Cognitive Science 29 (2005) 343–373

Rational Analyses of Information Foraging on the Web

Peter Pirolli

PARC, Information Sciences and Technologies Laboratory

Received 28 April 2004; received in revised form 13 December 2004; accepted 5 January 2005

Abstract

This article describes rational analyses and cognitive models of Web users developed within information foraging theory. This is done by following the rational analysis methodology of (a) characterizing the problems posed by the environment, (b) developing rational analyses of behavioral solutions to those problems, and (c) developing cognitive models that approach the realization of those solutions. Navigation choice is modeled as a random utility model that uses spreading activation mechanisms that link proximal cues (*information scent*) that occur in Web browsers to internal user goals. Web-site leaving is modeled as an ongoing assessment by the Web user of the expected benefits of continuing at a Web site as opposed to going elsewhere. These cost–benefit assessments are also based on spreading activation models of information scent. Evaluations include a computational model of Web user behavior called Scent-Based Navigation and Information Foraging in the ACT Architecture, and the Law of Surfing, which characterizes the empirical distribution of the length of paths of visitors at a Web site.

Keywords: Information foraging, Information scent, World Wide Web, Rational analysis, ACT–R, SNIF–ACT

1. Introduction

A prevalent strategy for engaging and adapting to modern physical and social environments is to acquire and use relevant external information. People (and technology designers) are faced with developing adaptive ways of acquiring and using external content to gain knowledge that improves decision making and problem solving. *Information foraging theory* (Pirolli & Card, 1999) addresses human–information interaction (HII) involving modern technologies such as the World Wide Web. The theory is concerned with human behavior and technology involved in gathering information for some purpose, such as making a medical decision, finding a restaurant, or solving a programming problem. This article presents an application of information foraging theory to the behavior observed in finding information on the Web. The focus

Requests for reprints should be sent to Peter Pirolli, PARC, Information Sciences and Technologies Laboratory, 3333 Coyote Hill Road, Palo Alto, CA 94304. E-mail: pirolli@parc.com is on understanding the behavior of users who are not expert in finding domain-specific information (cf. Bhavnani, 2002).

The human propensity to gather and use information to adapt to everyday problems in the world is a core piece of human psychology that has been largely ignored in cognitive studies. G. A. Miller (1983) argued that the human species might fruitfully be viewed as a kind of *informavore:* a species that hungers for information in order to gather it and store it as a means for adapting to the world. Humans, of course, are extreme in their reliance on information, with language and culture, and now modern technology, providing media for transmission within and across generations. Humans are the *Informavores Rex* of this era.

To develop theory and models of human information gathering behavior, information foraging theory has adopted the *rational analysis* program initiated by Anderson (1989, 1990, 1991). The rational analysis approach involves a kind of reverse engineering in which the theorist asks (a) *what* environmental problem is solved, (b) *why* is a given behavioral strategy a good solution to the problem, and (c) *how* is that solution realized by cognitive mechanism? The products of this approach include (a) characterizations of the relevant goals and environment, (b) mathematical rational choice models (e.g., optimization models) of idealized behavioral strategies for achieving those goals in that environment, and (c) computational cognitive models. This methodology is founded on the heuristic assumption that evolving, behaving systems are well designed (rational) for fulfilling certain functions in certain environments. Rational analysis is a variant form of an approach called *methodological adaptationism* that has also shaped research programs in behavioral ecology (e.g., Mayr, 1983; Stephens & Krebs, 1986; Tinbergen, 1963), anthropology (e.g., Winterhalder & Smith, 1992), and neuroscience (e.g., Glimcher, 2003).

In keeping with the rational analysis approach, this article begins with some general empirical characterizations about the structure of the Web as an environment for information gathering behavior. Then, rational analyses are developed to characterize optimal behavioral strategies for information foraging on the Web. Surprisingly, a branch of behavioral ecology called optimal foraging theory (Stephens & Krebs, 1986), which predicts the strategies used by animals and hunter-gatherers to forage for food, has been a valuable resource (among others) for developing rational analyses of how people forage for information. Finally, this article presents empirical evaluations of the optimization models and a computational cognitive model derived from the optimization analyses called Scent-Based Navigation and Information Foraging in the ACT Architecture (SNIF–ACT).

2. Aspects of the Web environment

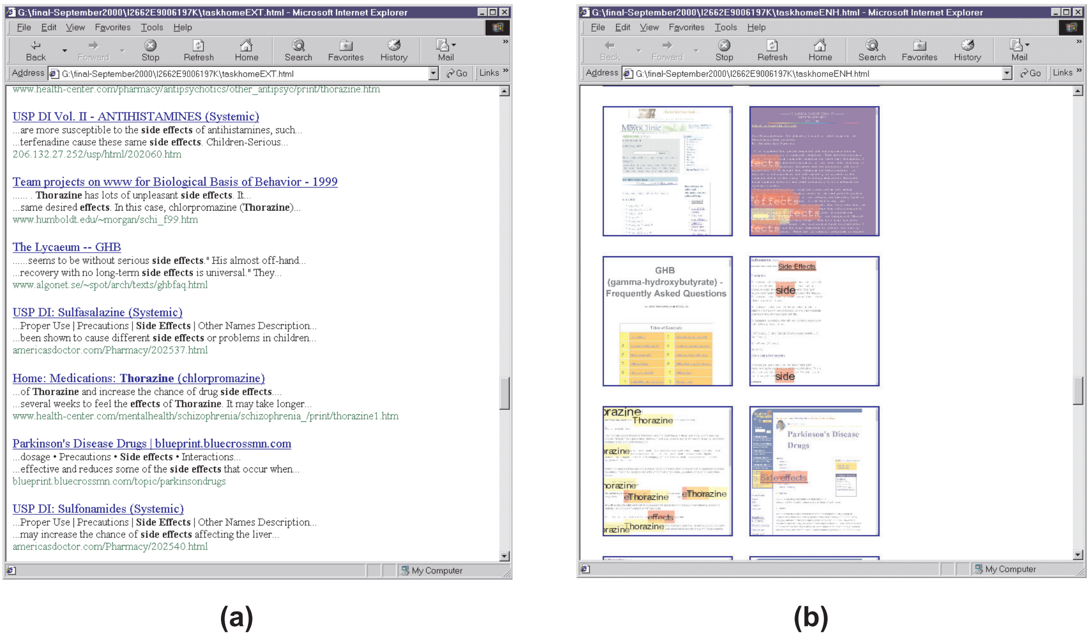
In this section, I review empirical observations about the Web to establish some general characterizations about task environments that involve the Web. These characterizations shape the rational analyses performed in the next section. As discussed in this section, people use the Web to acquire knowledge to improve ill-structured decision-making and problem solving. Interacting with the Web also involves costs—especially the opportunity cost of the time involved. Information foraging behavior is rational to the extent that it maximizes the value of the knowledge gained from the Web relative to the cost of interaction. Within this general framing of the value and costs of information foraging on the Web, there are two characteristic problems posed by the Web environment (and there are surely others): (a) the problem of choosing which links to follow based on available cues and (b) when to give up on a current Web locality (e.g., a Web site) and go to another.

2.1. *Ill-defined problems and the value of Web content*

When people are asked to report tasks they deem "significant" (Morrison, Pirolli, & Card, 2001), the majority of responses identify *ill-structured* problems (Reitman, 1964; Simon, 1973) requiring the acquisition of additional knowledge from external content sources. Ill-structured problems, such as choosing a medical treatment or buying a house typically require additional knowledge from external sources to better understand the starting state, to better define a goal, or to specify the actions that are afforded at any given state (Simon, 1973). People typically need to perform *knowledge search* (Newell, 1990) to solve their ill-structured problems (e.g., to define aspects of a problem space that permit effective or efficient problem space search). The Web is a potential source of valuable knowledge that can improve our range of adaptation because we can solve more problems, or solve problems using better approaches.

2.2. *Information scent cues for Web navigation*

The structure of the Web has evolved to exhibit regularities in the distribution of information resources and the navigation paths that lead to those resources. One regularity is the availability of labeled navigation links from one Web page to another (e.g., Fig. 1), and users appear to


G:\final-September2000\I2662E9006197K\taskhomeEXT.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail
Address G:\final-September2000\I2662E9006197K\taskhomeEXT.html Go Links
My Computer
(a)
G:\final-September2000\I2662E9006197K\taskhomeENH.html - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Forward Stop Refresh Home Search Favorites History Mail
Address G:\final-September2000\I2662E9006197K\taskhomeENH.html Go Links
MayoClinic
GHB
(gamma-hydroxybutyrate) -
Frequently Asked Questions
Side Effects
side
side
Thorazine
eThorazine
eThorazine
effects
Parkinson's Disease
Drugs
Side effects
My Computer
(b)

www.health-center.com/pharmacy/antipsychotics/other_antipsyc/print/thorazine.htm

USP DI Vol. II - ANTIHISTAMINES (Systemic)
...are more susceptible to the **side effects** of antihistamines, such...
...terfenadine cause these same **side effects**. Children-Serious...
206.132.27.252/usp/html/202060.htm

Team projects on www for Biological Basis of Behavior - 1999
.... **Thorazine** has lots of unpleasant **side effects**. It...
...same desired **effects**. In this case, chlorpromazine (**Thorazine**)...
www.humboldt.edu/~morgan/schi_f99.htm

The Lycaeum -- GHB
......seems to be without serious **side effects**." His almost off-hand...
...recovery with no long-term **side effects** is universal." They...
www.algonet.se/~spot/arch/texts/ghbfaq.html

USP DI: Sulfasalazine (Systemic)
...Proper Use | Precautions | **Side Effects** | Other Names Description...
...been shown to cause different **side effects** or problems in children...
americasdoctor.com/Pharmacy/202537.html

Home: Medications: **Thorazine** (chlorpromazine)
...of **Thorazine** and increase the chance of drug **side effects**....
...several weeks to feel the **effects** of **Thorazine**. It may take longer...
www.health-center.com/mentalhealth/schizophrenia/schizophrenia_/print/thorazine1.htm

Parkinson's Disease Drugs | blueprint.bluecrossmn.com
...dosage • Precautions • **Side effects** • Interactions...
...effective and reduces some of the **side effects** that occur when...
blueprint.bluecrossmn.com/topic/parkinsondrugs

USP DI: Sulfonamides (Systemic)
...Proper Use | Precautions | **Side Effects** | Other Names Description...
...may increase the chance of **side effects** affecting the liver...
americasdoctor.com/Pharmacy/202540.html

Fig. 1. Examples of information scent cues: (a) a typical set of search results in the form of textual summaries of linked documents and (b) Relevance Enhanced Thumbnails for the same set of search results.

prefer following links over other means of Web navigation (Katz & Byrne, 2003). Web page designs have evolved to associate (by human design or automated information systems) small snippets of text and graphics with such links. Those text and graphics cues are intended to represent tersely the content that will be encountered by choosing a particular link on one page and navigating to the linked page. When browsing the Web by following links, users must use these cues presented proximally on the Web pages they are currently viewing to make navigation decisions. These link cues are called *information scent.* For the Web user, there is uncertainty about the relation of the proximal cues to the linked information resources.

Fig. 1 presents some examples of information scent cues. Fig. 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, and a URL (Uniform Resource Locator). Fig. 1b illustrates an alternative form of search result representation (for exactly the same items in Fig. 1a) 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. In the complex network organization of the Web, small perturbations in the accuracy of information scent can cause qualitative shifts in the cost of browsing. This will be illustrated by application of an analysis of phase transitions in heuristic search developed by Hogg and Huberman (1987).

For a specific arrangement of Web content, interpage links, and information scent on a Web site, it is becoming possible to make model-based predictions about likely time costs of navigation—for each specific case. Examples of such models include SNIF–ACT (discussed later), Bloodhound (Chi et al., 2003), Cognitive Walkthrough for the Web (Blackmon, Polson, Kitajima, & Lewis, 2002), and MESA (C. S. Miller & Remington, 2004). The approach of Hogg and Huberman (1987), however, provides a way of characterizing the general relation between information scent and navigation costs for the asymptotic case. Consider an idealized case of browsing for information by surfing along links in the Web. The information structure generally will be a lattice of interlinked Web pages. At each visited page, the browsing will involve choosing a link to pursue from a set of presented links. Assume that the browsing takes the form of a hierarchically arranged search tree in which branches are explored, and if unproductive, the user returns to previously visited pages. To an approximation this matches observation (Card et al., 2001). Although the Web is a lattice, the search process over that lattice tends to follow a treelike form (i.e., a spanning tree is generated by the search process over a more general graph structure of the Web). This somewhat idealized case of hierarchically organized Web search is summarized in Fig. 2. The search tree may be characterized by a branching factor b corresponding to the average number of alternatives available at each decision point. The desired target information may occur at various depths, d, in the search tree (Fig. 2).

An exhaustive tree-search process would visit every leaf node. Such a full (exhaustive) search is indicated by the complete set of lines tracing the tree structure in Fig. 2. This exhaustive tree search process would visit b^d nodes. A random tree search that terminated on encountering a target (goal) node would search about half of the tree on average. Such a random search, indicated by the light lines in Fig. 2, would visit about $b^d/2$ nodes. Information

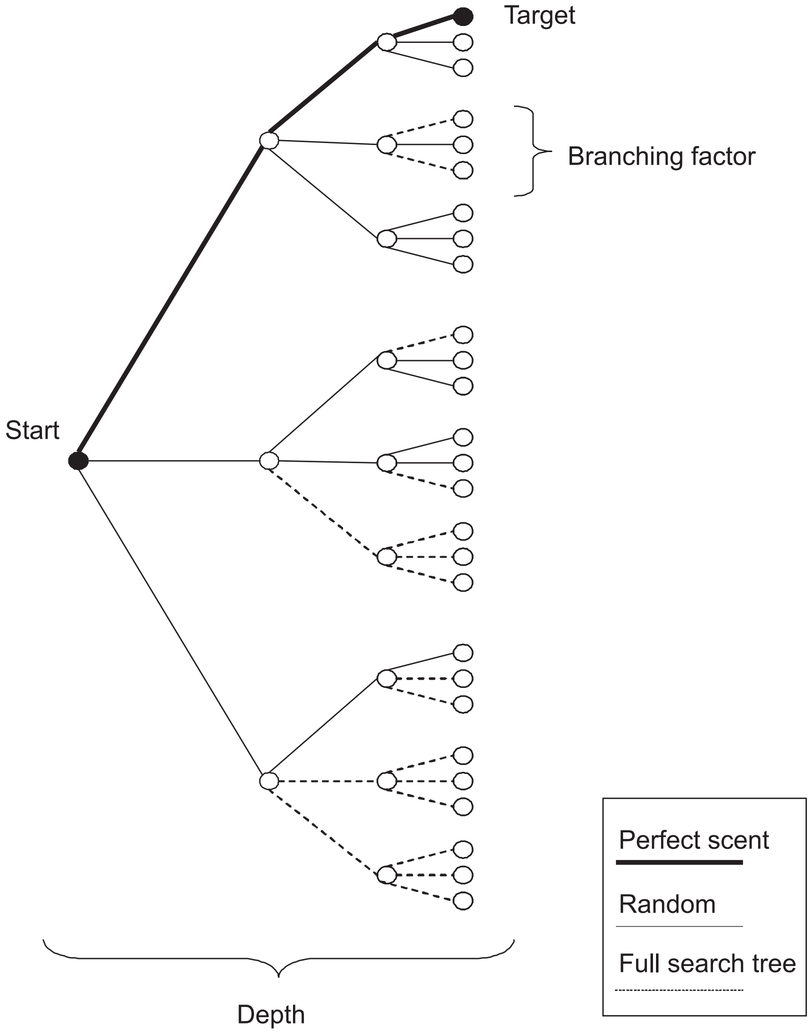


Fig. 2. Idealized search trees for heuristic search under perfect information scent (heavy solid lines) and no information scent (random, light solid lines). Dotted lines illustrate additional paths a full (exhaustive) search tree.

scent could be used as a heuristic to improve this search even further. At each node in the tree (corresponding to a Web page) some paths could be eliminated by consideration of the information scent associated with links to those paths. Improved information scent would improve the elimination of unproductive paths. If information scent were perfect, then the user would make no incorrect choices. Let us associate a false alarm factor, *f,* with the probability of failing to eliminate an incorrect link. At the extremes, perfect information scent would correspond to $f = 0$ (all wrong paths eliminated), and random guessing would correspond to $f = 1$ (no wrong paths eliminated). In Fig. 2 the heavily weighted line illustrates a search involving perfect information scent (f = 0) and the light line illustrates a random search (f = 1).

According to Hogg and Huberman (1987), the average number of nodes examined in such a hierarchical search process will be

$$N(d,b,f) = d + \left[\frac{(b-1)f}{2}\right] \sum_{s=1}^{d-1} \frac{1-(bf)^s}{1-bf} \tag{1}$$

Equation 1 captures a three-way interaction of information scent, *f,* branching factor, *b,* and depth, *d.* With perfect information scent ($f = 0$), the search cost is just *d.* In a random search ($f =$ 1) the search process must visit about half the nodes on average before target information is found.

Fig. 3 shows the effects of perturbations in false alarm factor more concretely by displaying search cost functions for a hypothetical Web site with branching factor b = 10. Search cost refers to the number of pages a user must visit before arriving at the desired page. The curves represent cost functions for links with false alarm rates of f = .015, .100, .125, and .150., which is about the range observed empirically in a study of information scent cues (Woodruff et al., 2002). One can see that the search cost regime changes very little as f ranges from .015 to .100,

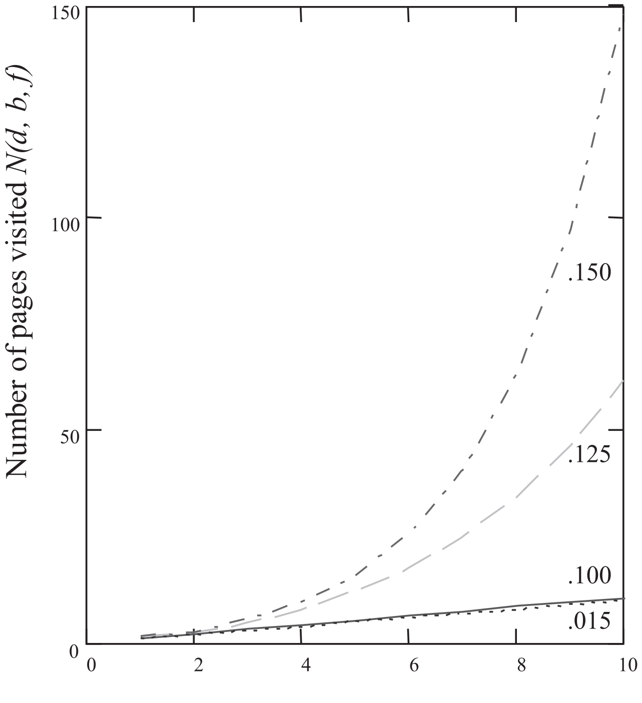


Fig. 3. Effects of perturbations of false alarm rates (f; indicated by labels next to each curve) on a hypothetical Web site with branching factor b = 10. A qualitative shift from linear search costs to exponentially increasing costs occurs at f = .100.

but changes dramatically as f becomes greater than .100. Indeed, for a branching factor of b = 10, there is a phase change from a linear search cost to an exponential search cost at the critical value of f = .100 (Hogg & Huberman, 1987). Small improvements in the false alarm factor associated with individual links can have dramatic qualitative effects on surfing large hypertext collections. Rational Web users should be motivated to maximize the accuracy of their judgments based on information scent to reduce search costs.

2.3. The hierarchical patch structure of the Web

There are a number of regularities about the distribution of links and content on the Web (e.g., Baldi, Frasconi, & Smyth, 2003). As discussed in this section, for an arbitrary task performed by a user, it appears that Web content tends to exhibit a patchy structure, in which clusters of relevant Web pages will be localized (i.e., it will be easy for a user to go from one page to another), but going from one cluster to another may require some effort. The patchy structure of the Web environment presents a kind of time allocation problem for the user: how long to forage in a patch of information before moving on to another.

One of the conventional models in the optimal foraging theory (Stephens & Krebs, 1986) deals with cases in which an organism faces an environment that has a patchy arrangement of food sources, such as birds that feed on berries that cluster on bushes that are more or less randomly scattered in the environment. The rational choice models in this area concern decisions of how long to forage in a patch (e.g., a berry bush) that exhibits diminishing returns before moving on to search for another patch. In keeping with the literature on foraging theory, I refer to groupings of information generically as *information patches*—the idea being that it is easier to navigate and process information that resides within the same patch than to navigate and process information across patches. The term *patch* suggests a locality in which within-patch navigation distances are smaller than between-patch distances. Pirolli and Card (1999) presented the classical patch model from optimal foraging theory and showed its application to user behavior in browsing a document clustering system called Scatter/Gather (Cutting, Karger, & Pedersen, 1993; Cutting, Karger, Pedersen, & Tukey, 1992).

Simon's (1962) work on the architecture of complexity argued that information systems tend to evolve toward hierarchical organizations. In part this has to do with robustness, but it also has to do with efficiency. Efficient hierarchically arranged information systems can emerge from decentralized social processes. Eiron and McCurley (2003) suggested that the link structure of the Web will tend to form localized information patches that reflect the social organizations surrounding the authorship process. For instance, the organization of groups within departments, within schools, within universities, might be reflected in a Web-site directory structure. Authors will tend to link to pages written by authors they know, and the likelihood of interauthor familiarity will decrease with tree distance in the hierarchy of the social organization, which is often reflected in the Web-site directory structure.

Empirical studies of the link structure of the Web indeed reveal a hierarchical patchy structure. In an analysis 616 million pages from 12.5 million Web sites sampled from a crawl of the Web conducted in 2002, Eiron and McCurley (2003) found that 75.0% of the links from Web pages go to other Web pages within the same Web site, and 54.8% of the within-site links go to Web pages in the same directory. Eiron and McCurley also found that the probability of occurrence of a hyperlink between two Web pages decreased exponentially with their distance in the directory structure of a Web-site host.

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 or cosine measure, *r,* on the vectors of the word frequencies in a pair of documents (Manning & Schuetze, 1999).[1] 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* =.02).

Fig. 4 illustrates these general findings of Davison (2000) in a concrete case. Fig. 4 shows how topical similarity between pages diminishes with the link distance between them. To produce Fig. 4, I used data collected from the Xerox.com Web site in May, 1998, and I computed the page-to-page content similarities for all pairs of pages at minimum distances of 1, 2, 3, 4, and 5° of separation. The similarities were computed by comparing normalized correlations of vectors of the word frequencies in a pair of documents (Manning & Schuetze, 1999). Fig. 4

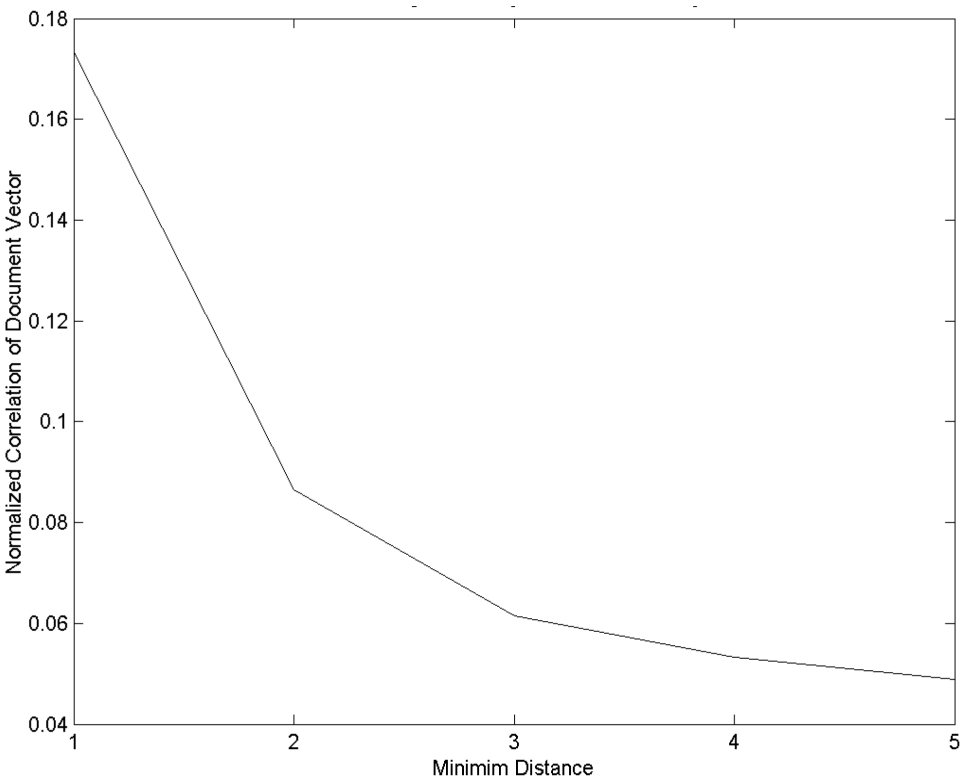


Fig. 4. The similarity (normalized correlation of document word frequency vectors) of pairs of Web pages as a function of the minimum link distances separating the pairs. Data collected from the Xerox.com Web site, May 1998.

shows that the similarity of the content of pages diminishes rapidly as a function of shortest link distance separating them.

Regardless of the ultimate evolutionary causes, it is clear that the Web is organized into a hierarchy of patches (Eiron & McCurley, 2003) and Web content is arranged into topically related patches (Davison, 2000). We may expect a rational information forager on the Web to be concerned with determining whether they are in a patch of information that is likely to yield results, or to give up and find another promising information patch.

2.4. Summary

We might assume that the adaptive forager chooses interaction methods and actions that tend to optimize the utility of information gained as a function of interaction cost, when feasible. Web users cannot have perfect knowledge of where to find desirable information. When navigating the Web by following links, users must use local proximal cues, called information scent, to make navigation choices. In general, it is expected that cues have only a probabilistic relation to distal desired information sources. The accuracy of navigation choices based on the assessments of information scent can have dramatic effects on the costs of information foraging, and an adaptive information forager might be expected to optimize such judgments and choices.

Information on the Web is arranged into hierarchical patches. Users need to detect whether a patch of information is relevant to their needs, and when to give up on a patch and seek another one. Assessing the local information scent cues over a series of visited pages might be one method for making these patch-leaving decisions. Information scent may also be a means 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 Fig. 4).

3. Rational analyses of foraging on the Web

The term *rational analysis* was inspired by rational choice theory in economics, in which people are assumed to be rational decision makers who optimize their behavioral choices in order to maximize their goals (utility). In rational analysis, however, it is not the person who is the agent of rational choice, but rather it is the selective forces of the environment that choose better biological and behavioral designs. Anderson used rational analysis to study the human cognitive architecture by assuming that natural information processing mechanisms involved in such functions as memory (Anderson & Milson, 1989) and categorization (Anderson, 1991) were well designed by evolutionary forces to meet the problems posed by the environment (see also, Oaksford & Chater, 1998).

This section presents rational analyses of some key aspects of information foraging on the Web. The first set of rational analyses concerns the use of information scent to navigate the link structure of the Web. The second set of rational analyses concerns decisions of when to leave an information patch for another. Before presenting these two sets of rational analyses, it is worth providing a general characterization of the optimization problem facing the information forager.

3.1. The principle of the extremization of information utility as a function of interaction cost

The information forager may be viewed as operating in a *task environment* and an *information environment.* The task environment (Newell & Simon, 1972) 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 external content that yields knowledge that permits people to more adaptively engage the task environment. Our particular analytic viewpoint on the information environment will be determined by the information needs that arise from the embedding task environment.

In modern society, people interact with information through technology that more or less helps them find and use the right knowledge at the right time. Information foraging theory assumes that people shape themselves and their technologies to be more adaptive in reaction to their goals and the structure and constraints of their information environments. A useful 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

$$\max\left[\frac{\text{Expected value of knowledge gained}}{\text{Cost of interaction}}\right] \tag{2}$$

This hypothesis is consistent with Resnikoff's (1989, p. 97) conclusion that natural and artificial information systems evolve toward stable states that maximize gains of valuable information per unit cost (when feasible). Cognitive systems engaged in information foraging will exhibit such adaptive tendencies, and they will prefer technologies that tend to maximize the value (or utility) of knowledge gained per unit cost of interaction. It is beyond the scope of this article to present theories of the utility of information, but discussion of this topic may be found in microeconomics (e.g., Stigler, 1961), artificial intelligence (e.g., Pearl, 1988, see especially pp. 313–327), and the foundations of cognitive science (Newell, 1990, Section 2.7).

3.2. Rational analysis of the use of information scent to navigate Web links

The rational analysis of the use of information scent assumes that the goal of the information forager is to use proximal external information scent cues (e.g., a Web link) to predict the utility of distal sources of content (i.e., the Web page associated with a Web link), and to choose to navigate the links having the maximum expected utility. This rational analysis decomposes into three parts: (a) a Bayesian analysis of the expected relevance of a distal source of content conditional on the available information scent cues, (b) a mapping of this Bayesian model of information scent onto a mathematical formulation of spreading activation, and (c) a model of rational choice that uses spreading activation to evaluate the utility of alternative choices of Web links. This rational analysis yields a spreading activation theory of utility and choice.

3.2.1. *Bayesian analysis of information scent*

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. The rational analysis of information scent is framed by a different assumption—that the information forager is making predictions about the expected value of different external actions—but it ends up with a derivation that parallels the rational analysis of human memory.

Fig. 5 presents an example for the purposes of discussion in this section. Fig. 5 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 that a distal source of content contains desired information based on the proximal cues available in the hypertext link labels.

Bayes's theorem can be applied to the information foraging problem posed by situations such as those in Fig. 5. The probability that a distal content structure has desired information features, *D,* given a structure of proximal features, *P,* can be stated using Bayes's theorem as,

$$\Pr(P|D) = \Pr(D) \bullet \Pr(P|D) \tag{3}$$

where Pr($P|D$) is the *posterior probability* of distal content 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's theorem in Equation 3 is

$$\frac{\Pr(D \mid P)}{\Pr(D \mid \sim P)} = \frac{\Pr(D)}{\Pr(\sim D)} \cdot \frac{\Pr(P \mid D)}{\Pr(P \mid \sim D)} \tag{4}$$

where Pr($D|{\sim}P$) is the probability of distal content 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($P \mid {\sim}D$) is the probability of P occurring given that D does not occur.

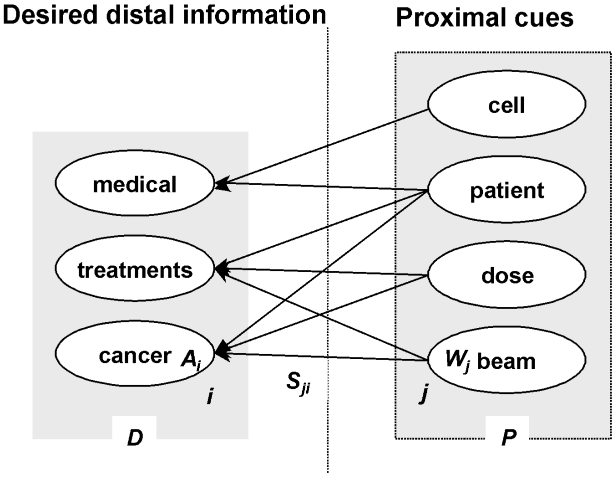


Fig. 5. A cognitive structure representing some desired goal information (D) and an encounter with some proximal information scent cues (P).

For concreteness, assume that the distal content structure *D* is the information contained on a Web page and that we represent *D* as a set of concepts corresponding to an information goal. Assume, likewise, that the proximal information scent structure *P* is represented as a set of the concepts corresponding to the words (images, etc.) on a Web link.

Following Anderson and Milson (1989), we make simplifying independence assumptions,[2] so that Equation 4 may be written as an odds equation for each individual feature *i* of the distal content structure *D* and each individual feature *j* of the proximal information scent structure,

$$\frac{\Pr(i \mid P)}{\Pr(\sim i \mid P)} = \frac{\Pr(i)}{\Pr(\sim i)} \cdot \prod_{j \in P} \frac{\Pr(i \mid j)}{\Pr(i \mid \sim j)} \tag{5}$$

where Pr(*i*|*j*) is the conditional probability of a distal concept or feature *i* occurring given that a proximal concept or 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.* Pr(*i*|~*j*) is the probability of *i* occurring when *j* does not occur. We may assume (as do Anderson & Milson, 1989) that usually the occurrence of concepts *i* and *j* together is extremely small relative to the occurrence of concept *i.* This means that the likelihood of concept *i* given that concept *j* is in among the information scent cues is approximately

$$\frac{\Pr(i \mid j)}{\Pr(i \mid \sim j)} \approx \frac{\Pr(i \mid j)}{\Pr(i)} \tag{6}$$

The approximation in Equation 6 is known as Pointwise Mutual Information (PMI) in the information retrieval literature (Manning & Schuetze, 1999, p. 178). As discussed later, PMI has been shown to be a very good predictor of human judgments of the similarity of word *i* and *j.*

With the approximation in Equation 6, Equation 5 may be rewritten as

$$\frac{\Pr(i \mid P)}{\Pr(\sim i \mid P)} = \frac{\Pr(i)}{\Pr(\sim i)} \cdot \prod_{j \in P} \frac{\Pr(i \mid j)}{\Pr(i)} \tag{7}$$

Taking the logarithms of both sides of Equation 5 leads to an additive formula,

$$\log\left[\frac{\Pr(i \mid j)}{\Pr(\sim i \mid j)}\right] = \log\left[\frac{\Pr(i)}{\Pr(\sim i)}\right] + \sum_{j} \log\left[\frac{\Pr(j \mid i)}{\Pr(i)}\right] \tag{8}$$

or

$$A_i = B_i + \sum_{j} S_{ji}, \tag{9}$$

where

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

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

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

3.2.2. Mapping the Bayesian rational analysis to spreading activation

Equation 9 provides the rational grounds for a spreading activation theory 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).

Fig. 5 presents a scenario for a spreading activation analysis. The chunks representing proximal cues are presented on the right side of Fig. 5. Fig. 5 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 9.

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

$$A_i = B_i + \sum_j W_j S_{ji}. \tag{10}$$

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 reflects attention (*source activation*) on chunk *j.* Note that Equation 10 reflects the log odds Equation 9 but now includes a weighting factor *W* that characterizes capacity limitations of human attention. One may interpret Equation 10 as reflection of a Bayesian prediction of the likelihood of one chunk in the context of other chunks. A_i in Equation 10 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 representation of a user's information goal provides an indicator of the likelihood that a distal source of information has desirable features based on the information scent cues immediately available to the user.

3.2.3. Relating the spread of activation to the evaluation of the utility of information foraging choices

To map this spreading activation model of information scent onto a model of rational choice (of navigation actions) involves the use of a random utility model (McFadden, 1974, 1978).[3] Random utility model (RUM) theory is grounded in classic microeconomic theory, and it has relations to psychological models of choice developed by Thurstone (1927) and Luce (1959). Its recent developments are associated with the microeconomic work of McFadden (1974, 1978).

For our purposes, a RUM consists of assumptions about (a) the characteristics of the information foragers making decisions, including their goal or goals, (b) the choice set of alternatives, (c) the proximal cues 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, $i \in G$ (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 Wj. 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

$$U_{J|G} = V_{J|G} + \varepsilon_{J|G} \tag{11}$$

where

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

is the summed activation, and where $\varepsilon_{J|G}$ is a random variable error term reflecting a stochastic component of utility. Thus, the utility $U_{J|G}$ is composed of a deterministic component $V_{J/G}$ and a random component $\varepsilon_{J|G}$. 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,

$$\Pr(J \mid G, C) = \Pr(U_{J|G} \geq U_{K|G}, \forall K \in C) \tag{12}$$

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), then Equation 12 takes the form of a multinomial logit

$$\Pr(J \mid G, C) = \frac{e^{\mu V_{J|G}}}{\sum_{K \in C} e^{\mu V_{K|G}}} \tag{13}$$

where μ is a scale parameter.[4] If there is only one alternative to choose (e.g., select J or do not select J), then Equation 12 takes the form of a binomial logit,

$$\Pr(J \mid G, C) = \frac{1}{1 + e^{\mu V_{J|G}}} \tag{14}$$

For a navigation judgment, we can now specify how the computation of spreading activation yields utilities by substituting Equation 10 into Equation 11:

$$\begin{aligned} U_{J|G} &= V_{J|G} + \varepsilon_{J|G} \\ &= \sum_{i \in G} A_i + \varepsilon_{J|G} \\ &= \sum_{i \in G} \left(B_i + \sum_{j \in J} W_j S_{ji} \right) + \varepsilon_{J|G} \end{aligned} \tag{15}$$

Equations 13, 14, and 15 provide a microeconomic model for the choice of Web links that is grounded in an underlying cognitive model of utility.

It should be emphasized that this theory of information scent is not a part of the standard ACT–R theory. The spreading activation model of information scent is not the same as the ACT–R theory of spreading activation, and the utility model based on information scent is not the same as the ACT–R model of utility. In ACT–R, spreading activation emits from goal chunks and is used to retrieve relevant chunks from memory. In the theory of information scent, cues from the external world are sources of activation; the activation levels of goal chunks are used to assess utility. In ACT–R, the utility assessments are based on the past successes and failures of actions. In the theory of information scent, the utility assessments are based on spreading activation. It is perhaps better to think of the spreading activation theory of information scent as being a kind of rational analysis of the categorization of cues according to their expected utility.

3.2.4. Estimating spreading activation strengths using PMI

As discussed in Pirolli and Card (1999), it is possible to automatically construct large spreading activation networks from on-line text corpora. In other words, we may analyze samples of the linguistic environment to provide the parameters of our cognitive models, a priori, rather than estimating those parameters, a posteriori, by fitting the models to behavioral data. The frequency of occurrence of words, and the co-occurrence frequency of words near one another can be used to estimate the base strengths, $B_{i,}$ and interchunk strengths, S_{ji} in Equation 10. In the SNIF–ACT model (Pirolli & Fu, 2003) discussed later, we estimated these strengths from the Tipster document corpus (Harman, 1993) and from the Web using a program that calls on the AltaVista search engine to provide data. Recently, Turney (2001) showed that PMI scores (which approximate S_{ji}) computed from the Web can provide good fits to human word-similarity judgments.

3.3. Rational analysis of foraging in information patches

What is known as the *conventional patch model* in optimal foraging theory (Stephens & Krebs, 1986) deals with the optimal policy for the amount of time to spend in food patches before leaving. Applications of variations of that model to information foraging are discussed in Pirolli and Card (1999). Here I present another variation that has ties to work investigating stochastic models of food foraging (McNamara, 1982) and studies of the aggregate behavior of large numbers of users browsing on the Web (Huberman, Pirolli, Pitkow, & Lukose, 1998).

3.3.1. A stochastic model of patch foraging

Assume that the experiential state of the information forager at time *i* is represented as a state variable $\mathbf{X}_{i,}$ and $\mathbf{X}_{i,}= \mathbf{x}$. is a particular state value. For current purposes, assume that this state variable includes some representation of the Web page that has just been revealed and perceived by the information forager. The utility *U* as a function of this user state, $U(\mathbf{x})$, of continued (optimal) link browsing in this information patch can be represented by the expectation,

$$U(\mathbf{x}) = E[U|\mathbf{X}_i = \mathbf{x}] \tag{16}$$

In keeping with the previous discussion, we might assume that $U(\mathbf{x})$ is determined by choosing links having the maximum expected stochastic utility according to the RUM model. The expected time cost, *t,* of future (optimal) link browsing can be represented as an expectation

$$t = E[T|\mathbf{X}_i = \mathbf{x}] \tag{17}$$

where *T* is a random variable representing future time costs. The value $U(\mathbf{x})$ of foraging for time *t* in this information patch (e.g., Web site) must be balanced against the opportunity cost

$C(t)$ of foraging for that amount of time. This defines what McNamara (1982) called the *potential function,* $h(\mathbf{x})$, for continued foraging in this patch,

$$h(\mathbf{x}) = U(\mathbf{x}) - C(t) \tag{18}$$

and the optimal forager is one who maximizes this potential function.

McNamara's (1982) model characterizes the opportunity cost $C(t)$ in terms of the overall long-term average rate of gain of foraging, R^*,

$$C(t) = R * t \tag{19}$$

In the case of information foraging on the Web, we might assume this refers to the overall average long-term rate of gain of foraging on the Web for similar tasks. The intuition behind Equations 18 and 19 is that the utility of foraging in this patch must be greater than or equal to the average rate of returns for foraging (otherwise continued foraging is incurring an opportunity cost). In other words, an information forager should continue foraging so long as,

$$U(\mathbf{x}) - R * t > 0 \tag{20}$$

The overall average rate of gain, R^*, could be characterized in terms of the mean utility $\overline{U}$ of going to a relevant Web site, the mean time spent setting up to go to the next relevant site (e.g., by using a search engine or guessing and typing URLs), $\bar{t}_s$, and the mean time spent foraging at the next new site, $\bar{t}$,

$$R^* = \frac{\bar{U}}{\bar{t}_s + \bar{t}} \tag{21}$$

Assuming Equation 21, we may rewrite the Inequality 20 as a rule to continue foraging so long as the rate of gain from this information patch (e.g., Web site) is greater than the expected rate of gain of going to another relevant information patch.

$$\frac{U(\mathbf{x})}{t} > R^* = \frac{\bar{U}}{\bar{t}_s + \bar{t}} \tag{22}$$

This decision to stop foraging in an information patch when the expected rate of gain drops below R^* is a stochastic version of the patch-leaving rule in the conventional patch model known as Charnov's marginal value theorem (Charnov, 1976), which was related to information foraging behavior in Pirolli and Card (1999). Note that the discussion here has implicitly assumed that the information forager has perfect knowledge of the relevant environmental values in Equations 19 through 22 (i.e., any learning is near asymptote). McNamara (1982) discussed how learning might be incorporated into this patch-leaving rule.

Note in Equation 22 that the average time, t, to go to a Web page that is within the same Web site as this page being visited may often be approximately the same as the time to go to a Web page at another Web site, $t_{s+}\bar{t}$. In such cases the decision rule to continue foraging in Equation 21 could be reduced to $U(\mathbf{x}) > \overline{U}$.

In English, this rule says

> forage in an information patch until the expected potential of that patch is less than the mean expected value of going to a new patch.

In the SNIF–ACT model, it is assumed that the expected potential of a patch is estimated from the links available on Web pages, relying again on spreading activation from information scent for this assessment. SNIF–ACT also assumes that the average expected value of a new patch is estimated from past experience on the Web.

3.3.2. From individual rationality to the aggregate behavior of Web foraging: The law of surfing

One interesting consequence of this formulation of foraging in information patches is that it leads to predictions (Huberman et al., 1998) concerning patterns of aggregate behavior on the Web—specifically what has been called the *Law of Surfing.* Such aggregate distributions are of practical interest because content providers on the Web often want to know how long people will remain at their Web sites (often referred to as the *stickiness* of a Web site). More generally, the ability to relate predictions about the emergent behavior of populations from the rational models of individuals is a way of bridging psychological science and the microeconomics of the Web.

The Law of Surfing characterizes the distribution of the length, *L,* of sequences of page visits by Web users (see also, Baldi et al., 2003, pp. 194–199). The Law of Surfing is based on the assumption that Web users can be modeled by a random walk process in which the expected utility from continuing on to the next state is stochastically related to the expected utility of this state,

$$U(X_t) = U(X_{t-1}) + \varepsilon_t \tag{23}$$

where ε_t are independent identically distributed (IID) Gaussian distributions. Note that we are discussing aggregates of different surfers with different stochastic utility functions. The constraint of IID Gaussian noise is expected from the central limit theorem. Fig. 4, which presents the similarity of Web pages as a function of number of degrees of separation supports the assumption that the utilities of adjacent pages are related as assumed in Equation 23.

From an initial page at a Web site we expect users to continue browsing (surfing) following the random walk specified in Equation 23 until the threshold specified in Inequality 20 is reached. In the limit this is the same as the analysis of first passage times in Brownian motion. The probability density function of *L,* the length of sequences of Web page visits, is distributed (as are the first passage times in Brownian motion) as an inverse Gaussian distribution (Seshardri, 1993),

$$f(L)=\sqrt{\frac{\lambda}{2\pi}}L^{-3/2}e^{-\frac{\lambda}{2\mu^2 L}(L-\mu)^2} \tag{24}$$

where the parameter μ is the mean, and λ is related to the mean and variance as $\lambda = \dfrac{\mu^3}{Var[L]}$.

The inverse Gaussian is a skewed distribution that looks very much like the more familiar lognormal distribution, and it has a number of interesting properties that are discussed later in application to data. Web-site developers are often interested in the amount of content that users will visit on their Web site (or the amount of time they will spend). The Law of Surfing is relevant to characterizing these data of interest.

3.4. Summary

The general problem posed by the Web environment is one of maximizing the gain of valuable knowledge in relation to the cost of information foraging. One significant subproblem concerns the rational choice of navigation links based on available information scent cues associated with those links. Another subproblem concerns the rational allocation to time to forage in a patch before moving on to another. This section presented rational analyses of behavioral solutions to these environmental problems that will guide the SNIF–ACT cognitive model presented in the next section.

The analyses presented here extend the rational analysis approach (e.g., Anderson, 1990; Oaksford & Chater, 1998) in several ways. The first extension includes a new rational analysis of navigation in information systems based on *information scent* (Pirolli, 1997). A second extension is one of focusing rational analyses on behavior that occurs in tasks that take substantially more time than the traditional focus of rational analysis. A third extension of rational analysis has been to relate the rational analysis of individual psychology to the emergent behavior of populations.

4. Empirical evaluations

This section begins with evaluations of how well spreading activation predicts human judgments of the expected utility of browsing actions. This spreading activation model of information scent is central to the SNIF–ACT model, which is presented next along with empirical evaluations of the model. The SNIF–ACT model implements strategies that approximate the rational analyses of link choice and information-patch leaving discussed previously. This section ends with an empirical evaluation of the Law of Surfing described previously.

4.1. Information scent

Pirolli and Card (1999) presented a cognitive model, ACT–Information Foraging (ACT–IF), that simulated users in a controlled experiment in which they performed information-seeking tasks with a browser called Scatter/Gather (Cutting et al., 1992, 1993). In that controlled experiment (also reported in, Pirolli, Schank, Hearst, & Diehl, 1996), each participant was required to collect as many documents as possible relevant to a sample of tasks generated by information retrieval experts (used in the Text Retrieval Conference [TREC] workshops; Harman, 1993). ACT–IF simulated the navigation actions of these participants on the Scatter/Gather browser using spreading activation networks that were automatically computed from the base rate frequencies and co-occurrence frequencies of words in the Tipster corpus.[5]

As a side evaluation of these automatically computed spreading activation networks, Pirolli and Card (1999) compared information scent computations from the ACT–IF simulation to participant ratings of the expected utility of *cluster summaries* presented on Scatter/Gather windows. Each cluster summary was a small piece of text that Scatter/Gather used to summarize a topically related set of documents to the user, for instance:

Cluster-0 (38940) cell, patient, radiation, dose, beam, disease, treatment,

> AP: Early Results In Hospital Patient Study Sho (aid, study, percentage, health,).
> DOE: Doses of secondary radiation appearing as (radiation, dose, exposure).
> AP: Poll: AIDS Test Confidentiality Opposed in (percentage, study, drug, report).

Each such cluster summary began with a line indicting the cluster label (cluster-0), the number of documents in the cluster (38,940), and a set of keywords summarizing the central concepts for the cluster of documents (cell, patient, radiation, dose, beam, disease, treatment). These keywords were automatically computed. The next three lines summarize the three most representative documents in the cluster by their source (e.g., “AP”), some words from the title, and a list of words representing the central concepts in the document.

Each Scatter/Gather window usually presented 10 of these cluster summaries, and each set of documents summarized by a cluster summary might range from a few documents to several hundred thousand documents. As described in detail elsewhere (Pirolli et al., 1996) users interacted with the system by collecting together and decomposing clusters to get to small sets of highly relevant documents. The side study in Pirolli and Card (1999) asked users to estimate the percentage of relevant documents contained in the set of documents represented by cluster summaries presented in Scatter/Gather windows that the users encountered as they pursued their experimental tasks. The comparison of users’ stated estimates to scores predicted by spreading activation in our ACT–IF simulation is presented in Fig. 6. ACT–IF used one scaling parameter used to adjust the empirical frequencies obtained from Tipster to set the spreading activation networks. The fit of ACT–IF to observed browsing behavior reported in Pirolli and Card (1999) contained no other free parameters estimated from user data. The fit to the user data in Fig. 6 uses the same spreading activation networks and a linear regression to map the activation values onto percentages.

Blackmon et al. (2002) also had success in modeling Web navigation behavior using latent semantic analysis (LSA) as a way of computing information scent, rather than using the specific spreading activation model proposed previously. Katz and Byrne (2003) also used LSA as a way of calculating information scent in a study of users’ preferences for using search or browsing. LSA is a technique that provides good fits to human similarity judgments for word pairs. In application to modeling Web browsing, the assumption is that the judgments about the relevance of Web links are the same thing as judging the similarity of the Web link to a user’s goal.

It appears that interchunk association strengths (S_{ij} in Equation 10) estimated by calculating PMI scores, yield similarity judgments that are comparable to LSA. Turney (2001) computed PMI scores (which approximate strength of association scores) using a standard Web search engine, by estimating P(i|j) from the number of documents in which the two words i and j occurred and P(i) from the number of documents in which just i occurred. On a version of the col-

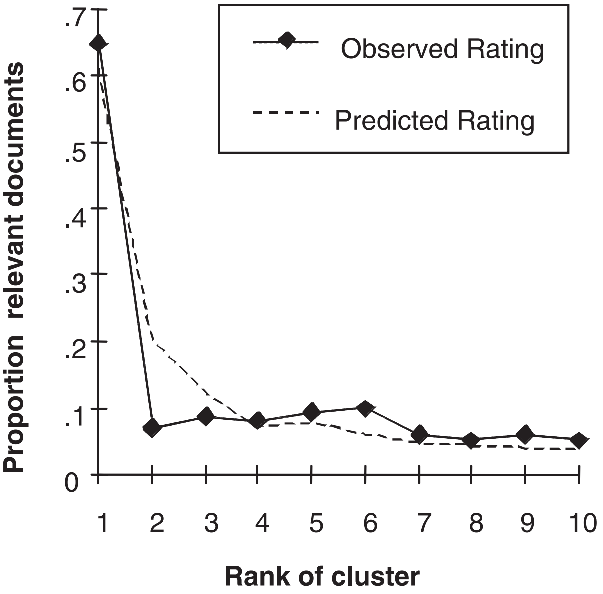


Fig. 6. Observed and predicted judgments of the expected proportion of relevant documents in a Scatter/Gather document cluster based solely on the text available in cluster summarized in Scatter/Gather windows. Observed data from $N = 24$ cluster summary ratings and predicted ratings from the ACT–IF spreading activation model of information scent operating on the same cluster summaries.

lege Test of English as a Foreign Language (TOEFL) of word similarity, students score 64.5%, LSA 64.4%, and Web-based PMI (or, equivalently, association strengths) 62.5%. Farahat, Pirolli, and Markova (2004) developed an efficient method of computing PMIs from a combination of a locally stored subset of the Web, with the Web used as a back-off for computing word similarities that are not stored locally, which scores 66.0% on the TOEFL test and outperforms LSA on other tests of human word-similarity judgments.

4.2. SNIF–ACT: Scent-based navigation and information foraging

A model called SNIF–ACT (Pirolli & Fu, 2003) was developed to simulate the participants in a Web study described in Card et al. (2001). Card et al. summarized extensive analyses of verbal protocol, WebLogger logs (which trace user and browser actions), and eye-tracking data collected from 4 participants in a Web study who worked on two tasks designed to be similar to ones reported by real-world users (Morrison et al., 2001). The tasks presented to users were

- *City task.* You are the chair of Comedic Events for Louisiana State University in Baton Rouge. 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.

- *Antz task.* 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.

A *user-tracing architecture* (Pirolli, Fu, Reeder, & Card, 2002) was developed to match the SNIF–ACT simulation to the detailed WebLogger logs and coded verbal protocols from the 4 participants working on these two tasks.

SNIF–ACT extends the ACT–R theory and simulation environment (Anderson & Lebiere, 2000). SNIF–ACT may also be considered as an extension of the ACT–IF model of the Scatter/Gather users discussed previously (Pirolli & Card, 1999), which uses the same information scent mechanisms but different production rules. ACT–R contains three kinds of assumptions about (a) knowledge representation, (b) knowledge deployment (performance), and (c) knowledge acquisition (learning).

There are two major memory components in the ACT–R architecture: a *declarative knowledge* component and a *procedural knowledge* component. Declarative knowledge corresponds to things that we are aware we know and that can be easily described to others, such as the state of a browser, the content of Web links, the functionality of browser buttons, or information goals. Declarative knowledge is represented formally as chunks in ACT–R. Declarative chunks in ACT–R have activation values. Activation spreads from this 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 specifies how declarative knowledge is transformed into active behavior. Procedural knowledge is represented formally as condition–action pairs, or *production rules.* 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 him or her. 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 this 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 the spreading activation model of information scent.

4.2.1. Utility and choice by information scent

In a SNIF–ACT simulation, information scent cues on a computer display activate chunks, and activation spreads 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 user data from Card et al. (2001). The major controlling variable in the model is the measure of information scent, which predicts two major kinds of actions: (a) which links on a Web page people will click on, and (b) when people decide to leave a site. We called the first kind of action *link-following* action, which was logged whenever a participant clicked on a link on a Web page. The second kind of action was called *site-leaving* action, which was 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.

4.3. 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 ac-

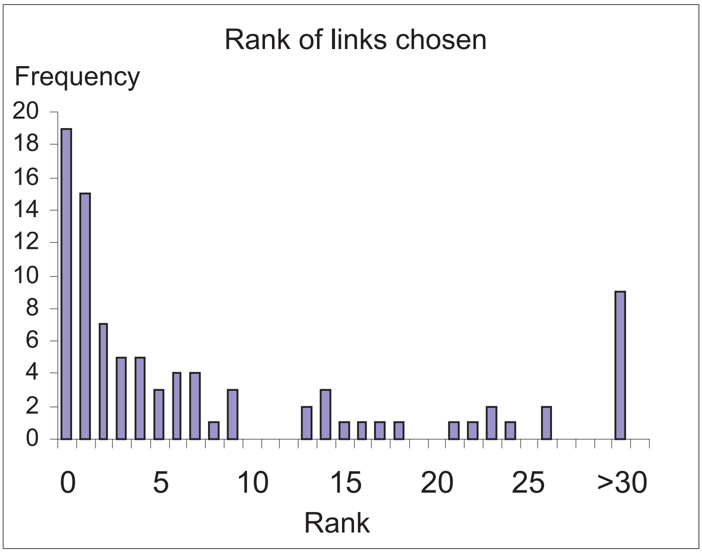


Fig. 7. Frequency that SNIF–ACT productions match link-following actions. The SNIF–ACT production rankings are computed at each simulation cycle over all links on the same Web page and all productions that match. The rankings of link-choice actions were produced by the spreading activation mechanisms for judging information scent.

tion 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 relation 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 . 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.

Fig. 7 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 Fig. 7. The distribution of the predicted link selection rates was significantly different from random selection χ^2 (30) = 18,589, $p < .0001$. This result replicates a similar analysis made by Pirolli and Card (1999) concerning the ACT–IF model prediction of cluster selection in the Scatter/Gather browser. The ability of the spreading activation model of information scent in SNIF–ACT to predict link choice on the Web supports the rational analysis presented around Equation 10.

4.4. 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

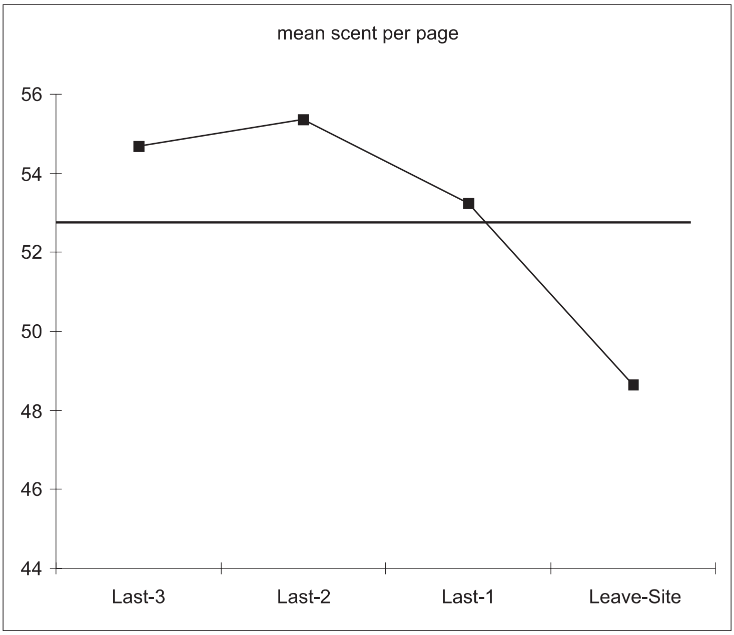


Fig. 8. Mean information scent of final four Web pages visited immediately before users left a Web site. Information scent was computed by spreading activation mechanisms. The dotted line indicates the average information scent of the first page visited on a new Web site.

Fig. 8. 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 Fig. 8 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.

Fig. 8 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. Fig. 8 suggests that spreading activation mechanisms compute activation values from information scent cues in order to reflect expected utilities associated with navigation choices. The participants in this Web study appeared to be following the patch-leaving rule developed around Equation 22.

4.5. Information patch-leaving and the Law of Surfing

SNIF–ACT is a cognitive model that approximates the rational analysis of individual Web foraging behavior. The Law of Surfing relates the rational analysis to the aggregate behavior. The Law of Surfing was tested in Huberman et al. (1998) and in addition validated in Lukose and Huberman (1998). Here, I review a couple of the evaluations in Huberman et al. to illustrate the general findings. Fig. 9 presents a typical empirical distribution of the length of paths taken by visitors to a Web site, along with a fitted inverse Gaussian distribution. Note the long positive tail, which yields a mean that is typically much larger than the mode. Fig. 10 plots an-

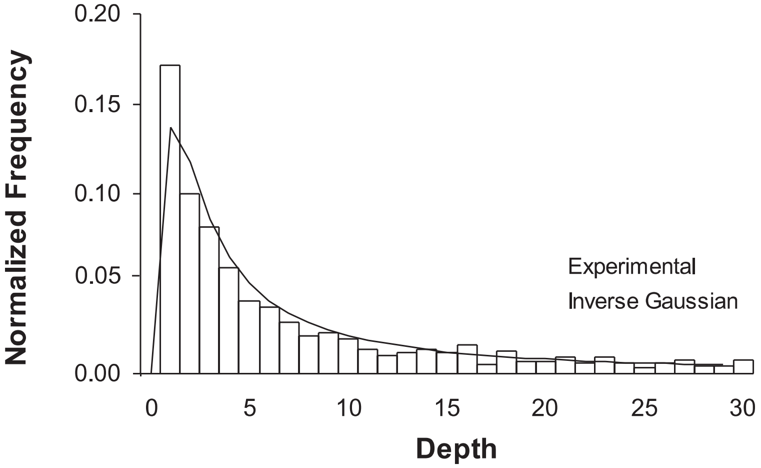


Fig. 9. The normalized frequency distribution of users as a function of depth of surfing. The observed data were collected at Boston University during late 1994 and early 1995. The fitted Inverse Gaussian distribution has a mean of = 51.19 and λ = 3.53.

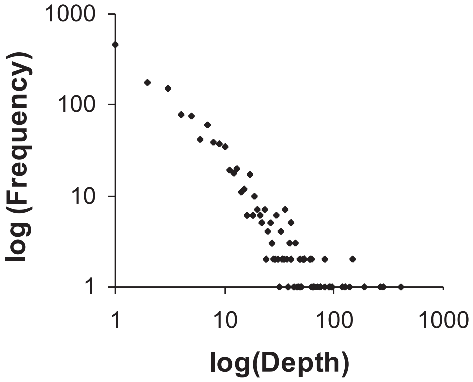


Fig. 10. The frequency distribution of surfing depths on log–log scales. Data collected from The Georgia Institute of Technology, August 1994.

other distribution of path lengths from another Web site in log-log coordinates. Figure 10 has the empirical appearance of a power-law distribution with an exponent of approximately –3/2. It turns out (Huberman et al., 1998) that this is characteristic of the inverse Gaussian distribution when $\mu <$ Var[L].

5. General discussion

Recently, Anderson (2002) drew on research in education to analyze this success of cognitive modeling. Like education, the study of information foraging provides another opportunity to develop scientific theories that simultaneously span behavioral phenomena that occur at the grain size of 100 msec to those that take 100 hr to unfold, and to produce theories and models that have practical relevance. For instance, this article integrates models of behavior associated with judgments about the relevance of individual Web links (which take approximately 1 sec) with models of Web navigation on tasks taking about .5 hr. These analyses could be integrated with "downward" models of the behavior associated with eye movements over pages, where there is already a promising rational analysis (Young, 1998), and "upward" to longer term information foraging tasks. Sense-making tasks, such as intelligence analysis (Pirolli & Card, 1999), involve trade-offs of exploration, enrichment, and exploitation over spans of hours to days to months to years that may be amenable to rational analyses and cognitive models.

Elsewhere, Anderson et al. (2004) revived the goal of "parameter-free" cognitive models (Card, Moran, & Newell, 1983). Often, cognitive models in psychology are fit to data by estimating parameters from the data themselves (e.g., through curve fitting). Information foraging models such as SNIF–ACT (Pirolli & Fu, 2003) and ACT–IF (Pirolli & Card, 1999) have achieved good fits to data using parameters set, a priori, from analysis of the information environment. The entire model of information scent and utility, save for a scaling factor estimated for ACT–IF, is determined by statistical analysis of natural language in the environment.

Theories of information foraging on the Web might be expected to improve current technology designs. The notion of information scent has already found its way into the vocabulary of Web designers (User Interface Engineering, 1999). Information foraging theory and the notion of information scent has be used to design improved search and browsing tools for the Web (Olston & Chi, 2003). Cognitive engineering models have been developed to automatically assess browser designs (Pirolli, 1998) and Web-site design (Chi et al., 2003). Given the demands in private industry and public institutions to improve the Web, and sparsity of relevant psychological theory, there is likely to be continuing demand for scientific foundations that may improve commerce and public welfare.

More generally, improvements to HII are improvements to human intelligence. This claim can be analyzed by working through Newell's (1990) discussions of knowledge and intelligence. Newell proposed that "intelligence is the ability to bring to bear all the knowledge that one has in service of one's goals" (p. 90).[6] By knowledge, Newell meant something we can attribute to a rational person that would be used to achieve his or her goals. Newell conceived of pure knowledge in a manner that transcended physical information processing limitations: In the idealized view of knowledge, everything in a body of knowledge (including all possible implications) is instantly accessible. However, people, or any physical system, can only approximate such perfect intelligent use of knowledge because the ability to bring forth the right knowledge at the right time is limited. The laws of physics limit the amount of information that can be stored or processed in a circumscribed location of space and time. Within those limits, however, intelligence increases with the ability to bring to bear the right knowledge at the right time.

Newell's (1990) discussions focused on unaided intelligent systems (people or computer programs) and the knowledge that they had available in their local memories. But there is a sense in which the world around us provides a vast external memory teeming with knowledge that can be brought forth to remedy a lack on the part of the individual. We can extend Newell's notion of intelligence and argue that intelligence is improved by enhancement of our ability to bring forth the right knowledge at the right time from the external world. Of course, the world (both physical and virtual) shapes the manner in which we can access and transform knowledge-bearing content, and thus shapes the degree to which we reason and behave intelligently. Psychological theories that provide a foundation for improved HII can provide a foundation for improving human intelligence.

Notes

1. Davison used two additional measures that yielded similar results.
2. Specifically (a) that the probability that the distal information is relevant to the user's goal is just the product of the probabilities that the individual components of the distal information are relevant to the goal, and (b) that the probability of one element *i* conditional on another element *j* is independent of all the other elements.
3. The information scent approach originally developed in Pirolli and Card (1999) is consistent with the RUM, although it was not recognized in that article.
4. 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.
5. These statistics were provided to us by our colleague Hinrich Schuetze.
6. Newell's technical definition was that "A system is *intelligent* to the degree that it approximates a knowledge-level system" (Newell, 1990, p. 90).

Acknowledgments

Portions of this research have been supported by an Office of Naval Research Contract No. N00014–96-C-0097 to P. Pirolli and S.K. Card and Advanced Research and Development Activity, Novel Intelligence from Massive Data Program Contract No. MDA904–03-C-0404 to S.K. Card and Peter Pirolli.

I would like to thank the reviewers Mark Steyvers and Michael Lee for their very constructive comments and suggestions. I also would like to thank Julie Heiser for her comments on earlier drafts. Hinrich Schuetze and Ed Chi provided technical assistance in generating data for the information scent computations.

References

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

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

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

Anderson, J. R. (2002). Spanning seven orders of magnitude: A challenge for cognitive modeling. *Cognitive Science, 26,* 85–112.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of mind. *Psychological Review, 11,* 1036–1060.

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

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, England: Wiley.

Bhavnani, S. K. (2002, April). *Domain-specific search strategies for the effective retrieval of healthcare and shopping information.* Paper presented at the Conference on Human Factors and Computing Sytems, Minneapolis, MN.

Blackmon, M. H., Polson, P. G., Kitajima, M., & Lewis, C. (2002). Cognitive walkthrough for the Web. *CHI 2002, ACM Conference on Human Factors in Computing Systems, CHI Letters, 4,* 463–470.

Card, S., Pirolli, P., Van Der Wege, M., Morrison, J., Reeder, R., Schraedley, P., et al. (2001). Information scent as a driver of Web behavior graphs: Results of a protocol analysis method for web usability. *CHI 2001, ACM Conference on Human Factors in Computing Systems, CHI Letters, 3,* 498–505.

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

Charnov, E. L. (1976). Optimal foraging: The marginal value theorem. *Theoretical Population Biology, 9,* 129–136.

Chi, E. H., Rosien, A., Suppattanasiri, G., Williams, A., Royer, C., Chow, C., et al. (2003). The Bloodhound Project: Automating discovery of Web usability issues using the InfoScent simulator. *CHI 2003, ACM Conference on Human Factors in Computing Systems, CHI Letters, 5,* 505–512.

Cutting, D. R., Karger, D. R., & Pedersen, J. O. (1993, July). *Constant interaction-time Scatter/Gather browsing of very large document collections.* Paper presented at the Sixteenth Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR 93, New York.

Cutting, D. R., Karger, D. R., Pedersen, J. O., & Tukey, J. W. (1992, June). *Scatter/gather: A cluster-based approach to browsing large document collections.* Paper presented at the Proceedings of the Fifteenth Annual International ACM Conference on Research and Development in Information Retrieval, SIGIR 92, New York.

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

Eiron, N., & McCurley, K. S. (2003, May). *Locality, hierarchy, and bidirectionality in the Web.* Paper presented at the Workshop on Algorithms and Models for the Web Graph, WAW 2003, Budapest, Hungary.

Farahat, A., Pirolli, P., & Markova, P. (2004). *Incremental methods for computing word pair similarity* (Tech. Rep. No. TR-04-6-2004). Palo Alto, CA: PARC.

Glimcher, P. W. (2003). *Decisions, uncertainty, and the brain: The science of neuroeconomics.* Cambridge, MA: MIT Press.

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

Hogg, T., & Huberman, B. A. (1987). Artificial intelligence and large-scale computation: A physics perspective. *156,* 227–310.

Huberman, B. A., Pirolli, P., Pitkow, J., & Lukose, R. J. (1998). Strong regularities in World Wide Web surfing. *Science, 280,* 95–97.

Katz, M. A., & Byrne, M. D. (2003). Effects of scent and breadth on use of site-specific search on e-commerce Web sites. *ACM Transactions on Computer-Human Interaction, 10,* 198–220.

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

Lukose, R. M., & Huberman, B. A. (1998, October). *Surfing as a real option.* Paper presented at the International Conference on Information and Computation Economies, Charleston, NC.

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

Mayr, E. (1983). How to carry out the adaptationist program? *American Naturalist, 121,* 324–334.

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

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* (pp. 75–96). Cambridge, MA: Harvard University Press.

McNamara, J. (1982). Optimal patch use in a stochastic environment. *Theoretical Population Biology, 21,* 269–288.

Miller, C. S., & Remington, R. W. (2004). Modeling information navigation: Implications for information architecture. *Human Computer Interaction, 19,* 225–271.

Miller, G. A. (1983). Informavores. In F. Machlup & U. Mansfield (Eds.), *The study of information: Interdisciplinary messages* (pp. 111–113). New York: Wiley.

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. *CHI 2001, ACM Conference on Human Factors in Computing Systems, CHI Letters, 3,* 163–164.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Harvard University Press.

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

Oaksford, M., & Chater, N. (Eds.). (1998). *Rational models of cognition.* Oxford, England: Oxford University Press.

Olston, C., & Chi, E. H. (2003). ScentTrails: Integrating browsing and searching on the Web. *ACM Transactions on Computer-Human Interaction, 10,* 177–197.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference.* Los Altos, CA: Kaufmann.

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

Pirolli, P. (1998, April). *Exploring browser design trade-offs using a dynamical model of optimal information foraging.* Paper presented at the ACM Conference on Human Factors in Computing Systems, CHI '98, Los Angeles.

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

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.

Pirolli, P., Fu, W., Reeder, R., & Card, S. K. (2002). A user-tracing architecture for modeling interaction with the World Wide Web. In M. D. Marsico, S. Levialdi, & L. Tarantino (Eds.), *Proceedings of the Conference on Advanced Visual Interfaces, AVI 2002* (pp. 75–83). Trento, Italy: ACM Press.

Pirolli, P., Schank, P., Hearst, M., & Diehl, C. (1996). Scatter/Gather browsing communicates the topic structure of a very large text collection. In *Proceedings of the Conference on Human Factors in Computing Systems, CHI '96* (pp. 213–220). Vancouver, BC: ACM Press.

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

Reitman, W. R. (1964). Heuristic decision procedures, open constraints, and the structure of ill-defined problems. In M. W. Shelly & G. L. Bryan (Eds.), *Human judgements and optimality* (pp. 282–315). New York: Wiley.

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

Seshardri, V. (1993). *The inverse Gaussian distribution.* Oxford, England: Clarendon.

Simon, H. A. (1962). *The architecture of complexity.* Paper presented at the Proceedings of the American Philosophical Society, volume 106 (pp. 467–482).

Simon, H. A. (1973). The structure of ill-structured problems. *Artificial Intelligence, 4,* 181–204.

Stephens, D. W., & Krebs, J. R. (1986). *Foraging theory.* Princeton, NJ: Princeton University Press.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy, 69,* 213–225.

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

Tinbergen, N. (1963). On the aims and methods of ethology. *Zeitschrift für Tierpsychologie,* 20, 410–463.

Turney, P. D. (2001, September). *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.

Winterhalder, B., & Smith, E. A. (1992). Evolutionary ecology and the social sciences. In E. A. Smith & B. Winterhalder (Eds.), *Evolutionary ecology and human behavior* (pp. 3–23). New York: de Gruyter.

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.

Young, R. M. (1998). Rational analysis of exploratory choice. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition.* Oxford, England: Oxford University Press.