www.webcriteria.com