18

From Plato to the World Wide Web

Information Foraging on the Internet

Wai-Tat Fu

Abstract

Generally speaking, two conditions make cognitive search possible: (a) symbolic structures must be present in the environment and (b) these structures must be detectable by a searcher, whose behavior changes based on the structures detected. In this chapter, information search on the Internet is used to illustrate how a theoretical framework of these two conditions can assist our understanding of cognitive search. Discussion begins with information foraging theory (IFT), which predicts how general symbolic structures may exist in an information environment and how the searcher may use these structures to search for information. A computational model called SNIF-ACT (developed based on IFT) is then presented and provides a good match to online information search for specific target information. Because a further component important to cognitive search is the ability to detect and learn useful structures in the environment, discussion follows on how IFT can be extended to explain search behavior that involves incremental learning of the search environment. Illustration is provided on how different forms of semantic structures may exist in the World Wide Web, and how human searchers can learn from these structures to improve their search. Finally, the SNIF-ACT model is extended to characterize directed and exploratory information foraging behavior in information environments.

Introduction

The ability to search for useful resources has long been taken as a prime indicator of intelligence. The first decade or so of research on artificial intelligence (AI) focused almost exclusively on the study of search processes (Nilsson 1971), and AI was almost synonymous for search. The problem of search, however, has a much longer history than AI research. In the *Meno*, for example,

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

Plato (in his account of Socrates) posed the problem of search (inquiry) as a basic conundrum for understanding human intelligence (Plato 380 BCE):

> *Meno*: And how will you inquire, Socrates, into that which you know not? What will you put forth as the subject of inquiry? And if you find what you want, how will you ever know that that is what you did not know?

In the *Meno*, Plato reasoned that whenever we are in a state where there is a lack of knowledge (or information), we are compelled to search (inquire) for the information, even though we may not know exactly what it is that we seek; thus we face the problem of evaluating whether the information we find is that which we lack. The problem posed by Plato is fundamental: *How do we know how to search for something that we do not know?* Plato's solution is that we possess preexisting knowledge that allows this search. Although the origin of this preexisting knowledge is debatable (Plato argued in favor of a previous existence), for our purpose, let it suffice to assume that an agent has the ability to detect symbolic structures and to behave differentially to the detected structures. Indeed, perhaps one major characteristic of cognitive search is the *intelligence* exhibited by this particular ability to detect structures in the environment, in the general sense that an intelligent search is one that allows the searcher to decide *how* to find *relevant* information. This may sound trivial for humans (or at least many believe that humans always know what they are searching for), yet answering the question "how can an agent search intelligently" proves to be very challenging. In fact, in areas such as AI and cognitive science, *search is considered to be a central process that makes intelligence possible*, and thus has been a primary emphasis of AI research for decades (see Newell and Simon 1976).

The focus of this chapter is on a specific kind of search: information search. My goal is to demonstrate how the nature of cognitive search is manifested through the systematic investigation of the complex activities involved when people search for information. Although the Internet was not invented when Plato contemplated the question of search, I will demonstrate how the major questions that he raised are still relevant and can be used to guide understanding and theoretical development of information search on the Internet. In particular, I emphasize two important aspects of information search: (a) the structure of the Internet's information environment and (b) the means by which humans detect this structure and use it to guide their search for information.

Information Search

Searching for information has increasingly become an indispensable part of our daily activities: from checking the weather to planning a trip; from finding recipes to conducting a literature review for scientific research. While these activities may seem mundane, many scientific questions lurk behind them and

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

are relevant (or even equivalent) to those in research on different kinds of cognitive activities, such as problem solving (finding a solution to an algebra problem), decision making (finding an apartment), or foraging (finding high- or low-calorie food).

One important aspect of search is that the searcher often needs to extract information incrementally from the task environment to find the desired resource. Consider an extreme case in which resources are randomly distributed in the task environment and the searcher is unable to detect any structures or patterns in the environment. In this situation, the searcher can be said to have no more intelligence than a random (without knowledge) searcher. In most cases, however, the environment has some structure or patterns that are detectible by the searcher, so that the searcher can use this structure to acquire knowledge about the environment and improve search. Knowledge about the environment allows the searcher to behave differentially based on the detected structure and to search selectively on one path instead of others so that a resource can be reached in as few steps as possible.

An important question in the study of cognitive search is how *intelligence* is manifested through the search process. Following Newell and Simon (1976), I argue that intelligence in information search can only be found when two conditions are met:

1. Detectable structures must be present in the information environment.
2. The searcher must be able to detect these structures and use them to control search activities.

The first condition is usually determined by the nature of the environment or the task, whereas the second involves the characteristics of the searcher. It is important to note that the amount of search is not a direct measure of the amount of intelligence being exhibited. To the contrary, I argue that the study of cognitive search is concerned with the question about how an agent is able to behave differentially as it acquires information from the environment, such that a large amount of search *would* be required if the agent *did not behave* in such a way. Before I elaborate on this point, by demonstrating a specific model of information foraging that captures the efficient nonrandom search behavior of people on the Internet, I will first highlight the evidence for structure in the Internet that facilitates search. Then I will demonstrate how an intelligent searcher can navigate these environments both to find specific information and to learn about the underlying information structure in an exploratory fashion.

Structure of the Internet's Information Environment

In a world in which the Internet has become pervasive, information search is almost synonymous with search on the World Wide Web, which is dominated by the use of search engines. Those who have experience with Internet

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

search engines find that they work reasonably well in most cases. The major reason why search engines perform well (and appear intelligent) can, indeed, be attributed to their ability to exploit the inherent structure of the Web's environment. Take the algorithm *PageRank*, used by the popular search engine *Google*, as an example. PageRank works by analyzing the link structure of Web pages to assign each page a *PageRank score* (Brin and Page 1998). The exact method to derive the PageRank score (and how it is combined with other methods) is beyond the scope of this chapter; however, the general idea is to derive the score of a page based on its links to other pages, such that each link represents a "vote" to the page. Pages that have higher PageRank scores have more weight in their votes. Thus, a page with a high PageRank score is linked to many pages that also have high PageRank scores. PageRank scores can then be used to rank the list of pages returned from a search engine; the assumption is that the page with the highest score will most likely lead the searcher to find the target information.

The primary reason why link structures in the WWW can be exploited to facilitate search is that Web pages tend to occur in "patches"; that is, there tends to be a few "hubs" that connect to many pages, and many pages that are only loosely connected to other pages. The formation of these hubs is often the result of two related processes:

1. There is a general tendency for people to link new pages to "authoritative" or "high-quality" pages; once a page becomes a hub, it attracts even more pages to link to it, creating a rich-gets-richer effect.
2. When a new page is created and linked to other pages, the linked pages tend to have related topics to the new page.

Because of these tendencies, the information environment becomes "patchy": pages that contain related topics tend to be within short distances of one another and a common "hub" (measured by the number of clicks required to move between two pages). This characteristic is commonly found in the growth of a *scale-free network* (Barabási 2009). A scale-free network has the characteristic that the size of the network (i.e., number of Web pages) may increase without significantly increasing the average distance between any two nodes (i.e., the number of clicks between two random Web pages). A common example of a scale-free network is the air transportation system: The existence of "hub" airports allows new airports to be introduced without significantly increasing the average number of transfers between any two random airports. As long as the new airports are connected to nearby hubs, one can reach any airport through the hubs.

In general, exploiting link structures (and identifying hub pages) in the WWW allows the searcher to navigate more quickly to the target information. This method works well as long as the information for which they are searching is connected to one another through hub pages. Search engines that exploit link structures work well *on average* for two interrelated reasons: (a) "rich

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

patches" (hub pages) exist in the information environment and (b) most people are interested in information in one of the rich patches. However, when searching for information outside of a rich patch (e.g., when searching for information related to topics that are not yet sufficiently well connected to form hub pages), the likelihood that search engines can find the information decreases dramatically (e.g., searching for "cognitive search" on the Web will be unlikely to lead to information as rich as the chapters in this volume).

In summary, search engines generally support the exploitation of popular, well-connected information patches (reflected by the link structures constructed by Web page designers), but they do not generally support exploration for new information that is not already connected to rich information patches.

Targeted Search in the Internet Information Environment

Information foraging theory is an important theory of human information search (Fu and Pirolli 2007; Pirolli 2007; Pirolli and Card 1999). It predicts how humans interpret information cues and use these to decide how to navigate in an information environment to find specific information targets (e.g., evaluating and selecting hyperlinks when navigating in the WWW; Fu and Pirolli 2007). IFT assumes that people adapt their information-seeking behavior to maximize their rate of gaining useful information to meet their task goals (similar to optimal foraging theories developed in behavioral ecology; Stephens and Krebs 1986), and selectively to proceed along information paths based on their utility (McFadden 1974) by following cues encountered in the information environment. The theory further assumes that adaptive information systems evolve toward states that maximize gains of valuable information per unit cost. Thus IFT provides a principled method to understand how humans detect and adapt to the information structures in an environment and differentially follow an information path based on the interpretation of the detected structures.

The crucial element in IFT is the measure of *information scent*, which is defined based on a Bayesian estimate of the relevance of a distal source of information (whether the target information can be found) conditional on the proximal cues (e.g., a text snippet, such as the title of a page or the link text on a search page). Search by following an information scent is adaptive because search strategies are sensitive to the inherent predictive structures in the environment. In other words, similar to foraging behavior by animals, the decisions on *where* to search for information (food) and *when* to stop searching (patch-leaving policy) are assumed to be adapted to the statistical structures of the information environment. The detection and utilization of the structures are characterized as an adaptive response to the demands and constraints imposed by the information environments.

To a certain extent, this adaptive principle integrates the two conditions for exhibiting intelligence in the above-mentioned search. First, IFT assumes that information structures exist and that these structures emanate from two sources. The first source is the semantic structures embedded in text, and these are inherent language structures from which we derive meaning (Kintsch 1998). When a searcher sees a text snippet of a hyperlink (e.g., the link text or short description of a link that is returned from search engines), he/she can infer the relevance of the information on the page by estimating the semantic similarities between the text snippet and the information goal. The second source is the link structures between Web pages. As discussed earlier, patches which contain similar information contents (in terms of topical or semantic relevance assumption) tend to be closer to each other (in terms of number of links between them). The second condition of IFT is that people detect these structures by interpreting information cues (e.g., text snippets) and, by inferring their relevance to their information goal through a process that is inherent in human semantic memory (Anderson et al. 2004), they reach a decision on an information path through a stochastic choice process (McFadden 1974). Therefore, when a searcher selects a link that has high semantic overlap between the link text and the information goal, the searcher is getting closer to the information patch that contains the target information (i.e., assuming that a hill-climbing strategy works well).

To illustrate how IFT captures the essence of intelligence exhibited by information search, let us consider one instance of IFT: a computational model called *SNIF-ACT* (Fu and Pirolli 2007), which was developed as an extension of ACT-R (Anderson et al. 2004). The model was fit to detailed moment-by-moment Web surfing behavior of individuals studying in a controlled laboratory setting. The basic structure of the SNIF-ACT model is identical to that of a cognitive model called the *Bayesian satisficing model* (BSM) (Fu 2007; Fu and Gray 2006), which was developed to explain individual learning and choice behavior in repeated sequential decision situations. BSM is composed of a Bayesian learning mechanism and a local decision rule. SNIF-ACT applies the BSM to Web information search and assumes that, when users evaluate links on a Web page, they will incrementally update their perceived relevance of the Web page according to a Bayesian learning process. A local decision rule then decides when to stop evaluating links. Evaluation of additional links continues until the perceived relevance of the new links is lower than the cost of evaluating them. At that point, the best link encountered thus far will be selected.

To illustrate the behavior of the model, consider a case where the model is facing a single Web page with multiple links. Three actions are possible, each represented by a separate production rule (hereafter referred to as a "production"; see Anderson et al. 2004): Attend-to-Link, Click-Link, and Backup-a-Page. Similar to BSM, these productions compete against each other according to the random utility theory (McFadden 1974). That is, at any point in time, the model will attend to the next link on the page, click on a link on a page, or decide to leave the current page and return to the previous page. The utilities

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

of the three productions are derived from the link likelihood equation and can be calculated as:

$$\text{Attend-to-Link: } U(n+1) = \frac{U(n) + IS(Link)}{1 + N(n)} \tag{18.1}$$

$$\text{Click-Link: } U(n+1) = \frac{U(n) + IS(BestLink)}{1 + k + N(n)} \tag{18.2}$$

$$\begin{aligned}\text{Backup-a-page: } U(n+1) = {} & MIS(\text{Previous Pages}) \\ & - MIS(\text{Links 1 to } n) - GoBackCost\end{aligned} \tag{18.3}$$

U(*n*) represents the utility of the production at cycle *n*. *IS*(*Link*) represents the information scent of the currently attended link, calculated by the method called *pointwise mutual information* (Manning and Schutze 1999), which calculates the semantic similarity of two sets of words by some function of their base frequencies and collocation frequencies in large corpus of text. *N*(*n*) represents the number of links already attended on the Web page after cycle *n* (one link is attended per cycle). *IS*(*BestLink*) is the link with the highest information scent on the Web page; *k* is a scaling parameter; *MIS*(*page*) is the mean information scent of the links on the Web page; and *GoBackCost* is the cost of going back to the previous page. The values of *k* and *GoBackCost* are estimated to fit the data.

Figure 18.1 illustrates how the probabilities of selecting the three productions change (Figure 18.1b) as the model sequentially processes links on a single page (Figure 18.1a). Initially the probability of choosing Attend-to-Link is high. This is based on the assumption that when a Web page is first processed, there is a bias in learning the utility of links on the page before a decision is made. However, as more links are evaluated, the utilities of the productions decrease (i.e., the denominator gets larger as *N*(*n*) increases). Because the utility of Attend-to-Link decreases faster than that of Click-Link—since *IS*(*Best*) = 10, but *IS*(*link*) decreases from 10 to 2—the probability of choosing Attend-to-Link decreases but that of Click-Link increases. The implicit assumption of the model is that since evaluation of links takes time, the more links that are evaluated, the more likely it is that the best link evaluated so far will be selected; otherwise, time cost may outweigh the benefits of finding a better link.

As shown in Figure 18.1, after four links on the hypothetical Web page have been evaluated, the probability of choosing Click-Link is larger than that of Attend-to-Link. At this point, if Click-Link is selected, the model will choose the best (in this case the first) link and move on to process the links on the next page. Since the selection process uses a stochastic choice rule (i.e., a softmax rule; see Fu and Pirolli 2007), Attend-to-Link may, however, still be selected. If this is the case, as more links are evaluated—that is, as *N*(*n*) increases—the probability of choosing Attend-to-Link and Click-Link decreases. If not, the

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

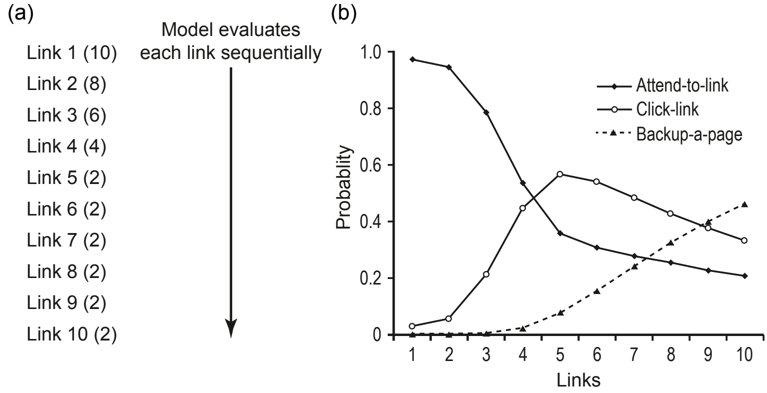

Figure 18.1 (a) A hypothetical Web page in which the information scent values (number in parenthesis) of links on the page decreases linearly from 10 to 2. (b) The probabilities of choosing each of the competing productions change as the model processes each additional link on the page; the mean information scent of the previous page was assumed to be 10.

probability of choosing Backup-a-Page is initially low because of the high GoBackCost. The utility for Backup-a-Page is calculated based on a moving average of the information scent encountered in previous pages. However, since the mean information scent of the links evaluated on the present page, *MIS*(links 1 to *n*), decreases relative to the information scent of links evaluated on previous pages, *MIS*(Previous Pages), the probability of choosing Backup-a-Page increases. This happens because the mean information scent of the current page is "perceived" to be dropping relative to the mean information scent of the previous page. In fact, after eight links are evaluated, the probability of choosing Backup-a-Page becomes higher than Attend-to-Link and Click-Link, and the probability of choosing Backup-a-Page keeps on increasing as more links are evaluated (i.e., as the mean information scent of the current page decreases). This demonstrates how competition between the productions can serve as a local decision rule that decides when to stop exploration.

Figure 18.2 shows the results of matching the SNIF-ACT model to the link selection data from a group of 74 users who conducted a search using the Yahoo! Web site (Fu and Pirolli 2007) across a range of information search tasks (e.g., "find the 2002 holiday schedule"). During the experiments, all pages visited were saved and all Web links on the pages selected by both the model and human subjects were extracted; the total frequencies of visits for each of these links are plotted in Figure 18.2. We see that the model provided good fits to the data ($R^2 = 0.91$), suggesting that the dynamic selection mechanism in the Bayesian satisficing model describes the human link selection process well.

In summary, the SNIF-ACT model demonstrates how IFT can be applied to search for specific information on the Internet. It thus creates a link with a wide range of other search domains found in this volume. Furthermore, the model

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

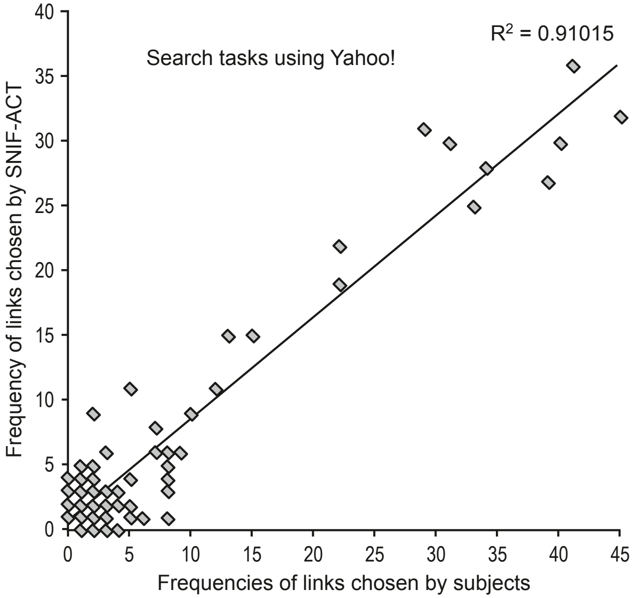

Figure 18.2 Scatterplot for the frequencies of links chosen by the SNIF-ACT model and human subjects when searching using the Yahoo! Web interface.

is very useful in predicting what links will be selected by users when they are engaged in different information search tasks. For example, the SNIF-ACT model can provide direct quantitative predictions on how likely users will find information produced by different designs of Web pages (e.g., what link text should be used, their layouts, etc.).

Exploratory Search in the Internet Information Environment

As discussed earlier, Web search engines are designed to exploit link structures to facilitate search of popular information. In many cases, people do indeed use the Web to retrieve simple facts, such as to search for the address of a restaurant, information about a person, or deadlines for filing taxes. When a person is engaged in a more open-ended or "ill-defined" search task, search engines may help, but they do not allow people to explore related information that is less popular and which may be found on pages far away from hub pages. In other words, while the "patchy" nature of Web pages allows search engines to find relevant information quickly in a patch (pages closely linked to certain hub pages), search engines are not designed specifically to facilitate search for information domains that are not patchy (i.e., information that does not overlap much in terms of topical relevance and thus may not share sufficient links to form a patch).

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

Going back to the questions posed in the *Meno*, while we may not know exactly what will be found, our ability to detect whether something is relevant often improves during the search process as we gather more information about the task and the environment. For example, when pursuing interest in the topic of "anti-aging," a person might find that it is related to many disjoint sets of topics: cosmetics, genetic engineering, nutrition, etc. When co-occurrences of these topics are repeatedly encountered during search, the searcher may be able to learn that these topics are related and relevant to the broader information goal. Typically, this does not reflect the design and intent of search engines; the ability to learn the association of topics may depend on whether the searcher can detect the structures (the co-occurrences of these topics returned from search engines) during the search. As more people use the Web to perform this kind of *exploratory information foraging*, new tools are being developed to augment search engines. However, before we discuss these new tools, I will briefly discuss the characteristics of exploratory information foraging.

In general, exploratory information foraging refers to the situation in which the information searcher has only a rough idea of the object of the search, and judgment of relevance depends greatly on the external information encountered during the search process. Furthermore, there is often no clear criterion for deciding when to stop searching. This is in sharp contrast to specific or targeted information foraging (Fu and Pirolli 2007; Pirolli and Card 1999), where the forager has a clear criterion for judging whether the target information is found. This criterion is mostly driven internally and is seldom changed or is dependent on the external information encountered during the search process. The challenge is: How do we extend the IFT so that it has the capability of learning incrementally to detect structures in the environment during the search process? Incremental changes to internal knowledge structures during search is perhaps one way that humans overcome the challenge posed in the *Meno*: how to search for "that which you know not."

Intuitively, the capability to learn from the environment seems a natural way of raising intelligence in cognitive search. For example, empirical research shows that one important difference between novice and expert chess players is that expert chess players have more stored information, which allows the expert players to recognize a large number of specific features and patterns of features on a chessboard, and information that uses this recognition to select actions based on the features recognized (e.g., Chase and Simon 1973a). While sophisticated chess programs often require search over tens of thousands of moves, chess masters seldom need to search more than a hundred of these potential moves. What makes this possible is apparently their ability to recognize patterns in the environment that are more *meaningful* than others. In other words, experiences accumulated from previous searches allow a person to derive a rich set of *semantic information* about the task environment that makes search more intelligent. Research on this in AI and cognitive science,

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

which allows artificial systems to develop semantic information to exhibit human-level intelligence, is still extremely limited.

Semantic Structures in the Information Environment

As discussed above, when retrieving popular information, search engines that exploit link structures tend to be effective in guiding users to the information. However, when searching for information not linked to "hub pages" (i.e., pages in which most people are interested), link structures do not help. For example, a recent study (Kang et al. 2010) found that when searching for information that was less popular, following the links between Web pages often led to a restricted set of information, presumably because less popular information tended to be distributed across patches in the information environment that were not directly connected by hyperlinks. In contrast, people who had more knowledge about the domain (e.g., experts) were able to utilize their knowledge to explore more efficiently for relevant information by coming up with better search terms; this allowed for better identification of relevant information patches, which subsequently allowed them to select better links and pages as they navigated in the information environment.

From observations, it appears that people can acquire knowledge during search. For example, during a Web search, a searcher may recognize certain Web sites to be more authoritative, or remember how to navigate from one page to another through a sequence of hyperlinks. Whereas traditional search engines fail to capitalize on this form of search knowledge, *social information Web sites* have been developed to allow searchers to share their knowledge with other users to facilitate exploration of information on the Web. In a recent study (Fu and Dong 2012), we examined how people learn from the social information Web site Del.icio.us—a social tagging Web site that allows people to assign short text descriptions to Web pages and share them on the site. The popularity of social tagging arises from its benefits for supporting information search and exploration. When these user-generated sets of tagged documents are aggregated, a bottom-up semantic structure, often referred to as *folksonomy*, is formed. Many argue that folksonomies provide platforms for users in a distributed information space to share knowledge among users, as social tags can reveal relationships among structures in the resources that others can follow. We have shown that as users interpret tags created collaboratively by others, these tags not only help a user explore for more relevant information, they also help the learning of the conceptual structures of other tags (Fu and Dong 2012; further details discussed in the next section).

Here I analyze this new form of social information ecology and show how emergent structures in such an ecology may guide the use of these semantic structures during exploratory information foraging (Fu and Dong 2010). To highlight the characteristics of these semantic structures, this form will be compared to link structures extracted using a method similar to the PageRank

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

algorithm (see earlier discussion). We conducted a simulation study on an existing system called Bibsonomy (Bibsonomy.org)—a public social tagging Web site that allows multiple users to post articles, Web pages, and other media to share with other users. The goal of this study was to show how different structures may help people perform exploratory information foraging.

We compared the empirical probability distribution functions of the predictive probabilities of topics and tags in each set of resources (see Figure 18.3). We defined experts broadly as people who have more domain-specific concepts which allow them to differentiate between relevant information or to make better inferences in their topic of expertise. Experts are thus generally more proficient in selecting tags that better describe topics in a resource. Quality of resources was defined as those that are most referred to by others, such as Web pages that have many links pointing to them (i.e., in-links). The definition of quality is therefore similar to that used by the PageRank algorithm, which assumes that each in-link is a "vote" by another page; the more votes a page receives, the higher its quality. By comparing how well expertise and quality can distinguish between resources, the goal is to test the extent to which experts

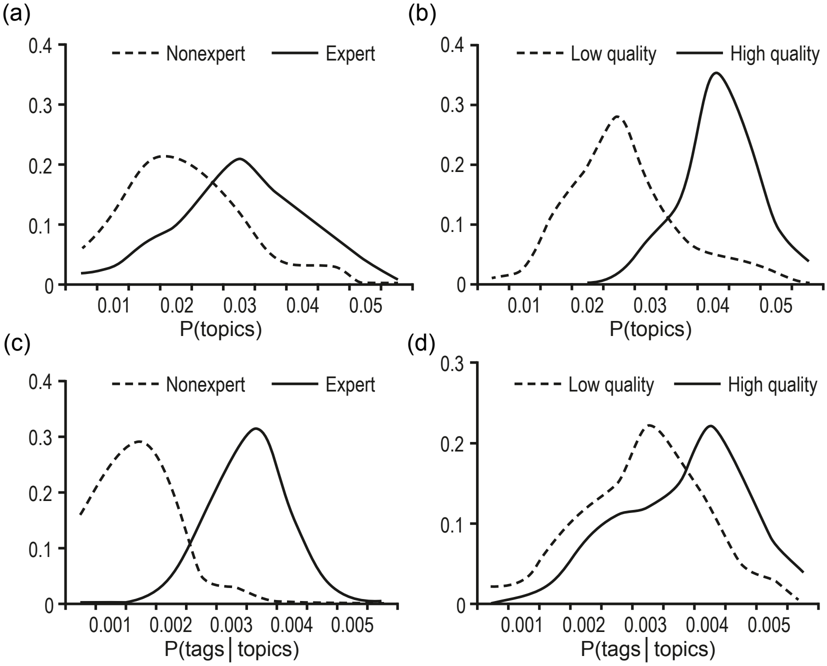

Figure 18.3 The empirical probability distribution function for the predictive probabilities of topics and tags in each of the four sets of resources. P(topics) represent the probability that a given topic to be found will be contained in the documents; P(tag|topics) indicate the probability that a tag is predictive of the topic.

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

(who index resources based on contents) and quality (which rank resources based on link structures) can help people explore for information.

The topic distributions between experts and nonexperts (Figure 18.3a) were less distinguishable than those between low- and high-quality resources (Figure 18.3b); however, the reverse was true for tag distributions (Figure 18.3c, d). This suggests that the quality of resources is generally better at predicting "hot" topics—that is, higher P(topics)— and that expert-generated tags tended to be more predictive of "cold" topics than resource quality. For example, resources tagged by a focused group of domain experts could contain cold topics associated with high-quality tags, but these resources were less likely picked up by quality (i.e., ranking of resources based on link structures). These results are consistent with the notion that content-based semantic structures are more useful for exploration of less popular topics, whereas link structures are more useful for finding popular information. For example, when presenting information cues (e.g., a recommended list of Web links as a person is exploring for information on the Web), a system can utilize either semantic structures (based on topical relevance of contents) or link structures (based on number of in-links and out-links) to select pages which may be relevant. In general, results show that following cues derived from semantic structures can more likely lead a person to explore patches that are less explored by others than those derived from link structures.

Detecting and Learning Semantic Structures during Search

Having demonstrated the different characteristics between semantic and link structures in social information Web sites for exploratory information foraging, let us turn to an experiment that directly tests how people utilize and learn from the semantic structures. In this experiment, we developed a set of exploratory information foraging tasks and observed the search behavior of people over a period of eight weeks (Fu and Dong 2012). In all tasks, participants started with a rough description of the topic and gradually acquired knowledge about the topic through iterative search-and-comprehend cycles. Participants were told to imagine that they had to write a paper and give a talk on the given topic to a diverse audience and that all kinds of questions related to the topic might be posed. Two general topics were chosen: (a) find out relevant facts about the independence of Kosovo (IK task) and (b) find out relevant facts about anti-aging (AA task). These two tasks were chosen because the IK task referred to a specific event; thus information related to it tended to be more specific, and there many Web sites contained multiple pieces of well-organized information relevant to the topic. By contrast, the AA task was more ambiguous and related to many disjoint areas, such as cosmetics, nutrition, or genetic engineering. Thus, Web sites relevant to the IK task contained more overlapping concepts (which can be found on the same pages) than those relevant to the AA task (which must be found on different pages).

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

Each of the eight participants performed one of the tasks for eight sessions over a period of eight weeks, with each session approximately one week apart. Participants were told to think aloud during the task in each session. Participants were instructed to provide a verbal summary of every Web page they read before they created any tags for the page. They could then bookmark the Web page and create tags for the page. After they finished reading a document, they could either search for new documents by initiating a new query or select an existing tag to browse documents tagged by others. This exploratory search-and-tag cycle continued until a session ended. All tags used and created during each session were extracted to keep track of changes in the shared external representations, and all verbal description on the Web pages were extracted to keep track of changes in the internal representations during the exploratory search process. These tags and verbal descriptions were then input as contents of the document. At the end of each session, participants were given printouts of all Web pages that they had read and bookmarked, and were asked to "put together the Web pages that go together on the basis of their information content into as many different groups" as the participants saw fit. These categories were then used to judge how much they had learned during the search.

To keep track of how people learn new knowledge during search, we extended the SNIF-ACT model to predict how the searcher incrementally learns to incorporate semantic structures extracted from the information system to improve their search. The idea is to assume that each searcher has a set of mental concepts *R* and a set of semantic nodes *S*. The information goal is to predict whether node S_j (some useful information) can be found by following a link with tags *T*. That is, the user is trying to estimate this probability, $P(S_j|R,T_k)$, when deciding which links can be broken down into two components:

$$P\left(S_j\,|R,T\right)=\sum_m \overset{\substack{\text{Predict internal rep}\\ \text{from external rep}\\ \downarrow}}{P\left(R_m\,|T\right)}\,P\left(S_j\,|R_m\right)\swarrow^{\substack{\text{Predict information}\\ \text{from a given mental}\\ \text{(internal) category}}} \tag{18.4}$$

In other words, to predict whether node S_j can be found in a particular document, one must first estimate $P(R_m|T)$: the probability that the document with tags *T* belongs to a particular concept R_m. The second estimate $P(S_j|Rm)$ involves the probability that *Sj* can be found in mental concepts R_m. This estimate depends on the "richness" of the mental concepts: the richer the set of mental concepts, the better the model will be able to predict whether the information can be found in the concept R_m. As the model incrementally learns to enrich the mental categories (for details, see Fu and Dong 2012), its ability to predict which links should be selected improves.

In addition, the model learns new concepts as it encounters new tags (and contents). Based on the rational model of categorization, a new concept R_0

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

is formed when $P(R_0|T)$ is larger than $P(R_m|T)$ for all m when a new tag is encountered. $P(R_0|T)$ and $P(R_m|T)$ can be calculated based on Bayes's theorem, and prior probabilities $P(R_m)$ and $P(R_0)$ can be estimated using standard methods which reduce a multi-class to a binary categorization problem. One such method is to define the prior probabilities based on a *coupling probability*—the probability (which is independent of items seen thus far) that any two items belong to the same category in the environment. The higher the coupling probability, the higher the likelihood that two items can be found in the same category (e.g., when two related items can be found on the same Web page), and the higher the likelihood that they will be in one of the existing categories. (It can be shown that when the number of viewed items increases, there is a higher tendency to put an new item into the largest category—a property that is consistent with the effect of base rate in categorization; see Fu and Dong 2012.) On the other hand, when the coupling probability is low, the likelihood that a new category will be formed will be higher, as the prior probability that any two items are from the same category is lower.

We tested the model against the data to understand how well this integration of learning and search is able to capture exploratory information foraging behavior. We used the verbal protocol data to perform model tracing to predict how well the model predicted search behavior. Figure 18.4 shows the proportion of new tags assigned to each page that participants bookmarked and the corresponding proportions of tags that were assigned by the model to the pages that it bookmarked.

Interestingly, even though participants assigned fewer tags in the AA task, the *proportions* of new tag assignment over the total number of tag assignments were higher in the AA task than in the IK task. This is consistent with the lower rate of return of relevant information in the AA task. This lower rate could be caused by the fact that existing tags were less informative for the AA task. Indeed, concepts extracted from the documents by the participants in the AA task were more often different from existing tags in the IK task, which suggests that the existing tags did not serve as good cues to information contained in the documents. The general trends and differences between the two tasks were closely matched by the model (average $R^2 = 0.83$, min = 0.62, max = 0.98). We also matched the categories of bookmarks selected by the participants as well as by the model and found good correlations between these categories (mean $R^2 = 0.83$). The current set of results demonstrates the good match of the model in keeping track of how the incremental extraction of semantic structures helped search performance. It clearly shows how the participants incrementally developed semantic structures to improve their exploration of information. Results, therefore, demonstrate how internal conceptual structures are influenced by external information structure, and how their interaction influences the success of exploratory information foraging.

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

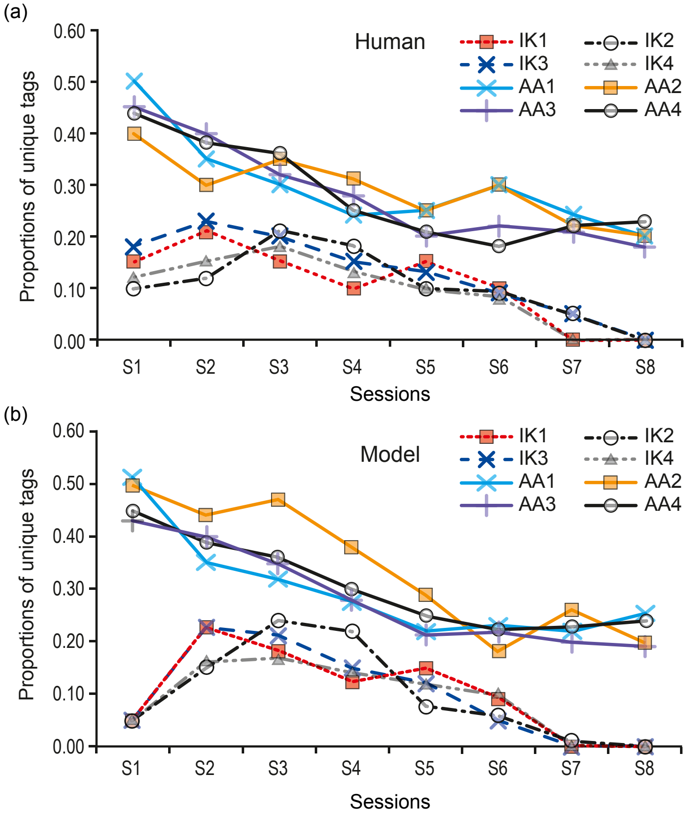

Figure 18.4 Mean proportions of unique tag assignment for the high-overlap (IK) and low-overlap (AA) tasks by participants (a) and the model (b) over eight sessions. IK1 represents participant 1 in the IK task and AA1 represents participant 1 in the AA task, etc.

General Discussion

Technology has greatly changed since Plato's *Meno*. Information search has become much more efficient through the help of the Internet and powerful Web search engines. These technological advances, however, merely *help* us carry out search much faster; they do not possess the same level of intelligence as humans or animals searching in their environments. Indeed, the vast amount of empirical research shows that cognitive search has distinct intelligence that is not yet completely understood.

In this chapter, empirical results and computational models were used to illustrate how structures exist in information environments, and how humans can detect and use these structures to guide their search. Results show that while exploiting link structures can facilitate simple fact retrieval, semantic structures are more important for people to learn and explore as their information

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

goals evolve. Analysis on the probability distribution of topics on the Web shows that, when resources are ranked based on link structures, the probability distribution of topics tends to be more distinct in "hot" topics than in "cold" topics; however, the reverse is true when resources are ranked by semantic structures derived from the contents. Empirical studies on how people perform exploratory information foraging show that people not only assimilate new information into their existing conceptual structures, they also develop new conceptual structures as they explore for information on the Web. Results further demonstrate the coupling of internal conceptual structures and external information structures during exploratory information foraging.

I have argued that for both fact retrieval and exploratory information foraging, two conditions are necessary for cognitive search: structures must exist and searchers must be able to detect and utilize them. In addition, different levels of intelligence can be observed as an organism searches in response to the structures detected in the environment. In the domains of AI and cognitive science, it is customary to believe that the critical test for understanding any behavior is to develop a machine (or program) that *exhibits the same level of intelligence as the behavior*. A major challenge is to capture the intelligence behind cognitive search to the extent that a machine can be developed to search (and learn) as humans or animals do in their natural environments. This is particularly true when search is initiated for an unknown object; for example, when a searcher is engaged in exploratory information foraging and information goals evolve as new information is found during the search process.

From "Cognitive Search: Evolution, Algorithms, and the Brain," edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.

From “Cognitive Search: Evolution, Algorithms, and the Brain,” edited by Peter M. Todd, Thomas T. Hills, and Trevor W. Robbins. 2012. Strüngmann Forum Report, vol. 9, J. Lupp, series ed. Cambridge, MA: MIT Press. ISBN 978-0-262-01809-8.