The Use of Proximal Information Scent to Forage for Distal Content on the World Wide Web

Peter Pirolli
PARC

To appear in
Alex Kirlik (Ed.) *Working with Technology in Mind: Brunswikian Resources for Cognitive Science and Engineering*.

The legacy of the Enlightenment is the belief that entirely on our own we can know, and in knowing, understand, and in understanding, choose wisely…

Thanks to science and technology, access to factual knowledge of all kinds is rising exponentially while dropping in unit cost…We are drowning in information, while starving for wisdom.

- E.O. Wilson (1998)

INTRODUCTION

Information Foraging Theory addresses how people will adaptively shape their behavior to their information environments, and how information environments can best be shaped to people. It is a psychological theory explicitly formulated to deal with the analysis of adaptation (good designs) and explicitly formulated to deal with how external content is used to gain valuable knowledge. This symmetrical focus on users and their information system environments has been inspired in many ways by methods and concepts championed in Brunswik's (1952) probabilistic functionalism. This chapter summarizes Information Foraging research on use of the World Wide Web (hereafter, the *Web*). This research involves concepts that are akin to ones developed by Brunswik including the use of *representative design*, in which realistic Web tasks are studied, a variant of a *Lens Model* of how people judge the relevance of navigation cues on the Web, and the use of an *idiographic-statistical* approach to model aspects of human cognition.

The particular focus of this chapter will be on a psychological theory of *information scent* (Pirolli, 1997, 2003; Pirolli & Card, 1999) that is embedded in a broader model (Pirolli & Fu, 2003) of information foraging on the Web. The notion of information scent also has been used in developing models of people seeking information in document-clustering browsers (Pirolli, 1997) and highly interactive information visualizations (Pirolli, Card, & Van Der Wege, 2003). Information scent refers to the cues used by information foragers to make judgments related to the selection of information sources to pursue and consume. These cues include items such as Web links or bibliographic citations that provide users with concise information about content that is not immediately available. The information scent cues play an important role in guiding users to the information they seek, and they also play a role in providing users with an overall sense of the contents of collections. The purpose of this paper is to present a theoretical account of information scent that supports the development of models of navigation choice.

Information Foraging Theory is an example of a recent flourish of theories in *adaptationist psychology* that draw upon evolutionary-ecological theory in biology (e.g., Anderson, 1990; Cosmides, Tooby, & Barow, 1992; Gigerenzer, 2000). The framework of adaptationist psychology involves the analysis of the structure of the environments faced by people and analysis of the design (or engineering) rationale of the psychological mechanisms and knowledge used by people to perform adaptively in those environments. Such theories are developed through methodological adaptationism that assumes that it is a good heuristic for scientists to presuppose that evolving, behaving systems are rational, or well-designed, for fulfilling certain functions in certain environments. In other words, there is an assumption of *ecological rationality* regarding the behavior of the system being observed (Bechtel, 1985; Dennett, 1983, 1988; Dennett, 1995; Todd & Gigerenzer, 1999). Many (e.g., Anderson, 1990; Dennett, 1995; Marr, 1982) have argued that in order to understand the cognitive mechanisms underlying behavior one has to begin by defining its function in relation to the environment. The adaptationist approach involves a kind of reverse engineering in which the analyst asks (a) *what* environmental problem is solved, (b) *why* is a given system a good solution to the problem, and (c) *how* is that solution realized by mechanism. This simultaneous concern with environment and organism echoes the approach of Brunswik. Information foraging theory (Pirolli & Card, 1999) is an adaptationist theory that assumes that human information seeking mechanisms and strategies adapt to the structure of the information environments in which they operate.

Next I present some aspects of the Web environment that will be relevant to developing a model of how people assess information scent cues in order to navigate.[1] Users' navigation in the Web environment can be seen as involving assessments of proximal information scent cues in order to make action choices that lead to distal information sources. This view is a variant of Brunswik's Lens Model (Brunswik, 1952). This Lens Model view is then used to frame the development of a stochastic model of individual utility judgment and choice that

[1] Others (e.g., Baldi, Frasconi, & Smyth, 2003; Barabàsi, 2002) provide additional analyses of the regularities in the structure and growth of the Web.

derives from a Bayesian analysis of the environment. This Bayesian analysis motivates a set of cognitive mechanisms based on spreading activation (Anderson, 1990; Anderson & Milson, 1989), and a decision model based on the Random Utility Model (RUM, McFadden, 1974, 1978). The general argument is that information foragers operate in an ecologically rational manner to make choices based on their predictive judgments (under uncertainty) based on information scent. This information scent model (spreading activation plus RUM) of the navigation choice behavior of individuals can be seen as a variant of the idiographic-statistical approach advocated by Brunswik (1952). Finally, I summarize a model of Web foraging (Pirolli & Fu, 2003) that addresses data (Card et al., 2001; Morrison, Pirolli, & Card, 2001) collected from Web users working on Web tasks in a study that attempted to follow Brunswik's tenets of representative design.

ASPECTS OF THE STRUCTURE OF THE WORLD WIDE WEB

Task Environment and Information Environment

It is useful to think of two interrelated environments in which an information forager operates: The *task environment* and the *information environment*. The classical definition of the task environment is that it "refers to an environment coupled with a goal, problem or task—the one for which the motivation of the subject is assumed. It is the task that defines a point of view about the environment, and that, in fact allows an environment to be delimited" (Newell & Simon, 1972, p. 55). The task environment is the scientist's analysis of those aspects of the physical, social, virtual, and cognitive environments that drive human behavior.

The information environment is a tributary of knowledge that permits people to more adaptively engage their task environments. Most of the tasks that we identify as significant problems in our everyday life require that we get more knowledge—become better informed—before taking action. What we know, or do not know, affects how well we function in the important task environments that we face in life. External content provides the means for expanding and improving our abilities. The information environment, in turn, structures our interactions with this content. Our particular analytic viewpoint on the information environment will be determined by the information needs that arise from the embedding task environment.

When people interact with information through technology, those technological systems act as *cognitive prostheses*[2] that more or less help us intelligently find and use the right knowledge at the right time. People shape themselves and their cognitive prostheses to be more adaptive in reaction to their goals and the structure and constraints of their information environments. A useful abstract way of thinking about such adaptation is to say that

> *Human-information interaction systems will tend to maximize the value of external knowledge gained relative to the cost of interaction.*

Schematically, we may characterize this maximization tendency as

$$\max\left[\frac{\text{Expected value of knowledge gained}}{\text{Cost of interaction}}\right] \tag{1}$$

This hypothesis is consistent with Resnikoff's (1989, p. 97) observation that natural and artificial information systems evolve towards stable states that maximize gains of valuable information per unit cost (when feasible). Cognitive systems engaged in information foraging will exhibit such adaptive tendencies.

Probabilistic Texture

In contrast to application programs such as text editors and spreadsheets, in which actions have fairly determinate outcomes[3], foraging through a large volume of information involves uncertainties—for a variety of reasons—about the location, quality, relevance, veracity and so on, of the information sought and the effects of foraging actions. In other words, the information forager must deal with a *probabilistically textured* information environment. The ecological rationality of information foraging behavior must be analyzed through the theoretical lens and tools appropriate to decision making under uncertainty. The determinate formalisms and determinate cognitive mechanisms that are characteristic of the

[2] This term is due to Ken Ford (personal communication, June 2003).
[3] Barring bugs, of course.

traditional approach to cognitive models in Human Computer Interaction (e.g., Card, Moran, & Newell, 1983) are inadequate for the job of theorizing about information foraging in probabilistically textured environments. Models developed in Information Foraging Theory draw upon probabilistic models, and especially Bayesian approaches, and bear similarity to economic models of decision making (rational choice) under uncertainty and engineering models.

A Lens Model Framework for Information Scent

Throughout the foraging literature, it is assumed that organisms base their foraging decisions on predictions about the typology and utility of habitat and the typology and utility of food items (e.g., species of prey or plants). There is considerable evidence that humans show preference for landscape features (aesthetics) that appear to predict high resource-providing environments (e.g., a savanna), as well as features associated with ease of exploration (Orians & Heerwagen, 1992). I assume that the information forager must also base decisions on assessments of information habitats and information items, and that these assessments are based on learned utility judgments. Information scent refers to those environmental cues that feed into the utility judgments made by information foragers,

Figure 1 presents some examples of information scent cues. Figure 1a is a typical page generated by a Web search engine in response to a user query. The page lists Web pages (search results) that are predicted to be relevant to the query. Each search result is represented by its title (in blue), phrases from the hit containing words from the query (in black), and a URL (in green). Figure 1b illustrates, an alternative form of search result representation that is provided by *relevance enhanced thumbnails* (Woodruff, Rosenholtz, Morrison, Faulring, & Pirolli, 2002), which combine thumbnail images of search results with highlighted text relevant to the user's query. Figure 1c is a Hyperbolic Tree browser[4] (Lamping & Rao, 1994; Lamping, Rao, & Pirolli, 1995). Each label on each node in the Hyperbolic Tree is the title for a Web page. Finally, Figure 1d presents the Microsoft Explorer browser with a typical file system view. Each label on a node represents a folder, file, or application.

[4] Marketed by Inxight Inc. as the Star Tree

Figure 1. Examples of proximal information scent cues.

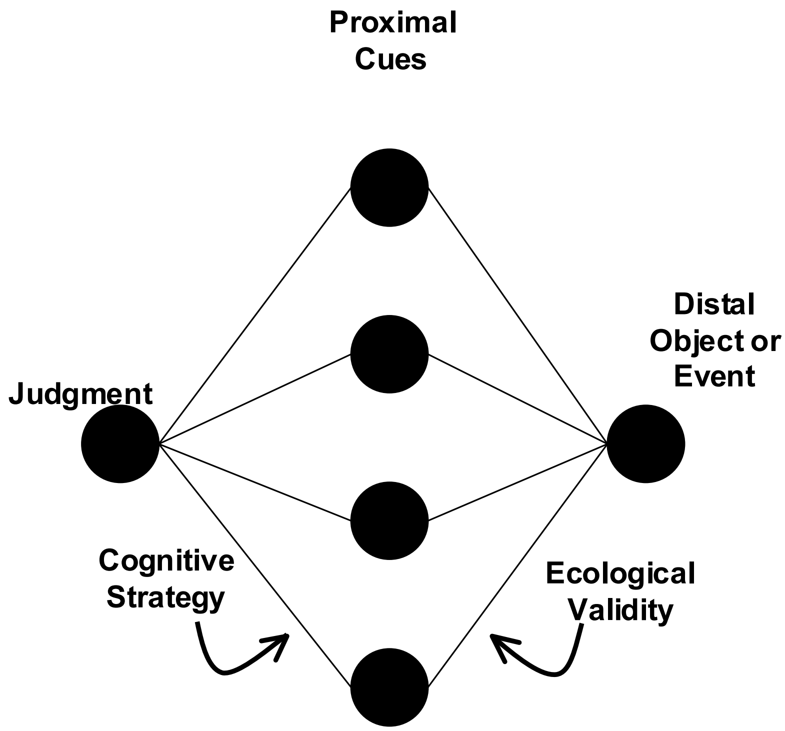


Figure 2. Brunswik's Lens Model

Egon Brunswik's (1956) ecological Lens Model can be used to frame our thinking about examples of information scent, such as those in Figure 1. The Lens model (Figure 2) characterizes the judgment problems facing organisms in probabilistically textured environments. In each example in Figure 1, the *distal* sources of information (e.g., Web pages, articles) are represented by some set of mediating cues (link summaries; relevance enhanced thumbnails; node labels; bibliographic citations). On the basis of these *proximal* cues, the user must make judgments about what is available and potential value of going after the distal content.[5] This distinction between proximal cues and distal objects is a well-known aspect of Brunswik's (1956) ecological theory of perception and judgment in a probabilistically textured environment. In Brunswik's Lens Model (Figure 2) the perception or judgment of a distal object or event is indirect and must be made on the basis of proximal cues. Brunswik (1956) advocated a detailed

---

[5] Furnas (1997) developed a theory of user interface navigation that uses notion of *residue*, which is very similar to the notion of proximal cues. Furnas develops a formal theory for the analysis of different kinds of user interaction techniques largely based on discrete mathematics.

analysis of the probabilistic relationships between observer and proximal cues (*cognitive strategy*) and between proximal cues and distal object or events (*ecological validity*).

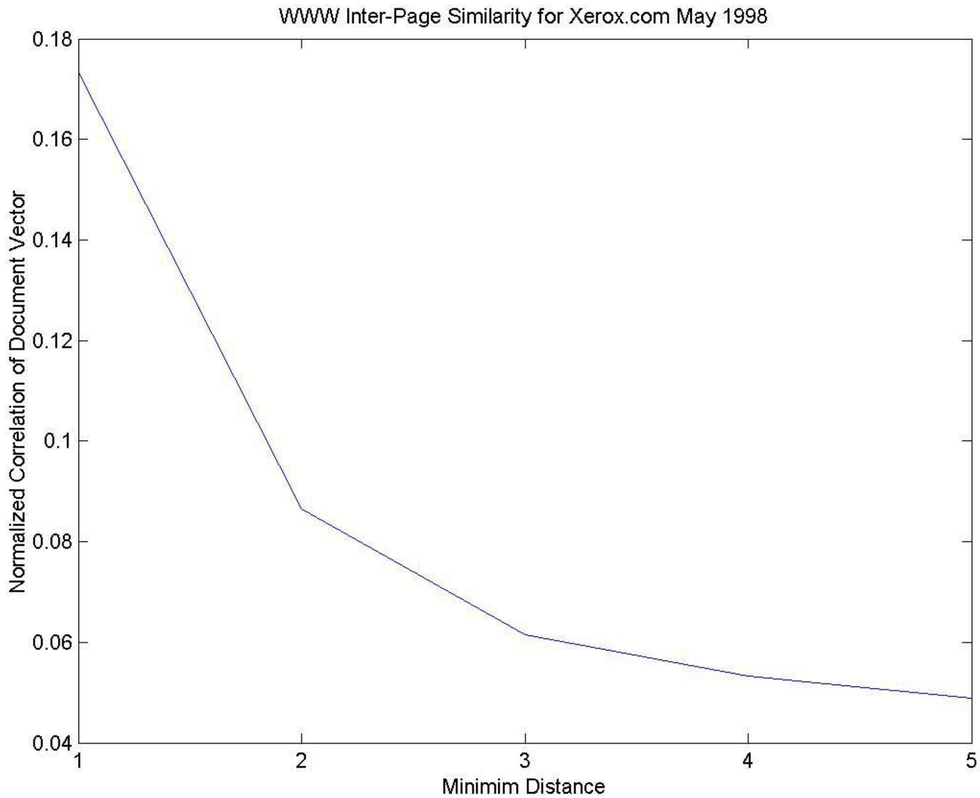


Figure 3. The similarity (normalized correlation of document word frequency vectors) of pairs of Web pages as a function of the minimum link distances separating the pairs. Data collected from the Xerox.com WWW site, May 1998.

Topical Patches and Diminishing Returns

Web users often surf the Web seeking content related to some topic of interest, and the Web tends to be organized into topical localities. Davison (2000) analyzed the topical locality of the Web using 200,998 Web pages sampled from approximately 3 million pages crawled in 1999. Davison assessed the topical similarity of pairs of Web pages from this sample that were: Linked (had a link between them), Siblings (linked from the same parent page), and Random (selected at random). The similarities were computed by a normalized correlation[6] or cosine measure, *r*, on the vectors of the word frequencies in a pair of documents (Manning & Schuetze, 1999).[7] The Linked pages showed greater textual similarity ($r = .23$) than Sibling pages ($r = .20$), but both were substantially more similar than Random pairs of pages ($r =.23$).

Figure 3 shows how topical similarity between pages diminishes with the link distance between them. To produce Figure 3 I used data collected from the Xerox.com Web site in May, 1998 (used for another purpose in, Pitkow & Pirolli, 1999) and I computed the page-to-page content similarities for all pairs of pages at minimum distances of 1, 2, 3, 4, and 5 degrees of separation. The similarities were computed by comparing normalized correlations of vectors of the word frequencies in a pair of documents (Manning & Schuetze, 1999). Figure 3 shows that the similarity of the content of pages diminishes rapidly as a function of shortest link distance separating them. Figure 3 suggests that the Web has topically related information patches.

Of course, what users actually see as "links" on the Web are sets of proximal cues such as those in Figure 1. Davison (2000) compared elaborated anchor text (the anchor plus additional surrounding text, having a mean of 11.02 terms) to a paired document that was either Linked (the page linked to the anchor) or Random (a random page). The normalized correlation (cosine) similarities were Linked $r = .16$ and Random $r \approx 0$. Davison's analysis of the correlation of proximal cues to distal content confirms our intuition that the cues have ecological validity (Figure 2).

Information scent cues described are expected to be useful to the information forager in at least two ways. First, information scent provides a way of judging the utility of following alternative paths (i.e., choosing a link from a set of links presented on a Web page). Second, information scent provides a way of detecting that one is moving out of a patch of topical relevance (i.e., by detecting that the quality of information scent is dwindling as in Figure 3).

---

[6] Manning and Schuetze (1999) show a mapping between the normalized correlation and a Bayesian analysis of the log likelihood odds of a document being relevant given a set of word cues representing the interest of a user.
[7] Davison used two additional measures that yielded similar results.

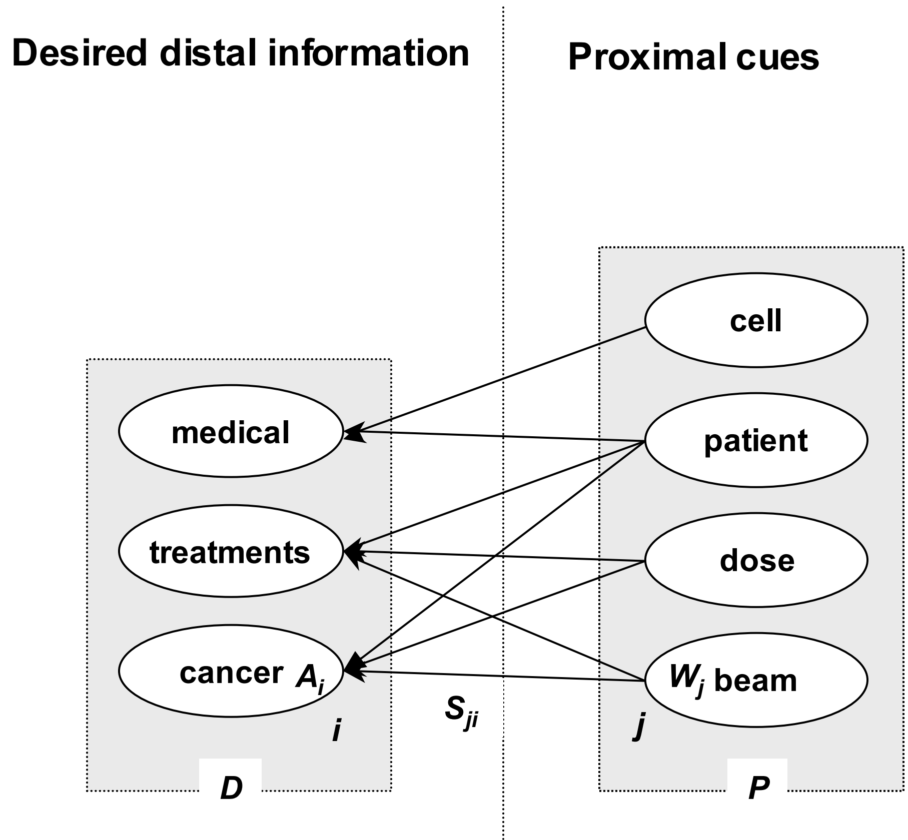


Figure 4. An examplar-based cognitive structure.

THE THEORY OF INFORMATION SCENT

Anderson (Anderson, 1990; Anderson & Milson, 1989) presented a rational analysis of human memory that serves as one basis for the theory of information scent presented here. Anderson and Milson (1989) proposed that human memory is designed to solve the problem of predicting what past experiences will be relevant in ongoing current proximal contexts, and allocating resources for the storage and retrieval of past experiences based on those predictions. Anderson and Milson developed a mathematical formulation that assumes that current proximal *context* factors and past *history* factors combine in forming those predictions about the relevance of past experience. Here, I assume that stored past experiences are retrieved, based on proximal features of the current context, and then used to predict the likelihood of distal features.[8]

Figure 4 presents an example for the purposes of discussion in this section. Figure 4 assumes that a user has the goal of finding distal information about medical treatments for cancer, and encounters a hypertext link labeled with the text that includes “cell”, “patient”, “dose”, and “beam”. The user’s cognitive task is to predict the likelihood of the desired distal information from the proximal cues available in the hypertext link labels.

Bayesian Analysis of Memory-based Prediction of Distal Features

Anderson and Milson (1989) applied Bayes Theorem to analyze the problem posed to human memory by the environment. By analogy, Bayes Theorem can be applied to the information foraging problem posed by situations such as those in Figure 4. The probability of a distal structure of features, *D*, given a structure of proximal features, *P*, can be stated using Bayes Theorem as,

$$\Pr(D|P) = \Pr(D) \cdot \Pr(P|D), \tag{2}$$

where $\Pr(D|P)$ is the *posterior probability* of distal structure *D* conditional on the occurrence of proximal structure *P*, $\Pr(D)$ is the *prior probability* (or *base rate*) of *D*, and $\Pr(P|D)$ is the *likelihood* of *P* occurring conditional on *D*. It is mathematically more tractable to conduct the analysis using *log odds*. The odds version of Bayes Theorem in Equation 2 is

$$\frac{\Pr(D \mid P)}{\Pr(D \mid \sim P)} = \frac{\Pr(D)}{\Pr(\sim D)} \cdot \frac{\Pr(P \mid D)}{\Pr(P \mid \sim D)}, \tag{3}$$

where $\Pr(D|{\sim}P)$ is the probability of distal structure *D* conditional on a context in which proximal structure *P* does not occur, $\Pr({\sim}D)$ is the prior probability of *D* not occurring, and $\Pr(P|{\sim}D)$ is the probability of *P* occurring given that *D* does not occur. With some simplifying assumptions and equation manipulation we can specify an odds equation for each individual feature *i* of the distal structure *D* and each individual feature *j* of the proximal structure,

[8] This move from predicting information needed in memory to predicting distal unobserved features in the current context extends the analysis of Anderson and Milson (1989).

$$\frac{\Pr(i \mid P)}{\Pr(\sim i \mid P)} = \frac{\Pr(i)}{\Pr(\sim i)} \cdot \prod_{j \in P} \frac{\Pr(i \mid j)}{\Pr(i)}, \tag{4}$$

where Pr(i|j) is the conditional probability of a distal feature i occurring given that a proximal feature j is present, Pr(~i|j) is the posterior probability of i not occurring when j is present, Pr(i) is the prior probability (base rate) of distal feature i occurring, Pr(~i) is the probability of i not occurring, and Pr(j|i) is the likelihood of proximal feature j occurring given distal feature i. Taking the logarithms of both sides of Equation 4 leads to an additive formula,

$$\log\left[\frac{\Pr(i \mid j)}{\Pr(\sim i \mid j)}\right] = \log\left[\frac{\Pr(i)}{\Pr(\sim i)}\right] + \sum_j \log\left[\frac{\Pr(j \mid i)}{\Pr(i)}\right] \tag{5}$$

or

$$A_i = B_i + \sum_j S_{ji}, \tag{6}$$

where

$$A_i = \log\left[\frac{\Pr(i \mid j)}{\Pr(\sim i \mid j)}\right],$$

$$B_i = \log\left[\frac{\Pr(i)}{\Pr(\sim i)}\right],$$

$$S_{ji} = \log\left[\frac{\Pr(j \mid i)}{\Pr(i)}\right].$$

This Bayesian analysis specifies the prediction problem facing the information forager in a probabilistically textured environment. The argument is that a rational (adaptive) solution on the part of the information forager would be to develop mechanisms and strategies that implement an approximation to the solution in Equation 6.

Mapping the Bayesian Rational Analysis to Spreading Activation

Having specified a Bayesian analysis of information scent in an ecologically rational way, we can illustrate the proposed cognitive mechanisms for performing these computations. Pirolli (1997) proposed a spreading activation model of information scent. Spreading activation models are neurally inspired models that have been used for decades in simulations of human memory (e.g., Anderson, 1976; Anderson & Lebiere, 2000; Anderson & Pirolli, 1984; Quillan, 1966). In such models, activation may be interpreted metaphorically as a kind of mental energy that drives cognitive processing. Activation spreads from a set of cognitive structures that are the current focus of attention through *associations* among other cognitive structures in memory. These cognitive structures are called *chunks* (Anderson & Lebiere, 2000; Miller, 1956; Simon, 1974).

Figure 4 presents a scenario for a spreading activation analysis. The chunks representing proximal cues are presented on the right side of Figure 4. Figure 4 also shows that there are associations between the goal chunks (representing needed distal information) and proximal cues (the link summary chunks). The associations among chunks come from past experience. The strength of associations reflects the degree to which proximal cues predict the occurrence of unobserved features. For instance, the word "medical" and "patient" co-occur quite frequently and they would have a high strength of association. The stronger the associations (reflecting greater predictive strength) the greater the amount of activation flow. These association strengths are reflections of the log likelihood odds developed in Equation 6.

Previous cognitive simulations (Pirolli, 1997; Pirolli & Card, 1999) used a spreading activation model derived from the ACT-R theory (Anderson & Lebiere, 2000). The activation of a chunk *i* is

$$A_i = B_i + \sum_j W_j S_{ji} , \tag{7}$$

where B_i is the base-level activation of *i*, S_{ji} is the association strength between an associated chunk *j* and chunk *i*, and W_j is reflects attention (*source activation*) on chunk *j*. Note that Equation 7 reflects the log odds Equation 6 but now includes a weighting factor *W* that characterizes capacity limitations of human attention. One may interpret Equation 7 as reflection of a Bayesian prediction of the likelihood of one chunk in the context of other chunks. A_i in Equation 7 is interpreted as reflecting the log posterior odds that *i* is likely, B_i is the log prior odds of *i* being likely, and S_{ji} reflects the log likelihood ratios that *i* is likely given that it occurs in the context of chunk *j*. This version of spreading activation was used in the past (Pirolli, 1997; Pirolli & Card, 1999) to develop models of information scent. The basic idea is that information scent cues in the world activate cognitive structures. Activation spreads from these cognitive structures to related structures in the spreading activation network. The amount of activation accumulating on the information goal provides an indicator of the likelihood of the distal features based on the proximal scent cues.

Learning the Strengths

As discussed in Pirolli and Card (1999), it is possible to automatically construct large spreading activation networks from on-line text corpora. In early research (Pirolli, 1997; Pirolli & Card, 1999), we derived these networks from the Tipster corpus (Harman, 1993). Unfortunately, the Tipster corpus does not contain many of the novel words that arise in popular media such as the Web. In the SNIF-ACT model (Pirolli & Fu, 2003) discussed below, we augmented the statistical database derived from Tipster by estimating word frequency and word co-occurrence statistics from the Web using a program that calls on the AltaVista search engine to provide data. Recently, Turney (2001) has shown that Pointwise Mutual Information (PMI) scores computed from the Web can provide good fits to human word similarity judgments, and PMI turns out to be equal to the inter-chunk association strength S_{ji} in Equation 7 (see Farahat and Pirolli, submitted for details).

The Random Utility Model

From proximal information scent cues, the user predicts the utility of a distal information source and makes choices based on the utilities of alternatives. This is achieved by spreading activation from the proximal cues (e.g., “cell”, “patient”, “dose”, and “beam” in Figure 4) and assessing the amount of activation that is received by the desired features of the information goal (e.g., “medical”, “treatments”, “cancer” in Figure 4). The information scent approach developed in Pirolli and Card (1999) is consistent in many ways with the Random Utility Model (RUM) framework (McFadden, 1974, 1978). For our purposes, a RUM consists of assumptions about (a) the characteristics of the information foragers making decisions, including their goal(s), (b) the choice set of alternatives, (c) the proximal cues (attributes) of the alternatives, and (d) a choice rule. For current purposes, we will assume a homogenous set of users with the same goal G with features, $i \varepsilon G$ (and note that there is much interesting work on RUMs for cases with heterogeneous user goals). Each choice made by a user concerns a set C

of alternatives, and each alternative J is an array of displayed proximal cues, $j \in J$, for some distal information content. Each proximal cue j emits a source activation W_j. These source activations spread through associations to features i that are part of the information goal G. The activation received by each goal feature i is A_i and the summed activation over all goal features is

$$\sum_{i \in G} A_i \ .$$

The predicted utility $U_{J|G}$ of distal information content based on proximal cues J in the context of goal G is:

$$U_{J|G} = V_{J|G} + \varepsilon_{J|G} \tag{8}$$

where

$$V_{J|G} = \sum_{i \in G} A_i$$

is the summed activation, and where $\varepsilon_{J|Q}$ is a random variable error term reflecting a stochastic component of utility. Thus, the utility $U_{J|Q}$ is composed of a deterministic component $V_{J|Q}$ and a random component $\varepsilon_{J|Q}$. RUM assumes utility maximization where the information forager with goal G chooses J if and only if the utility of J is greater than all the other alternatives in the choice set, i.e.,

$$U_{J|G} > U_{K|G} \text{ for all } K \in C.$$

Stated as a choice probability, this gives,

$$\Pr(J \mid G, C) = \Pr(U_{J|G} \geq U_{K|G}, \forall K \in C). \tag{9}$$

Because of the stochastic nature of the utilities $U_{K|G}$ it is not the case that one alternative will always be chosen over another.

The specific form of the RUM depends on assumptions concerning the nature of the random component ε_i associated with each alternative i. If the distributions of the ε_i are independent identically distributed Gumbel distributions (double exponential),[9] then Equation 9 takes the form of a multinomial logit

[9] This is also the assumption in ACT-R (Anderson & Lebiere, 2000).

$$\Pr(J \mid G, C) = \frac{e^{\mu V_{J|Q}}}{\sum_{K \in C} e^{\mu V_{K|Q}}}, \tag{10}$$

where μ is a scale parameter.[10] If there is only one alternative to choose (e.g., select *J* or do not select *J*) then Equation 9 takes the form of a binomial logit,

$$\Pr(J \mid Q, C) = \frac{1}{1 + e^{\mu V_{J|Q}}}. \tag{11}$$

Many off-the-shelf statistical packages (e.g., Systat, SPSS) include modules that can estimate the parameters of multinomial logistic models from discrete choice data.

Relating the Spread of Activation to Utility and Choice Probability

For a navigation judgment, we can now specify how the computation of spreading activation yields utilities by substituting Equation 7 into Equation 8:

$$\begin{aligned} U_{J|Q} &= V_{J|Q} + \varepsilon_{J|Q} \\ &= \sum_{i \in Q} A_i + \varepsilon_{J|Q} \\ &= \sum_{i \in Q} \left(B_i + \sum_{j \in J} W_j S_{ji} \right) + \varepsilon_{J|Q} \end{aligned} \tag{12}$$

As I noted above, the earlier specification (Pirolli & Card, 1999) of the discrete choice probability model in Equation 10 based on the spreading activation model of utility in Equation 12 was developed without knowledge of the RUM. It turns out that the ACT-R model of utility and choice (Anderson & Lebiere, 2000) is also a RUM, although the ACT-R model was also developed independent of RUM theory (J.R. Anderson, personal communication, July 2002). RUM is grounded in

---

[10] Note that in both Information Foraging Theory (Pirolli & Card, 1999) and ACT-R (Anderson & Lebiere, 2000) this equation was specified as a Boltzman equation with the substitution of 1/*T* for μ, where *T* is the “temperature” of the system.

classic microeconomic theory, and it has relations to psychological models of choice developed by Thurstone (1927) and Luce (1959). It's recent developments are associated with the Nobel Prize work of McFadden (1974; 1978). So, there are at least three reasons for casting information scent as a RUM:

1. There is prior empirical support for the applicability of the RUM to information scent (Pirolli & Card, 1999),
2. RUM provides a connection among models in psychology (e.g., ACT-R) microeconomic models of consumer preferences and choice, and human-information interaction models,
3. Because of it's connections to microeconomics and econometrics there is a considerable body of work on the statistical estimation of RUMs from choice and ratings data (e.g., Walker & Ben-Akiva, 2002), which could provide an influx of new measurement technology into cognitive science and human-computer interaction.

In the next section I summarize a study (Card et al., 2001) that illustrates, in a general way, the manner in which information scent controls the behavior of Web users. Then, I present a model (Pirolli & Fu, 2003) that uses the information scent mechanisms outlined in this section to simulate those Web users.

A LABORATORY STUDY OF WEB USERS

Card et al. (2001) conducted a study of Web in order to develop a base protocol analysis methodology and to begin to understand behavior and cognition involved in basic information foraging on the Web. These analyses were the basis for developing the SNIF-ACT computational cognitive model presented below. The procedure for the experimental design was partly inspired by the notion of representative design championed by Brunswik (1952; 1956).

Brunswik (1952; 1956) developed the method of representative design to insure that experimental results would be *ecologically valid*—that results would generalize from the laboratory to important tasks in real world. The basic idea behind this method is to utilize experimental conditions (materials, tasks, etc.) that reflect conditions of the world. Others (notably, Gibson, 1979; Neisser, 1976) have argued strongly that human psychology is exquisitely adapted to its environment, and consequently is best revealed through the study of tasks and stimuli representatively sampled from the environment.[11]

To develop a set of Web tasks for laboratory study, we (Morrison et al., 2001) conducted a survey to collect and analyze real-world Web tasks. We inserted a survey question into the GVU Tenth WWW User Survey (Kehoe, Pitkow, Sutton, & Aggarwal, 1999) conducted October-December 1998 by the Graphics, Visualization, and Usability Center at the Georgia Institute of Technology (henceforth, the *GVU Survey*). From 1994-1998 the GVU Survey collected information from on-line surveys on Internet demographics, culture, e-commerce, advertising, user attitudes, and usage patterns. The survey question (henceforth, the *Significance Question*) we used was a variation on those used in the Critical Incident Technique:

> *Please try to recall a recent instance in which you found important information on the World Wide Web; information that led to a significant action or decision. Please describe that incident in enough detail so that we can visualize the situation.*

The Critical Incident Technique originated in studies of aviation psychology conducted during World War II (Fitts & Jones, 1961), achieved wider recognition with the work of Flanagan (1954), and has evolved many variations in human factors (Shattuck & Woods, 1994), cognitive task analysis (Klein, Calderwood, & Macgregor, 1989), and usability (Hartson & Castillo, 1998), and Web use in particular (Choo, Detlor, & Turnbull, 1998). The key idea in this technique is to ask users to report a *critical incident*, which is an event or task that is a significant indicator (either positive or negative) of some factor of interest in the study. The Critical Incident Method, and its variants, provide a way of obtaining concrete descriptions of events or tasks that are identified as critical (important; significant) by typical people operating in real-world situations. It is not aimed at obtaining a random sample of tasks or events (which is a weak, general, method of understanding the world), but rather a sample of tasks or events that are revealing of the domain. As noted by Nielsen (2001) the use of Critical Incident

---

[11] See Gigerenzer (2000) for a discussion of psychology's focus on method and inferential statistics that is aimed at generalizing findings to populations of participants, but psychology's underdevelopment of complementary methodology and statistical machinery for generalizing findings to conditions of the world.

Technique on the Web is also useful in identifying important value-added tasks for Web provider and for gaining insights for innovations on the Web.

Table 1

Tasks used in the Web study.

Task Name	Task	Source Response to Significance Question
Java	You are a Java programmer working with the Java 2 Platform. You need to find out which Application Program Interfaces (APIs) have become obsolete. Find the site that describes all the Java 2 Platform v1.2.2 Deprecated APIs.	Reference the Java 1.2 API docs - I'm a Java programmer.
City	You are the Chair of Comedic events for Louisiana State University in Baton Rouge, LA. Your computer has just crashed and you have lost several advertisements for upcoming events. You know that The Second City tour is coming to your theater in the spring, but you do not know the precise date. Find the date the comedy troupe is playing on your campus. Also find a photograph of the group to put on the advertisement.	Searched and found (Using Yahoo Canada) for a comedy troupe web site to copy their photo for a poster to be printed and distributed locally for an upcoming community event."
Shop	You live at 3333 Coyote Hill Road, Palo Alto, CA 94304. After a particularly unpleasant experience at the supermarket you have decided that you would rather have someone else shop for you. Find out if there is an on-line supermarket in your area that will deliver local groceries to your home (i.e., not by UPS, FedEx, etc.).	Used the web to see if there was an on-line supermarket that could deliver groceries in my area
Haze	You are the Marketing Communications Manager for a small Silicon Valley firm. You are having problems with your printer and hope to find the solution on the web. You are printing out your document on a Canon 360PS (post-script) color printer. When it prints the graphics, a blue haze covers them. You do not know what is causing this problem: the printer driver, the post-script driver, or something else. You are printing from a computer running Microsoft Office 97 for Windows. Find the solution to this problem.	Found an updated driver for a printer that would not work correctly. Was able to get printer working within an hour of being presented with the problem.
Madonna	Your best friend has been a Madonna fan since middle school. For her birthday, you are interested in enrolling her in the official Madonna fan club, but you want to know the benefits. Find the web page that discusses all of the perks of being in the Madonna fan club.	I needed to find out when HANSON was coming out with their new album for my daughter. We went into their website and got the answer. We did this tonight.

Antz	After installing a state of the art entertainment center in your den and replacing the furniture and carpeting, your redecorating is almost complete. All that remains to be done is to purchase a set of movie posters to hang on the walls. Find a site where you can purchase the set of four Antz movie posters depicting the princess, the hero, the best friend, and the general.	Doing a little research on my boyfriend heritage and the history of the name "Gatling." I knew his great-grandfather had invented the Gatling Gun and the name had been passed down over generations. In a search engine, the word "Gatling" appeared as a movie. I looked up the movie, and went searching to purchase a movie poster from the movie "The Gatling Gun." It was not a popular movie, therefore the poster was difficult to find. I finally purchased one from a poster company (with a website) located in Boulder, CO. He believes it is the only GG movie poster around."

The survey yielded *N* = 2188 participant responses, and we developed a taxonomic analysis of the tasks that is presented in some detail in Morrison et al. (2001). A sample of tasks that involve finding a fact, product, software or document were selected for development into laboratory materials. Such tasks make up 25% of all the tasks reported in the survey, and it can be argued that finding information is a subgoal of any larger task conducted on the Web. The six tasks used in the laboratory study and the original survey responses on which they are based are presented in Table 1.

A total of *N* = 14 Stanford students participated in the study and worked on the six tasks in Table 1. The instrumentation and data analysis system that was used is depicted in Figure 5. Participants performed their Web tasks using a standard desk top computer running Windows 98 and the Internet Explorer Web browser on a 1024 X 768 pixel color monitor. As participants worked on Web tasks, a video recorder captured their think aloud protocols and the computer screen. A program called WebLogger (Reeder, Pirolli, & Card, 2001) collected and time-stamped all user interactions with the Internet Explorer Web browser including user keystrokes, mouse movements, scrolling, use of browser buttons, and all pages visited. WebLogger also collected and time-stamped all significant browser actions, including the retrieval and rendering of Web content. An eye tracker system was used to collect user eye movements, but those data will not be discussed here.

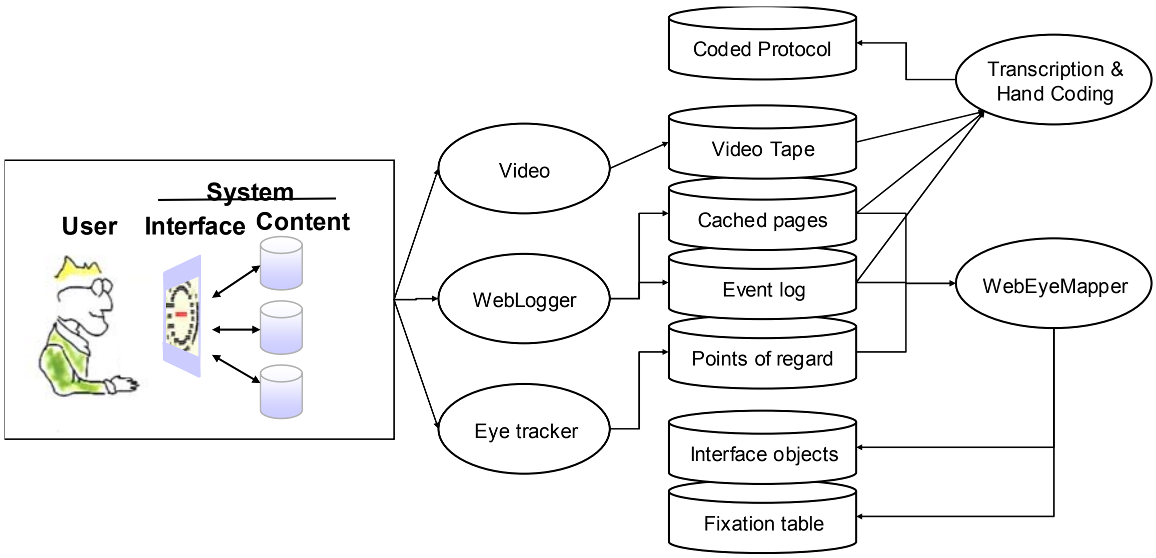


Figure 5. Instrumentation and analysis of data in the Web study.

Participants were given think aloud instructions and practice (Ericsson & Simon, 1984) Each question (in random order) was displayed on the computer screen to the left of the Internet Explorer and remained in view throughout the task. If participants did not complete the task within 10 minutes they were given a hint in form of a set of search terms that would help them find the target Web page. If participants did not complete the task within 15 minutes they were given the Web site where they could find the target information.

Results

Detailed protocol analyses were extremely time-consuming and consequently we decided to focus our efforts on protocols collected from four participants working on the Antz and City tasks listed in Table 1. These two tasks were representative of the complete set of tasks in the sense that they were near-median in task completion time and near-median in task completion time variance. The four participants were chosen because they had the most intact data on these two tasks.

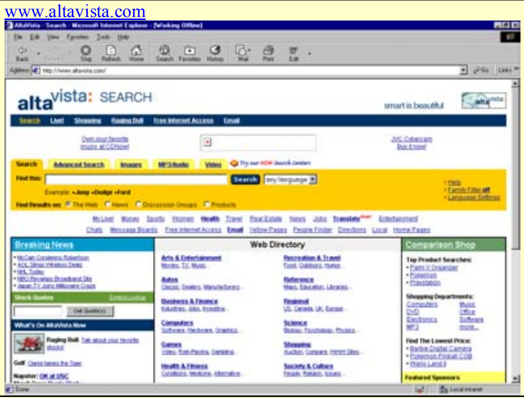
	SYSTEM	OBSERVED ACTIONS AND TRANSCRIPT	MODEL INTERPRETATION
		(Task #2:ANTZ)	
4:25:00		(Reads Question)	(O*READ-QUESTION)
		Ok, Um.	
		Let's go to Altavista because I like Altavista.	(G*GO-TO SITE "Altavista")
		Click: www.altavista.com in pulldown menu	(O*GO-TO SITE "www.altavista.com")
4:49.22	www.altavista.com 	Um, and let's look up Antz. And see what's generally available.	(G*SEARCH WEB www.altavista.com "Antz")
		Type search: Antz	(O*SEARCH WEB "www.altavista.com" "Antz")
4:55:00	http://www.altavista.com/cgi-bin/query?pg=q&sc=on&hl=on&q=antz&kl=XX&stype=stext	"The official Antz web ring," that sounds appropriate.	(O*EVAL LINK "The official Antz web ring" hi)
		Click: The Official ANTZ Webring link	(O*FOLLOW LINK "The official Antz web ring")
5:00.07	http://www.geocities.com/Area51/sShire/3303/ANTZ/webring.html	Uh, "list of sites,	(O*EVAL LINK "list of sites" null)
		random sites,	(O*EVAL LINK "random sites" null)
		how to join."	(O*EVAL LINK "how to join" null)
		Let's have a look at the list of sites.	(G*FOLLOW LINK "list of sites")
		Click: A List of Sites link	(O*FOLLOW LINK "list of sites")
5:06.19	http://www.webring.org/cgi-bin/webring?ring=z4195;list	"The anomaly, a few MX goodies mixed in,"	(O*EVAL LINK "The Anomaly" null "a few MX goodies mixed in")

Figure 6. A segment of a coded Web protocol transcript.

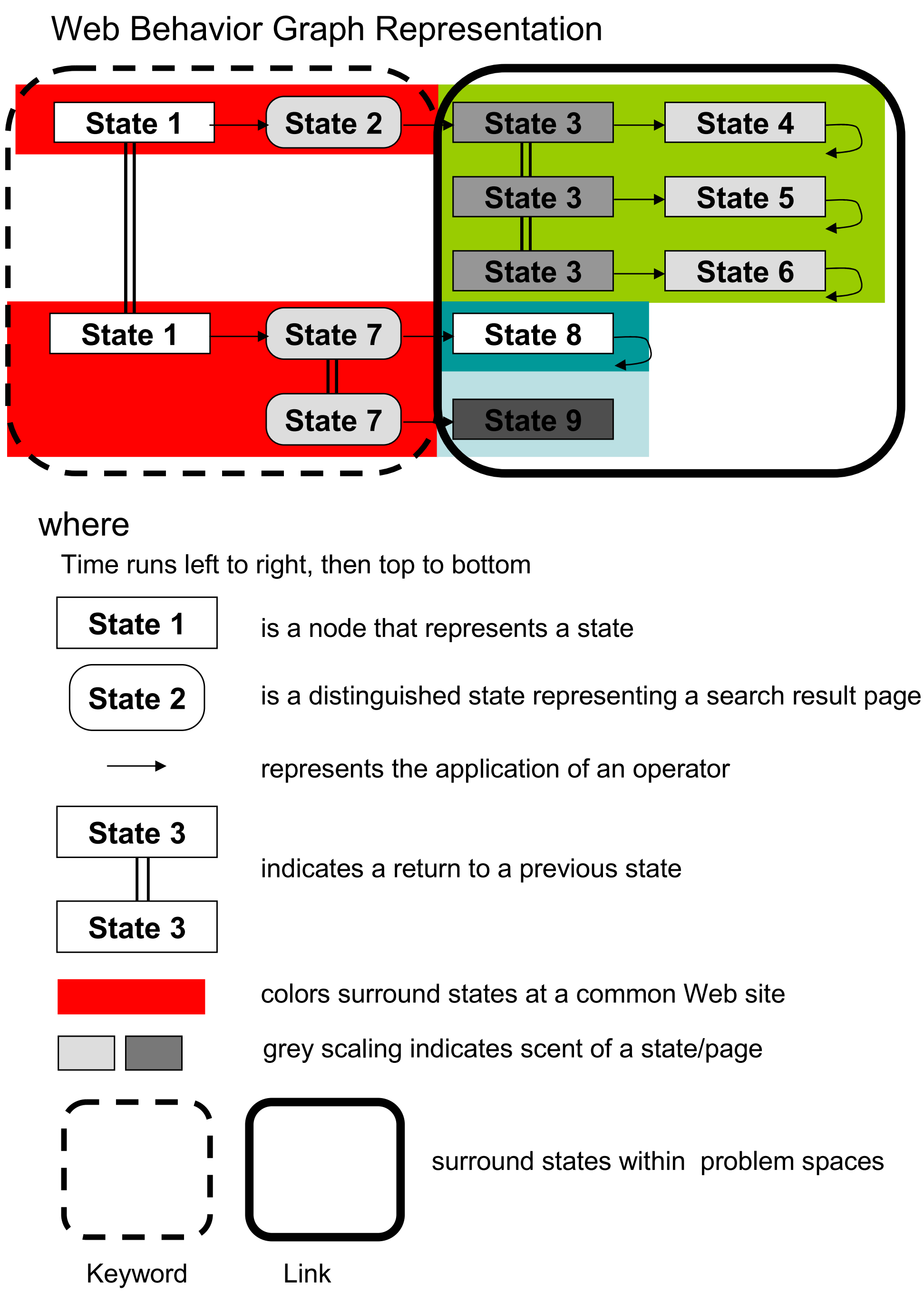


Figure 7. Schematic explanation of the components of a Web Behavior Graph.

Figure 5 shows how the data collected from participants were merged into a coded Web protocol transcript. Data were drawn from transcriptions of the video recordings of the think aloud protocol and the Weblogger recording of system and user actions with the browser (including the actual Web pages visited).

Several initial passes were made over a subset of the protocols to develop a Web Protocol Coding Guide[12]. A small portion of a coded Web protocol transcript is presented in Figure 6.[13] The leftmost column contains timing information, the second column contains information about the system state, including images of the page visited by the participant, the third column contains user actions (from the Weblogger recordings or the video) and verbalizations, and the fourth column contains codings of the protocol data contained in the previous columns, according to the Web Protocol Coding Guide. Sample protocols coded independently by two analysts yielded a 91% agreement in the partitioning of the protocol and a 93% agreement on the specific codings.

Problem spaces

Excluding the eye movement data, the participants' protocols suggest that three problem spaces structure the bulk of their Web foraging behavior.[14] There is

1. A *link* problem space in which the states are information patches (typically Web pages) and the basic operators involve moving from one patch to another by clicking on links or the back button,
2. A *keyword* problem space for in which the states are alternative legal queries to search engines, and the operators involve formulating and editing the search queries,
3. A URL problem space in which the states are legal URLs to be typed into the address bar of a Web browser and the operators involve formulating and editing the URL strings.

The participants' behavior in the Web Protocols can be visualized using a Web Behavior Graph (*WBG*, Card et al., 2001), which is an elaboration of the problem behavior graphs used in Newell and Simon (1972). Figure 7 presents a schematic explanation of the WBG graphical representation.

---

[12] The detailed Web Protocol Coding Guide is available on request.

[13] The Web Protocol Transcript presented in Figure 6 actually omits some details for the purposes of presentation here. For instance, I have only included one Web page image here, but the original transcript contains an image on every line of the transcript where a URL is listed in the System column. I have also omitted several timing indicators used to coordinate the transcript with the WebLogger files

[14] Other problem spaces are evident the protocols, for instance for navigating through the history list, but are much rarer than the ones discussed here. Other problem spaces could easily be added to the analysis.

Figure 8 presents the WBGs of the four participants' data on the Antz and City tasks.[15] Since the participants' protocols provide sparse data regarding their evaluations of the pages they visited, we had three independent judges rank the potential utility (information scent) of each page visited in the eight Web Protocols on a four point scale: No Scent (0), Low Scent (1), Medium Scent (2), or High Scent (3). The geometric mean of the judges ratings were taken to reduce the effect of outlying ratings. These ratings are plotted in the WBGs in Figure 8 on a scale of white (0), light gray (1), medium gray (2) and dark gray (3).

Following Information Scent

Inspection of Figure 8 reveals several phenomena. The Antz task is more difficult than the City task. Only 1/4 participants found a solution to the Antz task as opposed to 3/4 participants who found solutions to the City task. The WBGs for the Antz task show more branches and backtracking than the WBGs for the City task, which is an indication of less informed search. The participants on the City task move very directly to the target information, whereas the participants on the Antz task follow unproductive paths.

Antz task participants spend more time visiting search engines and in the Keyword problem space. On the Antz task participants generated about 3.25 separate sets of search results each, whereas on the City task they generated about 1.25 sets of search results each. One reason for the greater difficulty of the Antz task could be the poverty of information scent of links leading to the desired target information. We asked $N = 10$ judges to rate the information scent of links on the search results pages visited by participants in both the Antz and City tasks. The links were rated on a three-point scale of Not Relevant (0), Low Relevant (1), Medium Relevant (2), and Highly Relevant (3). A median split analysis showed that higher rated links were more likely to have been the ones selected by the study participant, $\chi^2_{(1)} = 15.46$, $p < .0001$. The links followed by the study participants had a lower information scent average on the Antz task ($M = 1.56$) than on the City task ($M = 2.44$) although the links not followed where about the same for the Antz task ($M = .65$) and the City task ($M = .62$).

Difficulty of foraging on the Web appears to be related to the quality of information scent cues available to users. Under conditions of strong information scent, users move rather directly to the target information, as is characteristic of

[15] WBGs can be automatically generated from WebLogger data.

the City task WBGs in Figure 8. When the information scent is weak, there is a more undirected pattern of foraging paths, as characteristic of the Antz WBGs in Figure 8, and a greater reliance on search engines.

Foraging in Information Patches

Although multiple Web sites were visited by all participants on both tasks, it is apparent that they tended not to flit from one Web site to another. There were more transitions within a Web site than between sites. The ratio of within-site to between-site transitions was $M = 2.1$ for the Antz task and $M = 5.2$ for the City task. Inspection of Figure 8 suggests that as the information scent of encountered Web pages declines at a site that there is a tendency for participants to leave the site or return to some previously visited state. From the Web Protocols I identified segments where participants visited three or more pages at a Web site that was not a search engine or Web portal. I found $N = 3$ three-page sequences and $N = 6$ five-page sequences (no other sequences were found). Figure 9 presents the scent ratings of the pages in these sequences as rated by the three expert judges discussed above. Each point in Figure 9 is the geometric mean of scent ratings of the visited pages produced by an independent panel of raters. Also plotted in Figure 9 is the geometric mean rating of the next page visited after leaving the site, and the geometric mean rating of all pages. When interpreting Figure 9, it is important to recall that the ratings form an ordinal scale and consequently the graph cannot be interpreted quantitatively. Figure 9 shows that initially the information scent at a site is high, and when that information scent becomes low, users switch to another site or search engine. A user assessing the potential rewards of foraging at a site based on information scent may at some point encounter diminishing returns and decide to go to another site. So long as the potential rewards are above some threshold, then the user continues foraging at the site. When the potential rewards pass below that threshold, then the user moves on to another site.

There also appears to be a relation between the amount of information scent first encountered at a Web site and the length of the sequence (the *run*) of page visits at the Web site. I identified $N = 68$ sequences of page visits (one or greater) at Web sites, and split these runs two ways: (a) run length ≥ 3 vs run length ≤ 2, and (b) start page information scent $\geq$ median vs start page information scent $<$ median. This median split analysis showed that starting with a high information scent was strongly associated with longer runs at a Web site, $\chi^2(1) = 8.69$, $p < .005$.

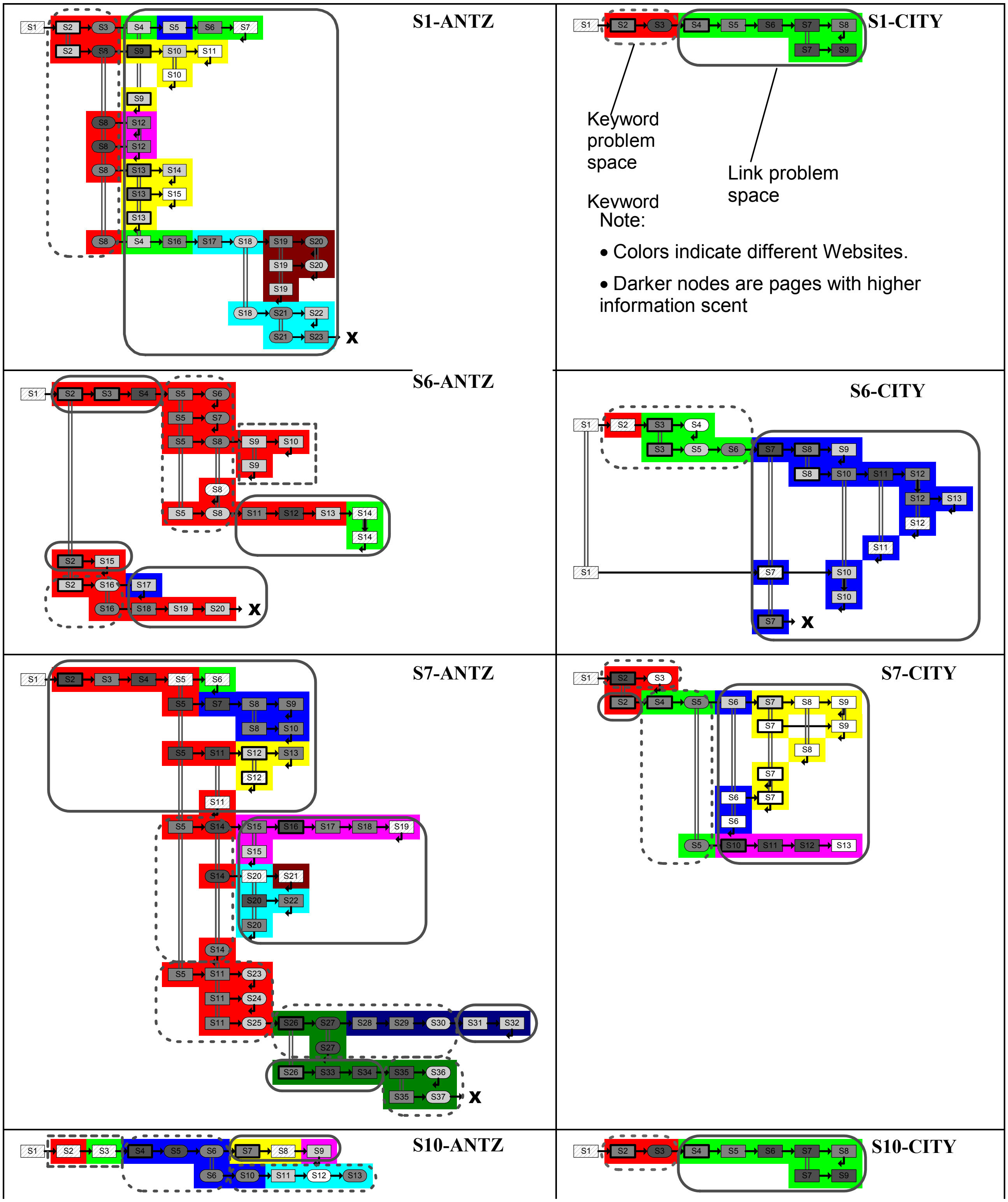


Figure 8. Web Behavior Graphs for four participants (rows) working on two tasks (columns).

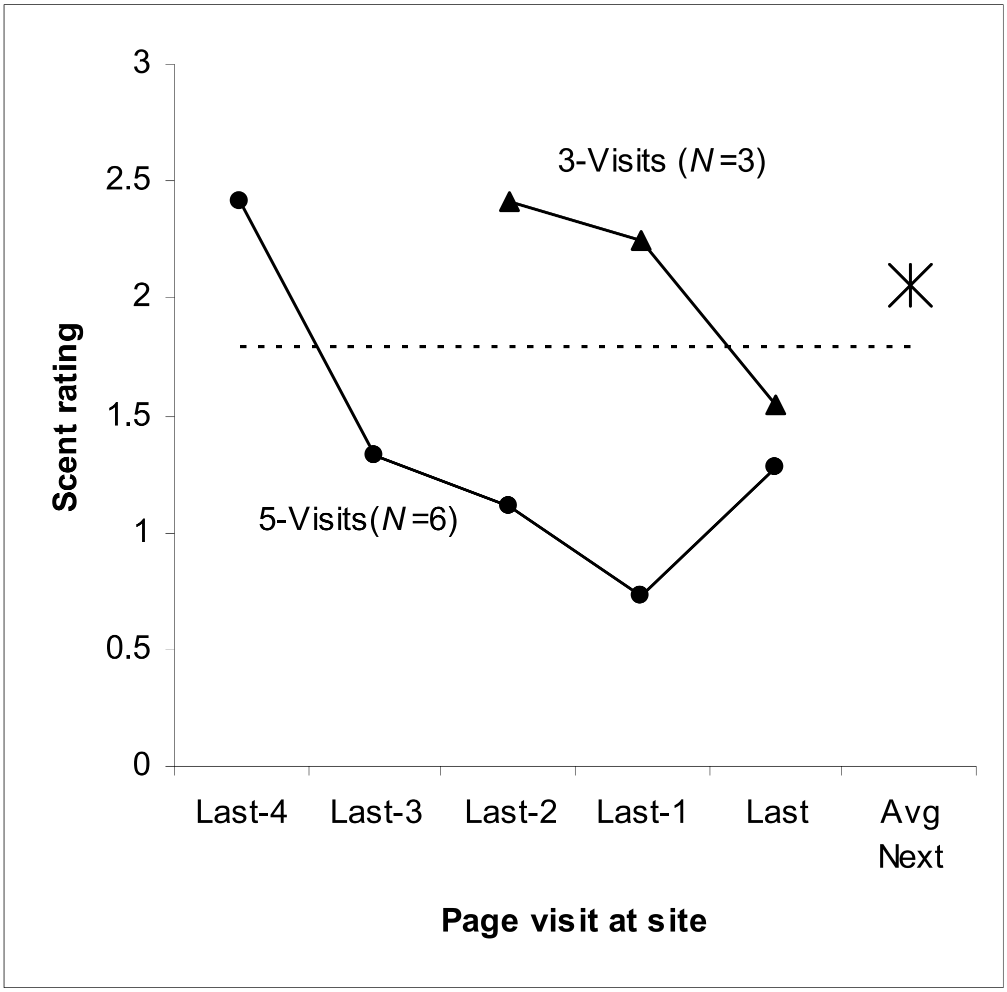


Figure 9. Information scent as a function of number of states prior to leaving the current Web site.

SNIF-ACT ARCHITECTURE

A model called SNIF-ACT (Pirolli & Fu, 2003) was developed to simulate the participants in the Web study presented above. SNIF-ACT extends the ACT-R theory and simulation environment (Anderson & Lebiere, 2000). ACT-R is a hybrid production system architecture designed to model human psychology. It is a hybrid because it combines production rule processing with neural-like connectionist processing. ACT-R contains three kinds of assumptions about: (1) knowledge representation, (2) knowledge deployment (performance), and (3) knowledge acquisition (learning). There are two major memory components in the ACT-R architecture: a *declarative knowledge* component and a *procedural* *knowledge* component. The distinction between these two kinds of knowledge is close to Ryle's (1949) distinction between *knowing that* and *knowing how*. One kind of knowledge (knowing that; declarative) is the kind that can be contemplated or reflected upon, whereas the other kinds of knowledge (know-how; procedural) is tacit and directly embodied in physical or cognitive activity.

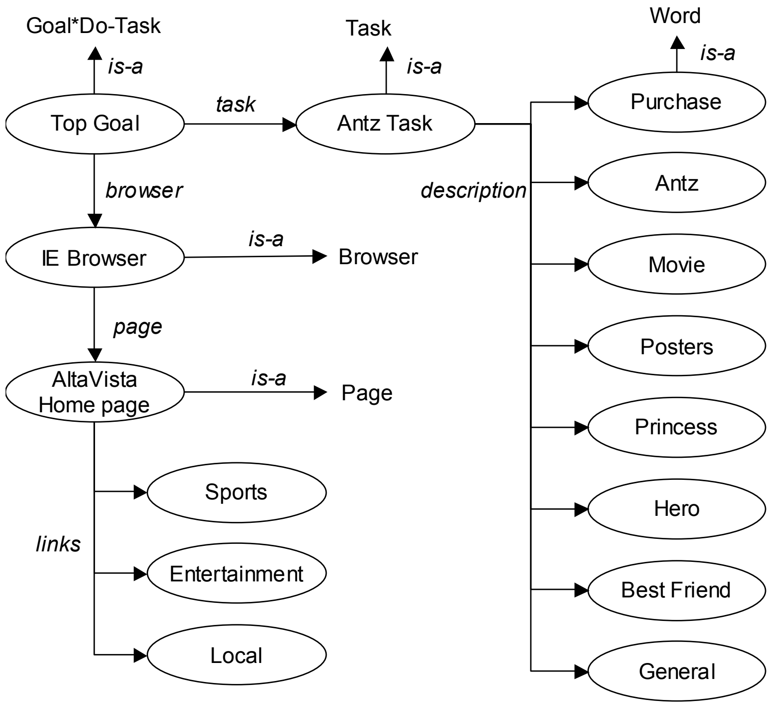


Figure 10. A subset of the declarative knowledge in the SNIF-ACT simulation of user S1 working on the Antz task. Chunks representing the tog goal, task description, and browser display are depicted.

Declarative Knowledge

Declarative knowledge corresponds to things that we are aware we know and that can be easily described to others, such as the content of Web links, or the functionality of browser buttons. Declarative knowledge is represented formally as chunks in ACT-R. Figure 10 provides a graphical representation of the declarative chunks in a SNIF-ACT simulation of a user performing the Antz task. The graphical notation in Figure 10 corresponds to the notation used for network graph representation of human memory in standard textbooks (Anderson, 2000). Each oval represents a chunk. Arrows indicate labeled relations among chunks, or indicate the type of chunks using the label "is-a." The chunk Top-goal represents he user's main goal, and it is related (*points*) to the chunk "Antz task" which points to a description of the concepts involved in the task. The Top-goal chunk also points to a chunk representing the user's perception of the Internet Exporer (IE) Browser, which points to the user's perception of the AltaVista home page that is displayed in the browser and the links that have been perceived on that page.

Declarative chunks in ACT-R have activation values. Activation spreads from the current focus of attention, including goals, through *associations* among chunks in declarative memory. Goals are also represented as chunks. At any point in time, ACT-R is focused on a single goal.

Procedural Knowledge

Procedural knowledge is knowledge (skill) that we display in our behavior without conscious awareness, such as knowledge of how to ride a bike, or how to point a mouse to a menu item. Procedural knowledge specifies how declarative knowledge is transformed into active behavior. Procedural knowledge is represented formally as condition-action pairs, or *production rules* (Figure 5). For instance, the SNIF-ACT simulations contain the production rule (summarized in English):

Use-Search-Engine:
IF the goal is Goal*Start-Next-Patch
  & there is a task description
  & there is a browser
  & the browser is not at a search engine
THEN
  Set a subgoal Goal*Use-search-engine

The production (titled Use-search-engine) applies in situations where the user has a goal to go to a Web site (represented by the tag Goal*Start-Next-Patch), has processed a task description, and has a browser in front of them. The production rule specifies that a subgoal will be set to use a search engine. The condition (IF) side of the production rule is matched to the current goal and the active chunks in declarative memory, and when a match is found, the action (THEN) side of the production rule will be executed. Roughly, the idea is that each elemental step of cognition corresponds to a production. At any point in time, a single production fires. When there is more than one match, the matching rules form a *conflict set*, and a mechanism called *conflict resolution* is used to decide which production to execute. The conflict resolution mechanism is based on a utility function. The expected utility of each matching production is calculated based on this utility function, and the one with the highest expected utility will be picked. In modeling Web users, the utility function is provided by information scent.

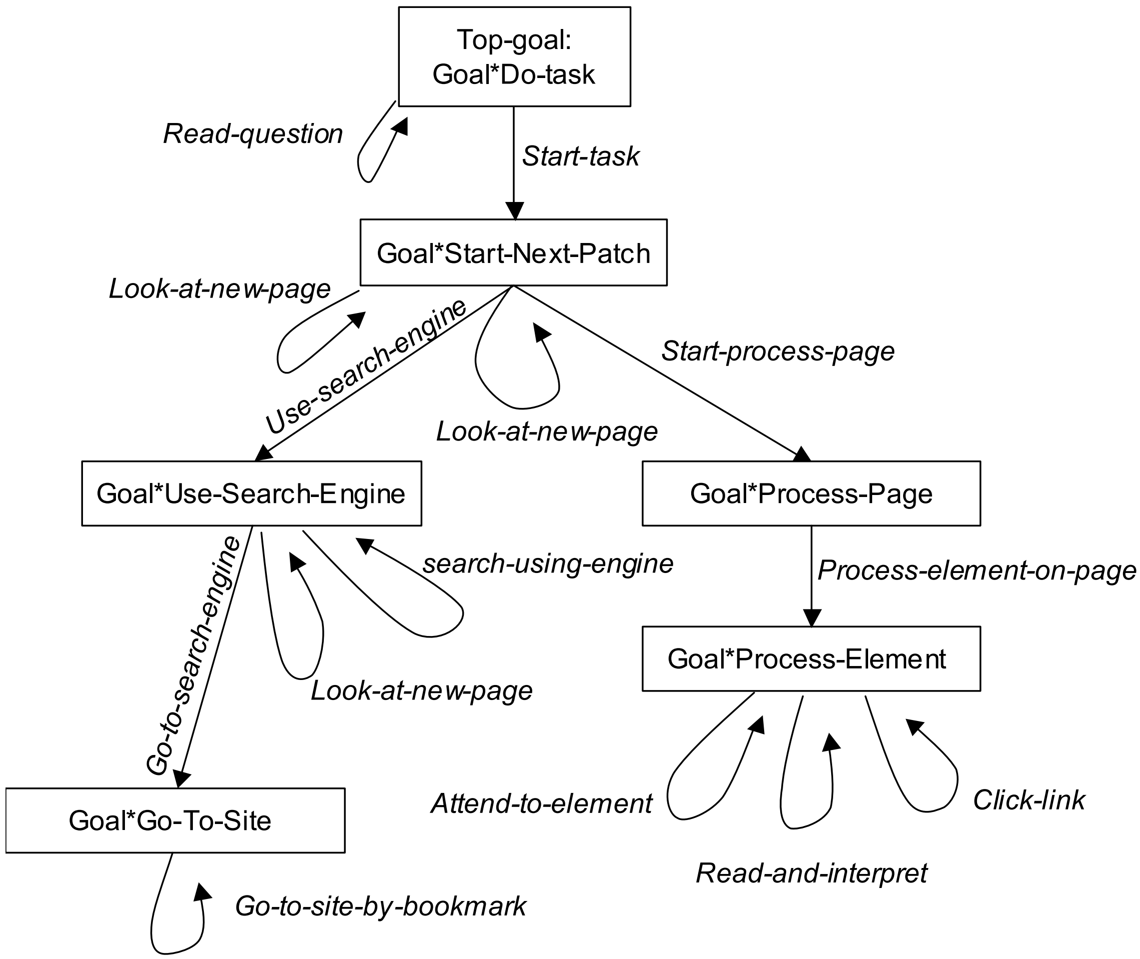


Figure 11. Goals and production rule execution in the initial portion of the SNIF-ACT simulation of user S1 working on the Antz task.

Figure 11 presents a small portion of a trace of the SNIF-ACT simulation of user S1 working on the Antz task. Each box represents a goal and each arrow represents the *firing* (match and execution) of a production rule in the SNIF-ACT simulation. The sequence of processing is depth-first, left-to-right in the diagram.

Utility and Choice: The Role of Information Scent

In a SNIF-ACT simulation, information scent cues on a computer display activate chunks and spread through the declarative network of chunks. The amount of activation accumulating on the chunks matched by a production is used to evaluate and select productions. The activation of chunks matched by production rules is used to determine the utility of selecting those production rules. This is the most significant difference between SNIF-ACT and ACT-R, which does not have an activation-based model of utility.

For instance, the following Click-link production rule matches when a Web link description has been read,

Click-link:
IF the goal is Goal*Process-element
& there is a task description
& there is a browser
& there is a link that has been read
& the link has a link description
THEN
Click on the link

If selected, the rule will execute the action of clicking on the link. The chunks associated with the task description and the link description will have a certain amount of activation. That combined activation will be used to evaluate the rule. If there are two Click-link productions matching against chunks for two different links, then the production with more highly activated chunks will be selected.

The predictions made by the SNIF-ACT model were tested against the log files of the data sets summarized in Figure 8. The major controlling variable in the model is the measure of information scent, which predicts two major kinds of actions: (1) which links on a web page people will click on, and (2) when people decide to leave a site. These kinds of actions were therefore extracted from the log files and compared to the predictions made by the model. We call the first kind of actions *link-following* actions, which were logged whenever a participant clicked on a link on a web page. The second kind of actions was called *site-leaving* actions, which were logged whenever a participant left a web site (and went to a different search engine or web site). The two kinds of actions made up 72% (48% for link-following and 24% for site-leaving actions) of all the 189 actions extracted from the log files.

Link-following actions

The SNIF-ACT model was matched to the link-following actions extracted from the data sets. Each action from each participant was compared to the action chosen by the simulation model. Whenever a link-following action occurred in the user data we examined how the SNIF-ACT model ranked (using information scent) all the links on the web page where the action was observed. We then compared the links chosen by the participants to the predicted link rankings of the SNIF-ACT model. If there were a purely deterministic relationship between predicted information scent and link choice, then all users would be predicted to choose the highest ranked link. However, we assume that the scent-based utilities are stochastic (McFadden, 1974, 1978) and subject to some amount of variability due to users and context (which is also consistent with ACT-R (Anderson & Lebiere, 2000)). Consequently we expect the probability of link choice to be highest for the links ranked with the greatest amount of scent-based utility, and that link choice probability is expected to decrease for links ranked lower on the basis of their scent-based utility values.

Figure 12 shows that link choice is strongly related to scent-based utility values. Links ranked higher on scent-based utilities tend to get chosen over links ranked lower. There are a total of 91 link-following actions in Figure 12. The distribution of the predicted link selection was significantly different from random selection $\chi 2$ (30) = 18,589, $p < 0.0001$. This result replicates a similar analysis made by Pirolli and Card (Pirolli & Card, 1999) concerning the ACT-IF model prediction of cluster selection in the Scatter/Gather browser.

Site-leaving actions

To test how well information scent is able to predict when people will leave a site, site-leaving actions were extracted from the log files and analyzed. Site-leaving actions are defined as actions that led to a different site (e.g. when the participants used a different search engine, typed in a different URL to go to a different web site, etc.) These data are presented in Figure 13. Each data point is the average of N = 12 site-leaving actions observed in the data set. The x-axis indexes the four steps made prior to leaving a site (Last-3, Last-2, Last-1, Leave-Site). The y-axis in Figure 13 corresponds to the average information scent value computed by the SNIF-ACT spreading activation mechanisms. The horizontal dotted line indicates the average information scent value of the page visited by users after they left a Web site.

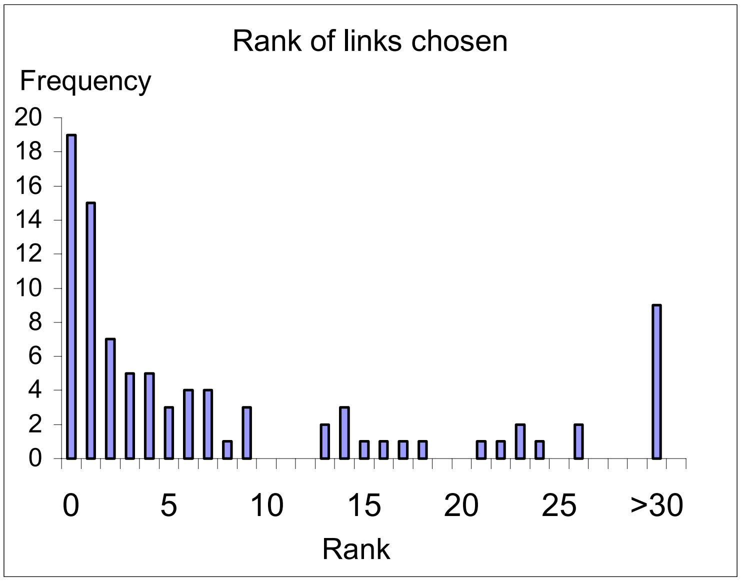


Figure 12. Frequency that SNIF-ACT productions match link-following actions. The SNIF-ACT production rankings are computed at each simulation cycle over all links on the same web page and all productions that match.

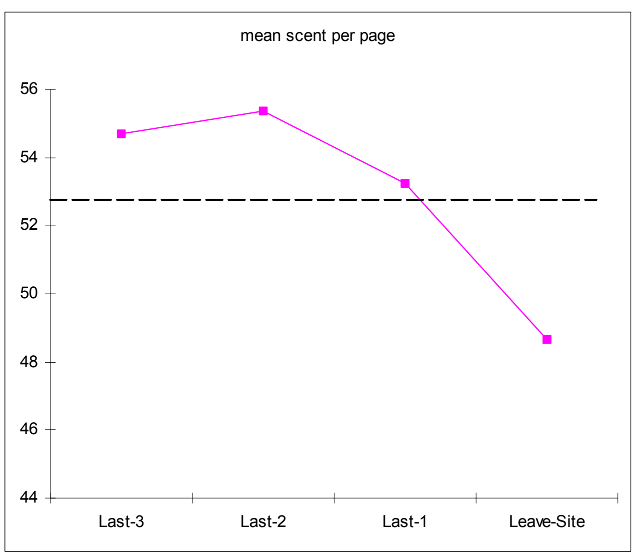


Figure 13. Mean information scent on the web page as a function of distance from the last web page before the users left the site.

Figure 13 suggests that users essentially assess the expected utility of continuing on at an information patch (i.e., a Web site) against the expected utility of switching their foraging to a new information patch. This is consistent with the predictions of information foraging theory and with previous findings on information patch foraging (Pirolli & Card, 1999). Figure 13 also suggests that spreading activation mechanisms compute activation values from information scent cues in order to reflect expected utilities associated with navigation choices. It is worth emphasizing that the formulation of the spreading activation networks employed in the information scent model were motivated by a Lens Model analysis of the environmental problem faced by the forager. Furthermore, the actual parameters of the spreading activation networks (the strengths) were set by an automated analysis of the statistical properties of the environment. The spreading activation networks used in the cognitive model reflect the probabilistic texture of the environment.

GENERAL DISCUSSION

The notion of information scent has played a role in understanding other areas of human-information interaction, and has played a role in practical application. Information scent was originally developed as part of information foraging models of a document clustering and browsing systems called Scatter/Gather (Pirolli & Card, 1999). A computational cognitive model called ACT-IF, very similar to the one presented here, modeled the navigation choices and patch foraging behavior in Scatter/Gather using information scent. Information scent was also used as an explanatory concept in understanding how users visually scan and interact with novel information visualization systems such as the Hyperbolic browser (Pirolli et al., 2003) and relevance enhanced thumbnails (Woodruff et al., 2002). The most significant current problem for the future development of the models concerns the analysis of non-text information scent cues, such as graphical icons, animations, and so on, and the relation of proximal information scent cues to non-text distal content such as video and music.

Another line of future theoretical development concerns learning. To date, cognitive models of information foraging such as ACT-IF and SNIF-ACT are built on spreading activation networks that represent associations among words. As discussed above, network strengths are estimated directly from statistics obtained from large corpora of language. The match of proximal cues to the desired information goal in Equations 10 through 12 can be viewed as implementing a kind of exemplar-based category match model. In fact, I derived the original equations presented in Pirolli and Card (199) from consideration of the examplar-based categorization models of Medin and Schaffer (1978) and Kruschke (1992). It seems unquestionable that people develop richer category representations of the world of information around them, and all large repositories (e.g., libraries or the Web) attempt to index into that category structure. Pirolli (2003) summarizes a nascent model called InfoCLASS that builds upon Anderson’s (1991) rational model of categorization. The goal of that model is to capture the observation that different kinds of information browsers, and different histories of information foraging, lead to different individual conceptions of what is out there.

Information scent has become a design concept in Web site design and usability (Nielsen, 2003; Spool, Scanlon, Snyder, & Schroeder, 1998; User Interface Engineering, 1999). Computational models of information scent have been used in the development of Bloodhound (Chi, Pirolli, Chen, & Pitkow, 2001; Chi, Pirolli, & Pitkow, 2000; Chi et al., 2003) which is a system that performs automated Web site usability analysis. A similar Web site usability system based on Latent Semantic Analysis (Landauer & Dumais, 1997) rather than spreading activation networks has also been developed (Blackmon, Polson, Kitajima, & Lewis, 2002). These automated usability systems basically take a hypothetical user goal (input as a set of words), compute the probability of navigation choice of links emanating from each page given that goal, flow simulated users through the Web site, and perform usability analyses such as estimates of the number of steps it would take on average to reach desired information from a given start page.

Particularly intriguing is the inverse problem of inferring user goals from observed navigation behavior. The RUM model in Equation 10 predicts choices given a goal and a set of alternative choices. Economists also attempt to infer goals given the observed choices made by people from set of alternatives (called *revealed preferences*). The Lumberjack system (Chi, Rosien, & Heer, 2002) is an attempt to use the theory of information scent to perform such goal inference. It may be possible to develop information environments that automatically tailor themselves to optimize the foraging behavior of users.

ACKNOWLEDGMENTS

This research has been funded in part by an Office of Naval Research Contract No. N00014-96-C-0097 to Peter Pirolli and Stuart K. Card, and a Advanced Research and Development Activity contract No. MDA904-03-C-0404.

REFERENCES

Anderson, J. R. (1976). *Language, memory, and thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*, 409-429.

Anderson, J. R. (2000). *Cognitive psychology and its implications: Fifth edition*. New York: Worth.

Anderson, J. R., & Lebiere, C. (2000). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*, 703-719.

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

Baldi, P., Frasconi, P., & Smyth, P. (2003). *Modeling the Internet and the Web*. Chichester, UK: Wiley and Sons.

Barabàsi, A.-L. (2002). *Linked: The new science of networks*. Cambridge, MA: Perseus Publishing.

Bechtel, W. (1985). Realism, instrumentalism, and the intentional stance. *Cognitive Science, 9*, 473-497.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). *Cognitive Walkthrough for the Web.* Paper presented at the Human Factors in Computing Systems, CHI 2002, Minneapolis, MN.

Brunswik, E. (1952). *The conceptual framework of psychology*. Chicago: University of Chicago Press.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press.

Card, S., Pirolli, P., Van Der Wege, M., Morrison, J., Reeder, R., Schraedley, P., & Boshart, J. (2001). *Information scent as a driver of Web Behavior Graphs: Results of a protocol analysis method for web usability.* Paper presented at the Human Factors in Computing Systems, Seattle, WA.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Chi, E., Pirolli, P., Chen, K., & Pitkow, J. E. (2001). *Using information scent to model user needs and actions on the web.* Paper presented at the Human Factors in Computing Systems, CHI 2001, Seattle, WA.

Chi, E., 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.* Paper presented at the Human Factors in Computing Systems, CHI 2000, The Hague.

Chi, E. H., Rosien, A., & Heer, J. (2002). *Lumberjack: Intelligent discovery and analysis of Web user traffic composition.* Paper presented at the ACM-SIGKIDD Workshop on Web mining for usage patterns and user profiles, WebKDD 2002, Edmonton, Canada.

Chi, E. H., Rosien, A., Suppattanasiri, G., Williams, A., Royer, C., Chow, C., Robles, E., Dalal, B., Chen, J., & Cousins, S. (2003). *The Bloodhound Project: Automating discovery of Web usability issues using the InfoScent simulator.* Paper presented at the Conference on Human Factors in Computing Systems, CHI 2003., Fort Lauderdale, FL.

Choo, C. W., Detlor, B., & Turnbull, D. (1998). *A behavioral model of information seeking on the Web: Preliminary results of a study of how managers and IT specialists use the Web.* Paper presented at the 61st Annual Meeting of the American Society for Information Science, Pittsburgh, PA.

Cosmides, L., Tooby, J., & Barow, J. H. (1992). Introduction: evolutionary psychology and conceptual integration. In J. H. Barkow & L. Cosmides & J. Tooby (Eds.), *The adapted mind: Evolutionary psychology and the generationof culture* (pp. 3–15). New York: Oxford University Press.

Davison, B. (2000). *Topical locality in the Web.* Paper presented at the Proceedings of the 23rd Annual International Conference on Information Retrieval, Athens.

Dennett, D. C. (1983). Intentional systems in cognitive ethology: The "Panglossian Paradigm" revisited. *6*, 343-390.

Dennett, D. C. (1988). *The intentional stance.* Cambridge, MA: Bradford Books, MIT Press.

Dennett, D. C. (1995). *Darwin's dangerous idea*. New York: Simon and Schuster.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol Analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Fitts, P. M., & Jones, R. E. (1961). Psychological aspects of instrument display: Analysis of factors contributing to 460 "Pilot Error" experiences in operating aircraft controls (1947), Reprinted in *Selected papers on human factors in the design and use of control systems* (pp. 332-358). New York: Dover Publications. Inc.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Furnas, G. W. (1997). *Effective view navigation.* Paper presented at the Human Factors in Computing Systems, CHI '97, Atlanta, GA.

Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston: Houghton Mifflin.

Gigerenzer, G. (2000). *Adaptive thinking: Rationality in the real world*. Oxford: Oxford University Press.

Harman, D. (1993). *Overview of the first text retrieval conference.* Paper presented at the 16th Annuam International ACM/SIGIR Conference, Pittsburgh, PA.

Hartson, H. R., & Castillo, J. C. (1998). *Remote evaluation for post-deployment usability improvement.* Paper presented at the Working Conference on Advanced Visual Interfaces, AVI '98, L'Aquila, Italy.

Kehoe, C., Pitkow, J., Sutton, K., & Aggarwal, G. (1999, May). *Results of the Tenth World Wide Web User Survey* [URL]. Retrieved December, 2003, from the World Wide Web: http://www.gvu.gatech.edu/user_surveys/survey-1998-10/tenthreport.html

Klein, G. A., Calderwood, R., & Macgregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics, 19*(3).

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*, 22-44.

Lamping, J., & Rao, R. (1994). *Laying out and visualizing large trees using a hyperbolic space.* Paper presented at the UIST '94, Marina del Rey.

Lamping, J., Rao, R., & Pirolli, P. (1995). A focus + context technique based on hyperbolic geometry for visualizing large hierarchies, *CHI '95, ACM Conference on Human Factors in Computing Systems*. New York: ACM.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Manning, C. D., & Schuetze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Marr, D. (1982). *Vision*. San Francisco: W.H. Freedman.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers of econometrics*. New York: Academic Press.

McFadden, D. (1978). Modelling the choice of residential location. In A. Karlqvist & L. Lundqvist & F. Snickars & J. Weibull (Eds.), *Spatial interaction theory and planning models*. Cambridge, MA: Harvard University Press.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review, 85*, 207-238.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Morrison, J. B., Pirolli, P., & Card, S. K. (2001). *A taxonomic analysis of what World Wide Web activities significantly impact people's decisions and actions.* Paper presented at the Conference on Human Factors in Computing Systems, CHI '01, Seattle, WA.

Neisser, U. (1976). *Cognition and reality*. San Francisco, CA: Freeman.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice Hall.

Nielsen, J. (2001, April, 2003). *The 3Cs of critical Web use: Collect, compare, choose* [URL]. Nielsen Norman Group. Retrieved December, 2003, from the World Wide Web: http://www.useit.com/alertbox/20010415.html

Nielsen, J. (2003, June, 2003). *Information foraging: Why Google makes people leave your site faster* [URL]. useit.com. Retrieved February, 2004, from the World Wide Web: http://www.useit.com/alertbox/20030630.html

Orians, G. H., & Heerwagen, J. H. (1992). Evolved responses to landscapes. In J. H. Barkow & L. Cosmides & J. Tooby (Eds.), *The adapted mind* (pp. 555-579). New York: Oxford University Press.

Pirolli, P. (1997). *Computational models of information scent-following in a very large browsable text collection.* Paper presented at the Conference on Human Factors in Computing Systems, CHI '97, Atlanta, GA.

Pirolli, P. (2003). A theory of information scent. In J. Jacko & C. Stephanidis (Eds.), *Human-computer interaction* (Vol. 1, pp. 213-217). Mahwah, NJ: Lawrence Erlbaum.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review, 106*, 643-675.

Pirolli, P., Card, S. K., & Van Der Wege, M. M. (2003). The effects of information scent on visual search in the Hyperbolic Tree Browser. *ACM Transactions on Computer-Human Interaction, 10*(1), 20-53.

Pirolli, P., & Fu, W. (2003). SNIF-ACT: A model of information foraging on the World Wide Web. In P. Brusilovsky & A. Corbett & F. de Rosis (Eds.), *User Modeling 2003, 9th International Conference, UM 2003* (Vol. 2702, pp. 45-54). Johnstown, PA: Springer-Verlag.

Pitkow, J. E., & Pirolli, P. (1999). *Mining longest repeated subsequences to predict World Wide Web suring.* Paper presented at the Second USENIX Symposium on Internet Technologies and Systems.

Quillan, M. R. (1966). *Semantic memory*. Cambridge, MA: Bolt, Bernak, and Newman.

Reeder, R. W., Pirolli, P., & Card, S. K. (2001). *Web-Eye Mapper and WebLogger: Tools for analyzing eye tracking data collected in web-use studies.* Paper presented at the Human Factors in Computing Systems, CHI 01, Seattle, WA.

Resnikoff, H. L. (1989). *The illusion of reality*. New York: Springer-Verlag.

Ryle, G. (1949). *The concept of mind*. London: Hutchinson.

Shattuck, L. W., & Woods, D. D. (1994). *The Ciritical Incident Technique: 40 years later.* Paper presented at the 38th Annual Meeting of the Human Factors Society.

Simon, H. A. (1974). How big is a chunk? *Science, 183*, 482-488.

Spool, J. M., Scanlon, T., Snyder, C., & Schroeder, W. (1998). *Measuring Website usability.* Paper presented at the Conference on Human Factors and Computing Systems, CHI '98, Los Angeles, CA.

Thurstone, L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Todd, P. M., & Gigerenzer, G. (1999). Simple heuristics that make us smart. *Behavioral and Brain Sciences, 22*, XXX-XXX.

Turney, P. D. (2001). *Mining the Web for synonyms: PMI-IR versus LSA on TOEFL.* Paper presented at the Twelfth European Conference on Machine Learning, ECML 2001, Freiburg, Germany.

User Interface Engineering. (1999). *Designing information-rich web sites*. Cambridge, MA: Author.

Walker, J., & Ben-Akiva, M. (2002). Generalized Random Utility Model. *Mathematical Social Sciences, 43*, 303-343.

Wilson, E. O. (1998). *Consilience*. New York: Knopf.

Woodruff, A., Rosenholtz, R., Morrison, J. B., Faulring, A., & Pirolli, P. (2002). A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks. *Journal of the American Society for Information Science and Technology, 53*, 172-185.