Characterizing Web Usage Regularities with Information Foraging Agents

Jiming Liu, *Senior Member, IEEE*, Shiwu Zhang, and Jie Yang

Abstract

Researchers have recently discovered several interesting, self-organized regularities from the World Wide Web, ranging from the structure and growth of the Web to the access patterns in Web surfing. What remains to be a great challenge in Web log mining is how to explain user behavior underlying observed Web usage regularities. In this paper, we will address the issue of how to characterize the strong regularities in Web surfing in terms of user navigation strategies, and present an information foraging agent-based approach to describing user behavior. By experimenting with the agent-based decision models of Web surfing, we aim to explain how some Web design factors as well as user cognitive factors may affect the overall behavioral patterns in Web usage.

Keywords

Web log; Web mining; Power law; Regularities; User behavior; Decision models; Information foraging; Autonomous agents; Agent based simulation.

I. Introduction

The contents and services on the World Wide Web (or the Web) have been growing at a very rapid rate. Up till now, there may have existed over one billion Websites on the Web at anytime, if projected based on the studies reported in [1], [2]. Viewing the Web as a large directed graph of nodes (*i.e.*, Web pages) connected with links (*i.e.*, hyperlinks), Huberman *et al.* [3] proposed a random-walk model to simulate certain regularities in user navigation behavior and suggested that the probability distribution of surfing depth (step) follows a two-parameter inverse Gaussian distribution. They conjectured that the probability of finding a group surfing at a given level scales inversely in proportion to its depth, *i.e.*, $P(L) \sim L^{-3/2}$.

In order to further characterize user navigation regularities as well as to understand the effects of user interests, motivation, and content organization on the user behavior, in this paper we will present an *information foraging agent*-based model that takes into account the interest profiles, motivation aggregation, and content selection strategies of users, and thereafter predicts the emerged regularities in user navigation behavior.

Jiming Liu (corresponding author) is with Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong; E-mail: jiming@comp.hkbu.edu.hk. Shiwu Zhang and Jie Yang are with Department of Precision Machinery and Instrumentation, University of Science and Technology of China, Hefei, China.

A. Organization of the Paper

The remainder of this paper is organized as follows: In Section II, we will provide a survey of the existing work in Web mining with a special focus on studies that deal with the regularities on the Web. This is followed by Section III that states the problems as well as important issues to be dealt with in our present study. Section IV presents the detailed formulation of our proposed information foraging agent model. Section V shows several experimental results on characterizing Web usage regularities. Section VI discusses the effects on the emergent regularities under different conditions in our model. Finally, Section VII concludes the paper by summarizing the key contributions and findings of this study.

II. RELATED WORK

This section provides an overview of research work related to Web mining. Generally speaking, Web mining is aimed to study the issues of (1) where and how information can be efficiently found on the Web and (2) how and why users behave in various situations when dynamically accessing and using the information on the Web.

A. Web Mining for Pattern Oriented Adaptation

The first major task in Web mining may be called Web mining for pattern-oriented adaptation, that is to identify the inter-relationships among different Websites, either based on the analysis of the contents in Web pages or based on the discovery of the access patterns from Web log files. By understanding such inter-relationships, we aim to develop adaptive Web search tools that help facilitate or personalize Web surfing operations.

This task is certainly justified as studies have shown that 85% of users use search engines to locate information [4]. Even though good search engines normally index only about 16% of the entire Web [2], an adaptive utility can still be useful to filter or rank thousands of Web pages that are often returned by search engines. For instance, some researchers have developed efficient search techniques that detect authorities, *i.e.*, pages that offer the best resource of the information on a certain topic and hubs, *i.e.*, pages that are collections of links to authorities [5], [6]. When it is difficult to directly find relevant information from search engines, navigating from from one page to another by following a hyperlink has become a natural way of information search. In this respect, it will be even more important to adaptively organize Web information in such a way that relevant information can be conveniently accessed.

A.1 Web Data Mining

As classified by Mobasher [7], Web data mining has traditionally been dealing with three problems: computing association rules, detecting sequential patterns, and discovering classification rules and data clusters. This classification of Web mining work has its counterparts in the field of *data mining*. Pitkow [8] summarized the previous work in Web mining with respect to different *data sources*, such as client, proxy, gateways, server, and Web. Cooley [9] presented a taxonomy of Web mining that distinguishes Web content mining from Web usage mining.

A.2 User Behavior Studies

Web usage mining deals with the analysis of Web usage patterns, such as user access statistical properties [10], [11], association rules and sequential patterns in user sessions [12], [13], [14], [15], [16], user classification and Web page clusters based on user behavior [17], [18], [19]. The results of Web usage mining can be used to understand user habits in browsing information as well as to improve the accessibility of Websites.

A.3 Adaptation

The primarily objective of Web mining for pattern-oriented adaptation is to help users to efficiently surf and retrieve information from the Web. One way to make information search efficient is to reduce the latency in information search by means of optimizing cache algorithms based on user browsing behavior characteristics on proxies or gateways [20], [21], [22], or by means of prefetching Web contents. Padmanabhan [23] proposed a method of predictive prefetching based on the analysis of user navigation patterns. However, this method is only useful if the relevant information contents at the next-level can be correctly predicted [24]. Some studies have examined the issue of Website workload [25], [26] and network traffic [27] in order to find ways to improve the efficiency of information response and propagation.

Other examples of Web mining for pattern-oriented adaptation include the studies on finding efficient search or personalization algorithms that directly work with the contents on the Web as well as the structure of the Web [28], [14].

B. *Web Mining for Model Based Explanation*

The second important task in Web mining can be referred to as Web mining for model-based explanation, that is to characterize user navigation strategies during Web surfing operations based on empirical regularities observed from Web log data. By experimenting with the decision models of Web surfing, we attempt to explain how various Web design factors as well as user cognitive factors may affect the overall behavioral patterns in Web usage.

B.1 Empirical Regularities on the Web

Recently, researchers have identified several interesting, self-organized regularities related to the Web, ranging from the growth and evolution of the Web to the usage patterns in Web surfing.

Many regularities are best represented by characteristic distributions following either a Zipf-like law [29] or a power law. That is, if probability P of a variant taking value k is proportional to $k^{-\alpha}$ where α is from 0 to 2. A distribution presents a heavy tail if its upper tail declines like a power law [30].

What follows lists some of the empirical regularities that have been found on the Web:

1. The popularity of requested and transferred pages across servers and proxy caches follows a Zipf-like distribution [20], [21], [11], [22].
2. The popularity of Websites or requests to servers, ranging from Web user groups to fixed user communities (such as within a proxy or a server) follows a power law [31], [21], [32].
3. The request inter-arrivals and Web latencies follow a heavy-tail distribution [26], [33], [19].
4. The distribution of document size either across the Web or limited to pages requested in a proxy or a certain user community exhibits a heavy tail [25], [20], [26], [11].
5. The number of pages either across all Websites or within a certain domain of the Web follows a power law [34].
6. The trace length of users within a proxy or a Website, or across the Web follows a power law [35], [3], [36], [37].
7. The dynamical response of the Web to a Dirac-like perturbation follows a power law [38].
8. The distribution of links (both incoming and outgoing) among Websites or pages follows a power law [39], [1], [40], [41], [42].

B.2 Regularity Characterization

Although researchers have empirically observed strong regularities on the Web, few of them have dealt with the issue of how such regularities emerge. Some black-box models of regularities consider only the input and output data correspondence for a "system", without explicitly addressing the rationale of underlying mechanisms. In [40], a random-network model with growth and preferential attachment factors is proposed that produces a power distribution of link number over Websites or pages. Huberman [43] showed that the power-law distribution of page number over various Websites can be characterized based on a stochastic multiplicative growth model coupled by the fact that Websites appear at different times and/or grow at different rates. He also presented a random-walk model to simulate user navigation behavior that leads to a power distribution of user navigation steps [3], [37]. Levene [36], [44] developed an absorbing Markov-chain model to simulate the power-law distribution of user navigation depth on the Web.

III. Problem Statements

The random-walk model [3], [37] and the Markov-chain model [44], [36] have been used to simulate statistical regularities as empirically observed from the Web. However, these models do not relate the emergent regularities to the dynamic interactions between users and the Web, nor do they reflect the inter-relationships between user behavior and the contents or structure of the Web. For instance, the random-walk model does not take into account the structure of the Web or the motivation of users to surf. Similarly, the absorbing Markov-chain model does not consider user interest profiles or information distribution on the Web. They are, by and large, black-box approaches that do not explicitly address the details of interacting entities.

The issues of user interest and motivation to navigate on the Web are among the most important factors that directly determine the navigation behaviors of users [45]. In our present study, we aim to take one step further by proposing a new computational model of Web surfing that takes into account the characteristics of users, such as interest profiles, motivations, and navigation strategies. By doing so, we attempt to answer the following questions:

1. Is it possible to experimentally observe regularities similar to empirical Web regularities if we formulate the aggregation of user motivation? In other words, is it possible to account for empirical regularities from the point of view of motivation aggregation?
2. Are there any navigation strategies or decision-making processes involved that determine the emergence of Web regularities, such as the distributions of user navigation depth?
3. If the above is validated, will different navigation strategies or decision-making processes lead to different emergent regularities? In other words, when we observe different power-law distributions, can we tell what are dominant underlying navigation strategies or decision-making processes that have been used by users?
4. What is the distribution of user interest profiles underlying emergent regularities?
5. Will the distribution of Web contents as well as page structure affect emergent regularities?
6. If we separately record users who can successfully find relevant information and those who fail to do so, will we observe different regularities?

In order to answer the above questions, we will develop a white-box model. This model should, first of all, incorporate the behavioral characteristics of Web users with measurable and adjustable attributes. Second, it should exhibit the empirical regularities as found in Web log data. Third, the operations in the model should correspond to those in the real-world Web surfing.

In the next section, we will present our white-box, information foraging agent-based model, for characterizing emergent Web regularities. Foraging agents are information seeking entities that are motivated to find certain information of their special interest from the pages of an artificial Web server.

IV. Information Foraging Agent Based Web Regularity Characterization

In our work, we are interested in finding the inter-relationship between the statistical observations on Web navigation regularities and the foraging behavior patterns of individual agents.

In what follows, we will introduce the notions and formulations necessary for the modeling and characterization of Web regularities with information foraging agents.

A. *Artificial Web Server*

In the agent-based Web regularity characterization, we view users as *information foraging agents* inhabiting in the Web server. The Web server is a collection of Websites connected by hyperlinks. Each Website contains certain information contents, and each hyperlink between two Websites signifies certain content similarity between them. The contents contained in a Website can be characterized using a multi-dimensional *Content Vector* where each component corresponds to the relative information weight on a certain topic. In order to build an artificial Web server that characterizes the topologies as well as connectivities of the real-world Web, we introduce the notion of an artificial Website that may cover contents related to several topics and each topic may include a certain number of Web pages. Such a Website may also be linked to other Websites of similar or different topics through URLs.

A.1 Web Server and Content Vector Representations

We consider a Web server as a graph consisting of nodes and links, as suggested in [42]. The nodes correspond to Websites and/or pages, whereas the links correspond to hyperlinks between them. The information contents in a certain node are represented using the weights of a *Content Vector* as follows:

$$\mathtt{C}_n = [cw_n^1, cw_n^2 \dots cw_n^i \dots cw_n^M] \tag{1}$$

where

$\mathtt{C}_n$: Content Vector for node n (*i.e.*, Website or page),

cw_n^i: relative content information weight on topic i,

M: number of topics.

To determine the content similarity between two nodes, we will make use of the following distance function:

$$D(\mathtt{C}_i, \mathtt{C}_j) = \left(\sum_{k=1}^{M} (cw_i^k - cw_j^k)^2 \right)^{1/2} \tag{2}$$

where $D(\mathtt{C}_i, \mathtt{C}_j)$ denotes the Euclidean distance between the Content Vectors of nodes i and j.

Thus, based on the preceding definition, we are able to specify the relationship between the contents of two nodes. For instance, when two nodes are linked through a hyperlink, it is reasonable to assume that the contents contained in the two nodes is somewhat related, that is to say, their Content Vector distance is below a certain positive threshold.

A.2 Content Distribution Models

Now that we have defined a means of representing node contents, our next question is how to describe the distribution of node contents with respect to various topics. In our present study, we will investigate the behavior of information foraging agents interacting with Web pages. The contents of those Web pages are distributed following a certain statistical law. Specifically, we will implement and contrast two models of Content Distribution: normal distribution and power-law distribution.

1. **Normal distribution**: The content weight cw_n^i with respect to topic j in node n is initialized as follows:

$$cw_n^i = \begin{cases} T + \mid X_c \mid, & \text{if } i = j, \\ \mid X_c \mid, & \text{otherwise.} \end{cases} \tag{3}$$

$$f_{X_c} \sim normal(0, \sigma_p) \tag{4}$$

$$T \sim normal(\mu_t, \sigma_t) \tag{5}$$

where

f_{X_c}: probability distribution of weight X_c,

$normal(0, \sigma_p)$: normal distribution with mean 0 and variance σ_p,

T: content (increment) offset on a topic,

μ_t: mean of normally distributed offset T,

σ_t: variance of normally distributed offset T.

In the above model, we assume that all content weights on the topic are non-negative. We can adjust σ_t and μ_t to get various topic distributions in Web pages; the smaller σ_t is or the larger μ_t is, the more focused the node will be on the topic.

2. **Power-law distribution**: In this model, the content weight of node n on topic j, cw_n^i, will follow a power law:

$$cw_n^i = \begin{cases} T + \mid X_c \mid, & \text{if } i = j, \\ \mid X_c \mid, & \text{otherwise.} \end{cases} \tag{6}$$

$$f_{X_c} \sim \alpha_p (X_c + 1)^{-(\alpha_p + 1)}, \; X_c > 0, \; \alpha_p > 0 \tag{7}$$

where

f_{X_c}: probability distribution of weight X_c,

α_p: shape parameter of a power-law distribution (also called a Pareto distribution),

T: content (increment) offset on a topic.

Similar to the model of a normal distribution, here we can adjust α_p to generate different forms of a power-law distribution.

A.3 Constructing an Artificial Web Server

Having introduced the notions of Content-Vector representation and Content Distribution models, in what follows we will discuss how to add links to an artificial Web server.

There are two major steps involved. First, we create several groups of nodes, where each group focuses on a certain topic. The distribution of the contents in the nodes follows a specific model as given above. We assume that an information agent starts its foraging from a Web homepage that contains links to the nodes of several topics. In our study, we assign this homepage equal distance to individual topics as follows:

$$cw_p^i = T_c, \ i = 1 \ldots M \tag{8}$$

where cw_p^i denotes the content weighting of the homepage on topic i. T_c denotes the content (increment) offset on the topic.

After initializing the Content Vectors, the next step is to build links between the nodes. As mentioned above, we assume that when there is a link between two nodes, the information contents of the two nodes should be related. Therefore, we will build a link between nodes only if the Content-Vector distance between them is below a positive distance threshold, r. r can be adjusted in order to generate Web clusters of different degrees of connectivity. In this respect, we refer to r as the *degree-of-coupling* (doc) of Websites. Increasing r leads to increasing the number of links in a Website (that is, the similarity between the contents of two linked nodes will decrease).

Now, let us summarize the key steps in constructing an artificial Web server as follows:

```
1.    For each topic i
2.        Create node Content Vectors
      End
3.    For each node i
4.        Initialize the link list of node i
5.        For each node j
6.            If D(C_i, C_j) < r
7.                Add node j to the link list of node i
8.                Add D(C_i, C_j) to the link list of node i
            End
        End
      End
```

A.4 Remarks on the Artificial Web Server

In the construction of our artificial Web server, we have assumed that two pages are similar if they are linked. This assumption has been found to be generally valid with respect to the real-world Web by several researchers [46], [47], [48]. For instance, in the studies reported in [46], Menczer has examined the relationship between content, linkage, and semantic similarity measures across a large number of real-world Web page pairs and has found that the Pearson's correlation coefficients between content and linkage similarity measures significantly positive. For instance, the content similarity measure can reach up to $0.4 \sim 0.6$ when the linkage similarity measure (a neighborhood function) is around 0.6. Both measures will have peaks around 0.9. Such a correlation is found to be significantly positive in the Web pages that deal with News, Home, Science, Sports, Reference, and Games among others. In [47], Menczer further formalizes and quantitatively validate two conjectures that are often taken for granted; they are:

1. the link-content conjecture that "a page is similar to the pages that link to it" and
2. the link-cluster conjecture that "pages about the same topic are clustered together."

Having said so, it should be pointed out that given the variety of kinds of links that are created in the real-world Websites, 'distance' may not always be a good indication of 'relevance' among Web pages. In some cases, two Web pages may be linked simply because one adds a special feature or service to another.

A.5 Dynamically Generated Web Pages

In the real-world Web, some portion of pages may be 'hidden' in databases; they are generated on the fly. In the artificial Web pages constructed in this study, we have not considered the dynamic generation of Web pages, but used only existing and continuing pages. Although our virtual Web pages may, to a certain extent, model the characteristics of the dynamically generated Web pages, there are still differences between them that deserve further experimental examinations taking both facets into consideration.

B. Foraging Agents

B.1 Interest Profiles

Each agent forages in the Web server with different interests in mind, *e.g.*, accessing a specific Website for an update on some contents, searching for information related to some topics, or simply wandering in the Web server to browse various topics. The Interest Profile of an agent will determine its behavior in Web surfing. In this section, we will describe how to model the Interest Profile of an agent using a multi-dimensional *Preference Vector* that specifies the interests of the agent in various topics. In addition, we will also introduce the measure of Interest Entropy to characterize whether or not an agent has a balanced Interest Profile.

Specifically, we define the Preference Vector of an agent as follows:

$$\mathtt{P}_m = [pw_m^1, pw_m^2 \ldots pw_m^i \ldots pw_m^M] \tag{9}$$

$$p_{mi} = \frac{pw_m^i}{\sum_{j=1}^{M} pw_m^j} \tag{10}$$

$$H_m = -\sum_{i=1}^{M} p_{mi} log(p_{mi}) \tag{11}$$

where

$\mathtt{P}_m$: Preference Vector of agent m,

pw_m^i: weight of preference on topic i,

H_m: Interest Entropy of user m.

In Equation 11, we define H_m in a similar way as we define the measure of entropy in information theory. Here, H_m indicates the breadth and balance of an agent's interests in different topics. The larger H_m is, the more evenly distributed the agent's interests will be. As a result, the agent is more likely to have multiple objectives and jump from one topic to another in its surfing. When the agent has equal interests in all topics, the value of H_m will be the largest, *i.e.*,

$$H_{max} = -\sum_{i=1}^{M} \frac{1}{M} log\left(\frac{1}{M}\right) = log(M) \tag{12}$$

As to be discussed in the next section, the quantity of Interest Entropy will affect the decision of an agent on which Web page to be selected among several others.

B.2 Interest Distribution Models

In order to investigate how different Interest Distributions may influence the behavior patterns of an agent's foraging, in our study we will specifically implement and observe two Interest Distribution models: normal distribution and power-law distribution. Thus, the Preference Vector of a foraging agent will be initialized as follows:

1. **Normal distribution**: The weight of a Preference Vector, pw_m^i, for agent m on topic i is defined as follows:

$$pw_m^i = X_p \tag{13}$$

$$f_{X_p} \sim normal(0, \sigma_u) \tag{14}$$

where $normal(0, \sigma_u)$ denotes the normal distribution with mean 0 and variance σ_u.

2. **Power-law distribution**: The probability distribution of agent m's preference weight on topic i, pw_m^i, is given as follows:

$$pw_m^i = X_p \tag{15}$$

$$f_{X_p} \sim \alpha_u (X_p + 1)^{-\alpha_u + 1}, \; X_p > 0, \; \alpha_u > 0 \tag{16}$$

where α_u denotes the shape parameter of a power-law distribution.

We can get various Interest Profiles of foraging agents by adjusting parameters σ_u and α_u.

B.3 Motivational Support Aggregation

When an information searching agent finds certain Websites in which the content is close to its interested topic(s), it will become more ready to *forage* to the Websites at the next level, that is, it gets more *motivated* to *surf* deeper. On the other hand, when the agent does not find any interesting information after some foraging steps or it has found enough contents satisfying its interests, it will stop foraging and leave the Web server. In order to model such a motivation-driven foraging behavior, here we introduce a support function, S_t, which serves as the driving force for an agent to forage further. When the agent has found some useful information, it will get rewarded, and thus the support value will be increased. As the support value exceeds a certain threshold, which implies that the agent has obtained a sufficient amount of useful information, the agent will stop further foraging. In other words, the agent is satisfied with what it has found. On the contrary, if the support value is too low, the agent will lose its motivation to forage further and thus leave the Web server.

Specifically, the support function is defined as follows:

$$S_{t+1} = S_t + \theta \cdot \Delta M_t + \phi \cdot \Delta R_t \tag{17}$$

where

- S_t: support value at time step t,
- ΔM_t: motivational loss at time step t,
- ΔR_t: reward received at time step t,
- θ, ϕ: weights of motivation and reward, respectively.

The initial support value, maximum and minimum support thresholds will be set, respectively, as follows:

$$\texttt{init\_support}_m = \frac{1}{2} \sum_{i=1}^{M} pw_m^i \tag{18}$$

$$\texttt{max\_support}_m = \sum_{i=1}^{M} pw_m^i \tag{19}$$

$$\texttt{min\_support}_m = 0. \tag{20}$$

where pw_m^i denotes the preference weight of agent m with respect to topic i.

C. Foraging in an Artificial Web Server

Generally speaking, the hyperlinks inside a Web page are connected to other pages covering the same or similar topics. The words or phrases that are used in the hyperlinks usually indicate the topics of the linked pages. In the process of information foraging, an agent will examine the hyperlinks and then predict which of the linked next-level pages may contain more interesting contents. In so doing, the predictability of different agents may be different, depending on their navigation strategies used.

Earlier research on closed hypertext systems, databases, and library information systems have suggested that there are possibly three browsing strategies: search browsing (directed search where the goal is known), general-purpose browsing (consulting sources that have a high likelihood of items of interest), and serendipitous browsing (purely random) [49]. In this section, we will provide the computational models of three navigation strategies to be used by information agents, and describe how the agents will update their Interest Profiles, motivation and reward functions during information foraging.

C.1 Navigation Strategies

Suppose that agent m is currently in page n that belongs to topic j (also referred to as domain here). There are h hyperlinks inside page n, among which h_1 hyperlinks belong to the same topic as page n and h_2 hyperlinks belong to other topics. We can describe the strategies of foraging agents in selecting the next-level Web page, *i.e.*, selecting hyperlink k out of h hyperlinks, in terms of different selection probabilities, as follows:

1. **Random agents**: Random agents have no strong interests in any specific topics. They wander from one page to another. In so doing, their decisions in selecting the next-level pages are random. The probability of reaching node k, p_k, at the next step can be written as follows:

$$p_k = \frac{1}{h} \qquad\qquad k = 1 \ldots h \tag{21}$$

2. **Rational agents**: Most foraging agents behave rationally. Rational agents have specific interested topics in mind and they forage in order to locate the pages that contain information on those topics. When they reach a new Website, they will try to decide whether or not the content sufficiently

matches their Interest Profiles and, if not, predict which page at the next level will be likely to become a more interesting one. In predicting the next-level contents, they will examine the titles of various hyperlinks inside the current page. Thus, the probability, p_k, of reaching the next-level node k given the Interest Entropy of agent m, H_m, can be computed as follows:

$$D^*(\mathtt{P}_m, \mathtt{C}_k) = \begin{cases} D(\mathtt{P}_m, \mathtt{C}_k), & \text{if } k \in h_1, \\ \frac{H_m}{H_{max}} D(\mathtt{P}_m, \mathtt{C}_k), & \text{if } k \in h_2. \end{cases} \tag{22}$$

$$p_k = \frac{D^*(\mathtt{P}_m, \mathtt{C}_k)^{-1}}{\sum_{j=1}^{M} D^*(\mathtt{P}_i, \mathtt{C}_j)^{-1}} \quad k = 1 \dots h \tag{23}$$

where $D^*(\mathtt{P}_m, \mathtt{C}_k)$ denotes the weighted distance between the preferences of agent m and the contents of node k given the agent's Interest Entropy H_m.

We note that Web pages can contain many outgoing links. In such a case, it will not be effective if rational agents select the next-level pages directly applying the above calculation of p_k. In order to maintain the predictability of agents, we modify the above probability definition as follows:

$$\begin{aligned} & Q_j = D^*(\mathtt{P}_m, \mathtt{C}_j) - \mathtt{mean}_{\forall l \in h}(D^*(\mathtt{P}_m, \mathtt{C}_l)), && j = 1 \dots h \\ & \text{Remove } j \text{ from set } h, && \text{if } Q_j \geq 0 \end{aligned} \tag{24}$$

$$p_k = \frac{Q_k}{\sum_{j=1}^{h} Q_j} \tag{25}$$

3. **Recurrent agents**: Recurrent agents are those who are familiar with the Web structure and know the whereabouts of interesting contents. They may have frequently visited such Websites. Each time when they decide to forage further, they know exactly the whereabouts of the pages that closely match their Interest Profiles. In this case, the probability of selecting a Web page at the next step can be defined as follows:

$$p_k = \begin{cases} 1, & \text{if } D^*(\mathtt{P}_m, \mathtt{C}_k) = \mathtt{min}(D^*(\mathtt{P}_m, \mathtt{C}_j)), \qquad j = 1 \dots h \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

C.2 Preference Updating

The preference of an agent changes over time, depending on how much information on interesting topics the agent has found and how much the agent has absorbed such information. Generally speaking, the update of the preference weights in the agent's Interest Profile reflects the change of the agent's continued interest in certain topics.

When agent m reaches and finishes reading page n, its interest will change according to page's Content Vector. The specific updating mechanism is defined as follows:

$$\begin{aligned} &\mathtt{P}_m(\tau) = \mathtt{P}_m(\tau-1) - \lambda \cdot \mathtt{C}_n, \\ &pw_m^i(\tau) = 0, \qquad for\ \ pw_m^i(\tau) < 0, \ \ i = 1 \ldots M \end{aligned} \tag{27}$$

where λ denotes an absorbing factor in [0,1] that implies how much information is accepted by agents on average. $\mathtt{P}_m(\tau)$ and $\mathtt{P}_m(\tau-1)$ denote an agent's Preference Vector after and before accessing information in page n, respectively.

C.3 Motivation and Reward Functions

As mentioned in Section IV-B.3, the motivational support for an agent plays an important role in information foraging. Depending on the support value, the agent will decide whether or not to forage further to the next-level Web pages. In what follows, we will elaborate on how the motivational support is aggregated based on the associated motivation and reward functions.

Recall that there are three terms in Equation 17. The first term S_t denotes the influence of initial and previously aggregated foraging support. The second term ΔM_t denotes the motivational (or patience) loss in information foraging. It changes along with the latency, *i.e.*, the time to find information. The third term ΔR_t denotes the reward received after finding relevant information.

There are many ways to compute ΔM_t, which can be generally characterized as follows:

$$\Delta M_t = -(\Delta M_t^c + \Delta M_t^v) \tag{28}$$

where ΔM_t^c denotes the constant decrement in ΔM_t at each time step, and ΔM_t^v the variable factor that dynamically changes at each time step. In our study, we adopt the following model of ΔM_t^v:

1. As earlier studies have shown that the empirical distribution of waiting time to access Web pages follows a log-normal distribution [4], [33], it is reasonable to believe that the distribution of motivational loss will also be a log-normal function:

$$f_{\mathtt{log}(\Delta M_t^v)} \sim normal(\mu_m, \sigma_m) \tag{29}$$

where μ_m and σ_m denote the mean and variance of the log-normal distribution of ΔM_t^v, respectively.

2. The patience or interest of an agent in carrying on information foraging decreases as the number of required foraging steps increases. Thus, we adopt the following mechanism for dynamically updating the motivation function:

$$\Delta M_t^v = \alpha_m e^{\gamma_m \mathtt{step}} \tag{30}$$

where α_m and γ_m denote the coefficient and rate of an exponential function. `step` denotes the number of pages/nodes that an agent has continuously visited.

Next, let us define the reward function in Equation 17. In our study, we model the reward received by an agent at each time step as a function proportional to the relevant information that the agent has absorbed. In our model, since the change of the agent's preference weights reflects the information that the agent has gained, we can write the reward function as follows:

$$\Delta R_t = \sum_{i=1}^{M} (pw_m^i(\tau - 1) - pw_m^i(\tau)). \tag{31}$$

Note that the reward, ΔR_t, for an agent is always greater or equal to zero. It provides the agent with the energy to forage on the Web. On the other hand, the motivational loss, ΔM_t, of the agent is always negative, which prevents the agent to forage further. Therefore, the total support for an agent at the current time step can be aggregated based on the support received at the previous time steps and the changes in the above-mentioned motivational loss and reward functions.

C.4 Remarks on Motivational Loss

In our present work, the experimental results will be obtained based on the assumption that the motivational loss in Web surfing follows a log-normal distribution. This assumption was in part inspired by the EPIC (Executive-Process/Interactive-Control) model of verbal working memory. Readers who are interested in EPIC are referred to [50] for details.

C.5 Foraging

Having defined the artificial Web server, the Interest Profile, and the support function of an agent, in what follows we will provide an outline of steps for simulating information agents foraging in the artificial Web server. We assume that the agents will start to forage from a homepage that contains links to other Web pages of various topics. When the support for an agent is either below a lower bound or above an upper bound, the agent will stop information foraging; otherwise, it will select and move to the next-level page. The specific steps are summarized as follows:

1. **Initialize** the nodes and links in an artificial Web server
2. **Initialize** information foraging agents and their Interest Profiles
3. **For** each agent m
4. **While** the support for the agent $S <$ `max_support`$_m$ and $S >$ `min_support`$_m$
5.   Find the hyperlinks inside node n that the agent is presently in
6.   Select, based on p_k, the hyperlink that connects to the next-level page
7.   Forage to the selected page
8.   Update the preference weights in the agent's Interest Profile based on Equation 27
9.   Update the support function of the agent based on Equation 17

```
            End
10.         If the support for the agent S > max_support_m
11.            Agent m is satisfied with the contents and leaves the Web server
            Else
12.            Agent m is dissatisfied and leaves the Web server
            End
        End
```

In the next section, we will present simulated foraging results and compare them with some real-world observed empirical datasets for validation.

V. Experimentation and Validation

In this section, we will describe several experiments in which the preceding given model of information foraging agents are implemented and simulated in an artificial Web server. The objective of these experiments is to validate the agent model using some empirically obtained real-world Web log datasets. Specifically, we want to examine whether or not the strong regularities that emerged from empirical Web log datasets can be generated in the simulations using the information foraging agents. If so, we can claim that the computational model proposed based on the idea of information foraging *characterizes* the behavior of human Web surfing that generates empirical Web regularities.

A. Experiment

In our experiment, we apply the steps as outlined in the preceding section to initialize and control information foraging agents. As the agents undertake their foraging sessions in the Web server, we will record their surfing-depth (step) distribution and the rank-frequency distribution of link clicks. The frequency of link clicks refers to the number of times for which a link is passed through by the agents. It is also called *link-click-frequency*.

In Experiment 1, we initialize $5,000$ agents foraging according to the above-given motivational support and decision models for three categories of foraging agents. In this experiment, we assume that the Interest Profiles of the agents follow a *power-law distribution* and the contents of Web pages on various topics follow a *normal-like distribution*. The detailed experimental parameters are given in Table I.

Figures 1 and 2 present the statistical distributions of foraging depth and link-click-frequency obtained in Experiment 1 for recurrent and rational agents, respectively.

From Figures 1 and 2, we can note that there do exist strong regularities in the behavior of agents foraging in the Web server. The *cumulative probability distribution of agent steps in accessing*

Degree-of-coupling, r	0.7
Number of agents	5,000
Number of nodes	254
Number of topics	10
T_c	0.1
ΔM_t^c	0.2
ΔM_t^v	1st
α_u	1.5
ϕ	1
λ	0.6
μ_m	5.97
μ_t	1.0
σ_m	0.8
σ_p	0.25
σ_t	0.2
θ	1

TABLE I
THE PARAMETERS FOR EXPERIMENT 1.

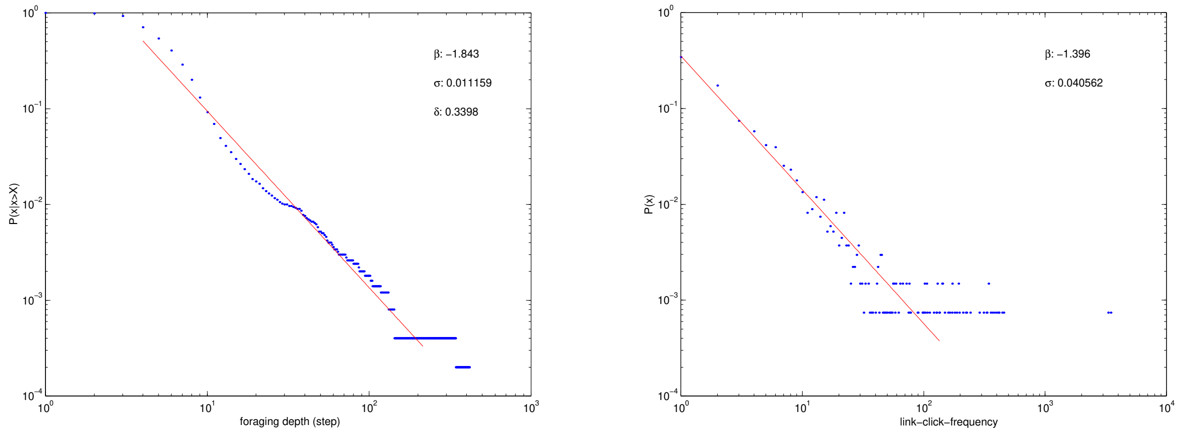
pages follows a heavy tail. Thus, the probability of agent foraging depth slowly decreases. It is interesting to observe from Figures 1(b) and 2(b) that the distributions of link-click-frequency exhibit a power law. A similar result on the distribution of Website popularity has been empirically observed and reported in [43].

In obtaining the lines of Figures 1 and 2, we apply a weighted linear-regression method, in which we assign the probability at each depth or link-click-frequency with the frequency of the depth or link-click-frequency occurrence. This implies that the higher the occurrence rate of a depth or a link-click-frequency is, the higher the weight will be.

B. Model Validation Using Real World Web Logs

In order to validate our model, we will use some real-world Web log datasets and compare their corresponding empirical distributions with those produced by the information foraging agents as mentioned above.

The first dataset is NASA Web server log that recorded all HTTP requests received by the NASA Kennedy Space Center Web server in Florida from 23:59:59 August 3, 1995 to 23:59:59 August 31, 1995. The data is available at `http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html`. Before we plot the distributions, we first filter the dataset by keeping only the requests that asked for the `html` files. This allows us to remove the noisy requests that were not directly sent by users such as the requests for image files. Here, we regard a *user session* as a sequence of a user's continuously browsed pages on the Web, which can be derived from the filtered dataset. To obtain the user sessions, we assume that the continuous requests from the same IP correspond to the same user. We also assume that a user session ends if the idle time of a user exceeds a threshold of 30



(a) (b)

Fig. 1. Recurrent agents in Experiment 1. (a) Cumulative distribution of agent foraging depth (step), where '·' corresponds to experimental data and '—' corresponds to a linear-regression fitted line. The tail of the distribution follows a power-law distribution with power $\beta_c = -1.843$ and the residual of linear regression $\sigma = 0.01$. δ denotes agents' satisfaction rate (*i.e.*, the ratio of the number of satisfied agents to the total number of agents what have surfed on the Web). (b) Distribution of link-click-frequency (link click refers to the total times for which agents pass through a link). The tail follows a power-law distribution with power $\beta_l = -1.396$, as obtained by weighted linear regression.

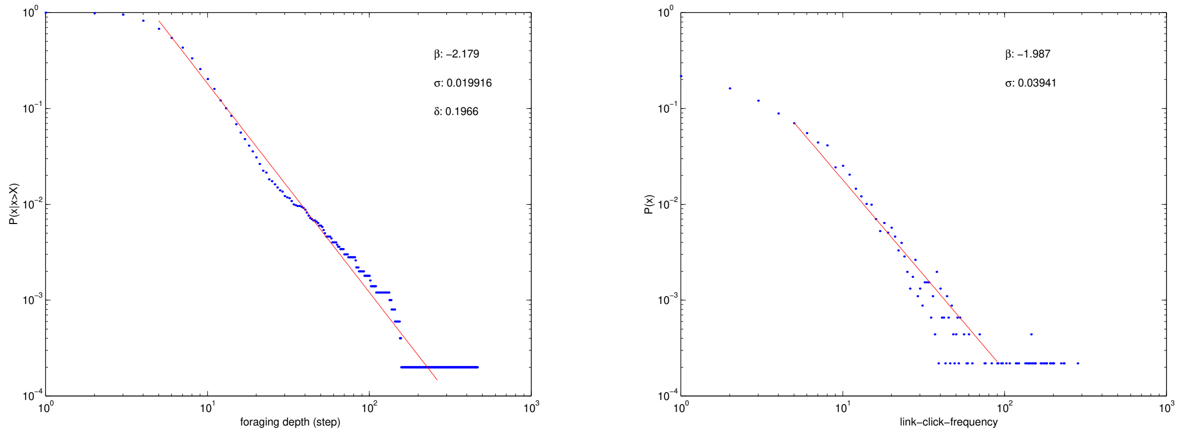
minutes.

In the filtered NASA dataset, there are 333,471 requests in 118,252 user sessions. The average depth of surfing by users is 2.82 *requests per user*. In addition, there are 1,558 nodes and 20,467 links found in the dataset that were visited by users. The average links per node is around 13.

The distributions of user surfing depth and link-click-frequency for the NASA dataset are shown in Figures 3(a) and (b), respectively.

The second dataset is from the Website of a laboratory at Georgia Institute of Technology (GIT-lab), which recorded the requests from March 26, 1997 to May 11, 1997. We preprocess the data in the same way as we did for the NASA data. As a result, we have found that there are 24,396 requests contained in the filtered dataset, and 6,538 user sessions, an average of 3.73 requests per user. Also, there are 1,147 nodes and 6,984 links visited by users. The distributions of user surfing depth and link-click-frequency for the GIT-lab dataset are shown in Figure 4.

Now, let us compare the empirical distributions of Figures 3 and 4 with the distributions of Figures 1 and 2 generated by information foraging agents in an artificial Web server. We can note that the results are similar, from the shapes of distributions to the parameters of the fitted functions. The NASA dataset reveals emergent regularities closer to those produced by rational agents as in Figure 2, whereas the GIT-lab dataset presents emergent regularities closer to those produced by recurrent agents as in Figure 1. These results demonstrate that our white-box model,



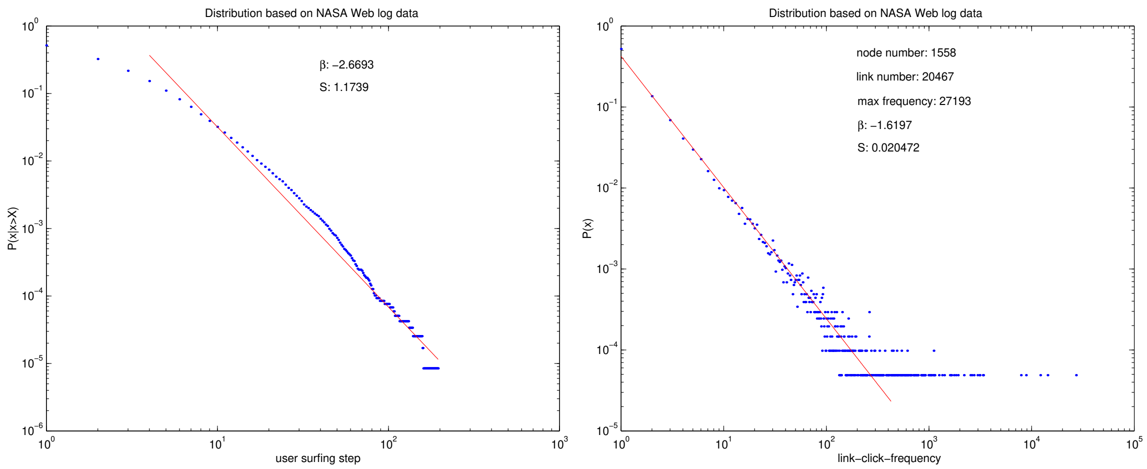
(a) (b)

Fig. 2. *Rational* agents in Experiment 1. (a) Cumulative distribution of agent foraging depth (step), where '·' corresponds to experimental data and '−' corresponds to a linear-regression fitted line. The tail of the distribution follows a power-law distribution with power $\beta_c = -2.179$ and the regression residual $\sigma = 0.02$. δ denotes agent's satisfaction rate. (b) Distribution of link-click-frequency. The distribution follows a power-law distribution with power $\beta_l = -1.987$, as obtained by weighted linear regression.

incorporating the behavioral characteristics of Web users with measurable and adjustable factors, does exhibit the regularities as found in empirical Web data. The foraging operations in the model correspond to the surfing operations in the real-world Web server.

In addition to the distributions of user steps in accessing pages and link-click-frequency, we are also interested in the *distribution of user steps in accessing domains or topics* − an issue of great importance that has never been studied before. We define agent steps in accessing domains as the number of domains that an agent has visited, and define an agent's *satisfaction rate* as the ratio of the number of satisfied agents to the total number of agents after they have completed surfing. Figure 5 presents the distributions of steps in accessing domains by recurrent and rational agents in Experiment 1, respectively. From Figure 5, we can readily observe that the cumulative probability distributions of agent steps in accessing domains follows an exponential function.

We have further obtained an empirical dataset that recorded user behavior in accessing the domains of a Website. The dataset is a Web log file for the *Microsoft* corporate Website, recording the domains or topics of `www.microsoft.com` that anonymous users visited in a one-week timeframe in February 1998. The dataset is available from `http://kdd.ics.uci.edu/databases/msweb/msweb.html`. In this dataset, there are 294 main domains and 32,711 users, with an average of 3 steps per domain. The number of user sessions is 6,336. The average number of links among domains passed through by the users is 6,336/294, or 21.55. The distribution of user steps in accessing domains is shown in Figure 6. Now, if we compare Figure 5 with Figure 6, we can note that the domain-visit



(a) (b)

Fig. 3. Distributions based on real-world *NASA* Web log data. (a) Cumulative distribution of user surfing step. The distribution follows a heavy tail with the tail's scale of $\beta_c = -2.669$. The linear-regression residual s is about 1.17. (b) Distribution of link-click-frequency. It agrees well with a power law of power $\beta_l = -1.62$, as obtained by weighted linear regression.

regularity generated by our model characterizes the empirically observed domain-visit regularity well.

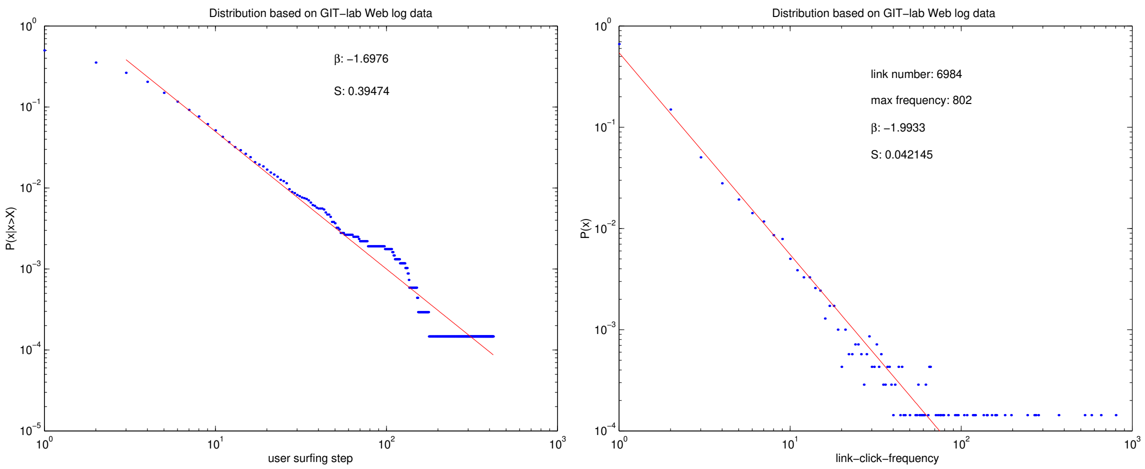
VI. Discussion

In the preceding section, we have presented a model of information foraging agents and shown how this model is derived and used to characterize empirical Web regularities. In this section, we will further investigate the inter-relationships between the emergent Web regularities as computed from our model and the characteristics of various user Interest Profiles and Content Distributions.

A. *Foraging Depth*

One of the main objectives in our research is to find out how the regularities in user navigation may be affected by the Content Distributions on the Web. In Experiment 2, we assume that the Content Distribution in the Web nodes follows a *power-law* and we keep all other parameters as in Experiment 1. We are interested in examining the influence of different Content Distribution models on agent navigation behavior. The specific parameters for this experiment are given in Table VI-A.

Now, let us compare the distributions of agent foraging depth in accessing Web pages as obtained from Experiments 1 and 2. Figure 7 shows the foraging depth distributions of recurrent and rational agents, respectively, from Experiment 2. We note that the two plots in Figure 7 are almost the same as those in Figures 1(a) and 2(b), respectively. Therefore, we suggest that the regularity



(a) (b)

Fig. 4. Distributions based on real-world *GIT-lab* Web log data. (a) Cumulative distribution of user surfing step. The distribution exhibits a heavy tail with the tail's scale of $\beta_c = -1.698$. The linear-regression residual s is about 0.395. (b) Distribution of link-click-frequency. It agrees well with a power law of power $\beta_l = -1.993$, as obtained by weighted linear regression.

of agent foraging depth in accessing Web pages may not be affected by the models of Content Distributions in the Web nodes.

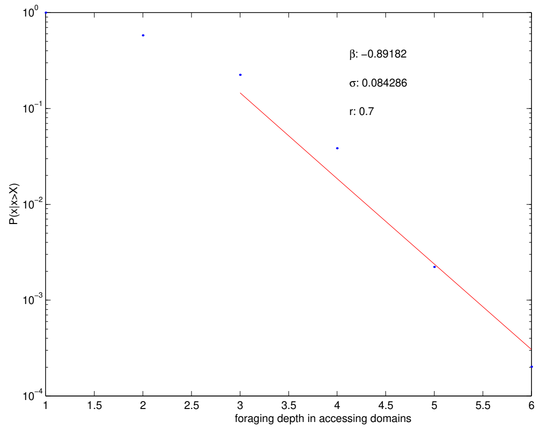
Next, we will examine the effect of agent Interest Profiles on the Web regularities. For this purpose, we will conduct Experiment 3, in which the Interest Profiles of agents are created based on a *normal-distribution* model. We will set all other parameters the same as Experiment 1. The specific parameters are given in Table III.

Figures 8(a) and (b) present the distributions of agent foraging depth in accessing Web pages by recurrent and rational agents, respectively, as obtained in Experiment 3. From Figure 8, we note that both distributions exhibit an exponential function. As the only difference between the settings of Experiments 1 and 3 is the distribution model of Interest Profiles used in the agents, we suggest that the regularities of power-law distributions observed in agent foraging depth in accessing Web pages are largely resulted from the power-law distribution of agent interests in various topics.

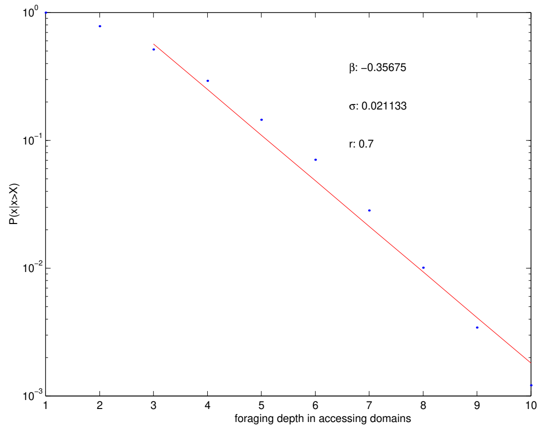
B. *Link-Click-Frequency*

Next, let us take a look at the link-click-frequency distributions in the earlier-mentioned experiments. Figures 9 and 10 present the distributions obtained in Experiments 2 and 3, respectively. As shown in the figures, the distributions of link-click-frequency remain to be a *power law* under the conditions of different agent Interest Distribution and Content Distribution models.

It should be pointed out that the above results can be established for recurrent and rational agents only. In the case of random agents, the regularities in link-click-frequency will disappear.



(a)



(b)

Fig. 5. Agents visiting *domains* in Experiment 1. (a) Cumulative distribution of foraging depth in accessing domains by *recurrent* agents, where '$\cdot$' corresponds to experimental data and '$-$' corresponds to a linear-regression fitted line. The distribution follows an exponential function with exponent $\beta_d = 0.892$ and residual $\sigma = 0.08$. (b) Cumulative distribution of foraging depth in accessing domains by *rational* agents. The distribution follows an exponential function with a *smaller* exponent $\beta_d = 0.357$ and residual $\sigma = 0.02$.

Figures 11(a) and (b) show the plots of link-click-frequency for random agents in Experiments 1 and 2, respectively.

In fact, if we compare Figure 11(a) with Figure 1(b) and Figures 2(b) and 11(b) with Figure 9, we can observe that from random agents to recurrent agents, the *power law* in link-click-frequency distribution will become more and more obvious. The only distinction among the different categories of agents in our information foraging model is their ability to predict which of linked next-level pages may contain more interesting contents. Thus, we suggest that the power-law distribution of link-click-frequency may be affected by the content predictability of the agents.

C. Degree-of-Coupling (doc)

In Section IV-A.3, we introduced a parameter for setting minimum similarity between two linked Web pages, called *degree-of-coupling* (doc), r. The larger the value of r is, the more links among Web pages belonging to different topics as well as the more links per each Web page. Given a certain r, the topology of an artificial Web server is determined. Agents with multiple interests will more readily forage from the contents on one topic to the contents on another topic. On the other hand, agents with a single interest will become more obsessive to select a direction from many hyperlinks within a page.

Figure 12 shows that the average number of links will increase as r increases. This result concerning Web structure is commonly found on the real-world Web. The question that remains is

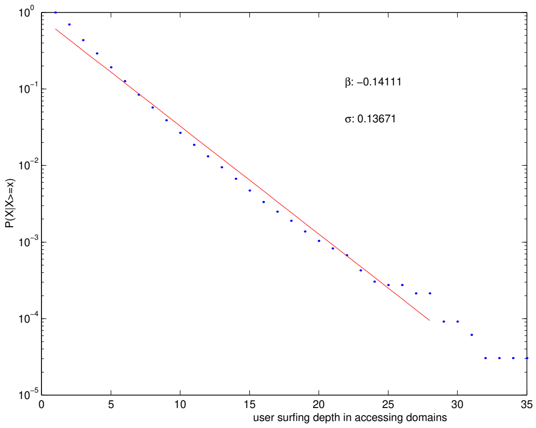


Fig. 6. Real-world *Microsoft* Web log data. Cumulative distribution of user step in accessing domains. The distribution follows an exponential function with $\beta_d = -0.141$. The regression residual σ is about 0.137.

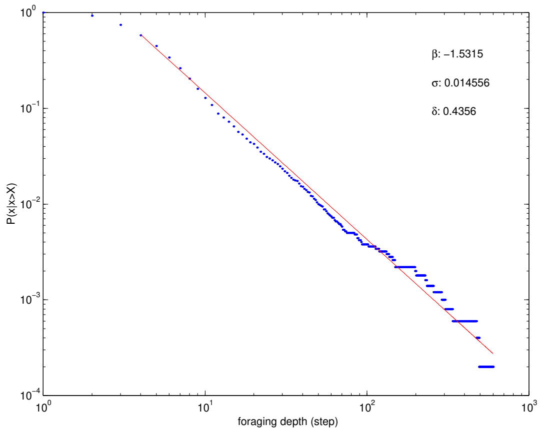
Degree-of-coupling, r	0.7
Number of agents	5,000
Number of nodes	246
Number of topics	10
T_c	0.1
ΔM_t^c	0.2
ΔM_t^v	1st
α_p	3
α_u	1.5
ϕ	1
λ	0.6
μ_m	5.97
μ_t	1.0
σ_m	0.8
σ_t	0.2
θ	1

TABLE II

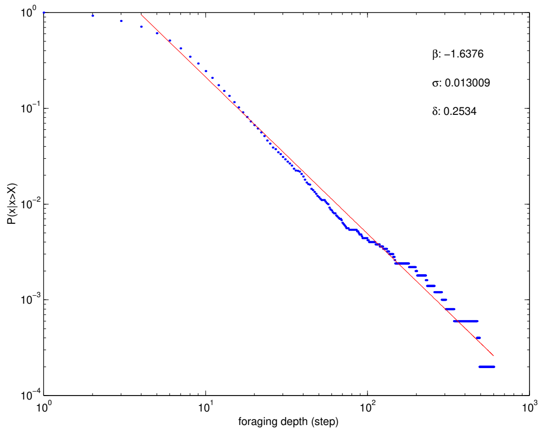
THE PARAMETERS FOR EXPERIMENT 2.

what will be a reasonable *degree-of-coupling* for agents. We believe that there should be an ideal r value in our model. In order to answer this question, we will conduct Experiment 4 to examine the results under different r values. In Experiment 4, we will gradually alter r, while keeping the rest of parameters the same as those in Experiment 2.

Figures 13(a) and (b) show the *power* values in the observed power-law distributions of foraging depth and the average foraging steps, with respect to degree-of-coupling, r, respectively. From Figure 13(a), we find that the power β_c is increasing with some fluctuations. From Figure 13(b), we note that the values of average step by rational agents are higher than those of recurrent agents. The explanation for this result is that the ability to find relevant information by rational agents is weaker than that by recurrent agents, and thus rational agents must go through more pages in order



(a)



(b)

Fig. 7. Agent foraging depth observed in Experiment 2, where *the Content Distribution follows a power law*, different from that of Experiment 1. (a) Cumulative distribution of foraging depth in accessing Web pages by *recurrent* agents. '·' corresponds to experimental data, and '$-$' corresponds to a linear-regression fitted line. The obtained distribution follows a power law with power $\beta_c = -1.532$ and residual $\sigma = 0.0145$. (b) Cumulative distribution of foraging depth by *rational* agents. The distribution follows a power law with power $\beta = -1.638$ and residual $\sigma = 0.013$.

to be satisfied. Consequently, their satisfaction rate will be lower than that of recurrent agents, as shown in Figure 14(a).

Website owners usually hope that users can stay longer or surf deeper at their Websites while viewing information, and at the same time, satisfy their interests. Figure 14(b) shows the combined measure of agent foraging depth and satisfaction rate. From Figure 14(b), we observe that in order to get an optimal effect, the value of degree-of-coupling, r, should be set to $0.7 \sim 0.8$. In such a case, the average link number per node is about $9 \sim 16$, as shown in Figure 12.

D. *Mixed Agent Population*

In real-world Web surfing, different users who visit a certain Website can have distinct navigation strategies. Some users may fall in the category of recurrent users, while others may be new comers. When the new comers feel that the Website contains or leads to some contents of interest to them, they will become more likely to visit the Website again. It is important for the designer of a Website to recognize from emergent Web regularities the underlying dominant navigation strategies of users.

So far, we have observed the regularities produced by three categories of information foraging agents with various Interest Profile Distributions. It may be noted that recurrent and random agents are two extreme cases, whereas rational agents have the ability to predict the next-level contents that is between the abilities of recurrent and random agents. The fact that all categories

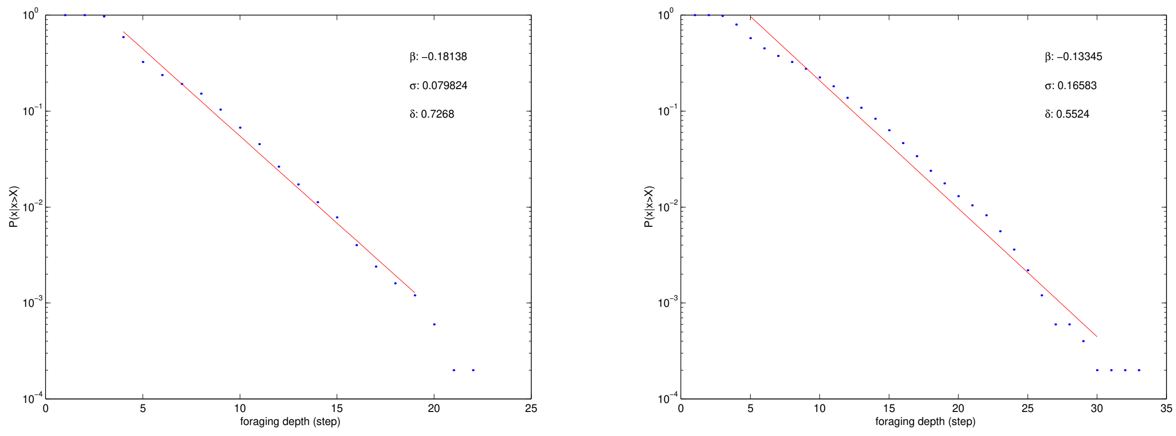
Degree-of-coupling, r	0.7
Number of agents	5,000
Number of nodes	254
Number of topics	10
T_c	0.1
ΔM_t^c	0.2
ΔM_t^v	1st
ϕ	1
λ	0.6
μ_m	5.97
μ_t	1.0
σ_m	0.8
σ_p	0.25
σ_t	0.2
σ_u	0.5
θ	1

TABLE III
THE PARAMETERS FOR EXPERIMENT 3.

of users may be involved in bringing about the emergent regularities in Web surfing has led us to the following question: What will be the distributions of foraging depth and link-click-frequency if all three categories of information agents are involved? In order to examine this case, we have conducted Experiment 5, where all three categories of agents, *i.e.*, recurrent, rational, and random agents, are involved and the number of agents in each group is $5,000$. Figure 15 presents the results of Experiment 5.

From Figure 15(a), it can be observed that there exists a strong regularity in the probability distribution of foraging depth in accessing Web pages in the case of mixed agent population. The obtained result is consistent to the regularities found in empirical Web log datasets. Figure 15(b) presents the distribution of foraging depth in accessing domains, which, like the real-world statistics, follows an exponential function. Figure 15(c) shows the power-law distribution of link-click-frequency. In Figure 15(c), the occurrence point of the most probable link-click-frequency is not at 1. This is because the number of agents is too large as compared to the number of links.

To summarize, emergent regularities can readily be observed when information foraging agents make use of different navigation strategies. As far as the satisfaction rate is concerned, the mixed agent population is relatively easier to satisfy than rational agents, but more difficult than recurrent agents, as we have already shown in Figure 14(b). One way to increase the level of satisfaction rate would be to improve the descriptions of hyperlinks such that they are topic-specific and informative to foraging agents.



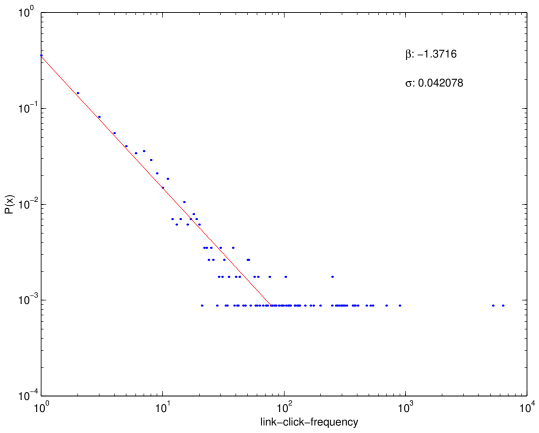
(a) (b)

Fig. 8. Agent foraging depth observed in Experiment 3, where *the Interest Profiles of agents follow a normal-distribution model*, different from that of Experiment 2. (a) Cumulative distribution of foraging depth in accessing Web pages by *recurrent* agents. '·' corresponds to experimental data, and '−' corresponds to a linear-regression fitted line. The obtained distribution follows an *exponential function* with exponent $\beta_c = -0.18$ and residual $\sigma = 0.08$. (b) Cumulative distribution of foraging depth by *rational* agents. The distribution follows an *exponential function* with exponent $\beta = -0.133$ and residual $\sigma = 0.17$.

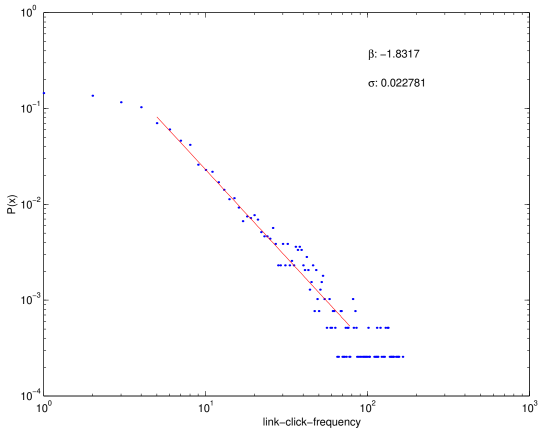
E. Satisfaction vs. Unsatisfaction

In the preceding sections, we have considered and classified agents with different navigation strategies depending on whether they are proficient users (recurrent), content explorers (rational), or curious users (random). In each case, an agent will leave the Web server either with the contents it has found or without any success. In Experiment 6, we are interested in the difference in the foraging-depth distributions between satisfied and unsatisfied agents. We will use the same agent data and the same Web server as those in Experiment 2, except that the motivation update mechanism for agents will be defined using Equation 29. The parameters of Experiment 6 are given in Table IV.

Figure 16 shows the distributions of satisfied and unsatisfied *recurrent* agents, whereas Figure 17 shows the distributions of satisfied and unsatisfied *rational* agents. From both figures, we can observe that the regularities can be found in both satisfied agents and unsatisfied agents cases. This experiment also demonstrates that the regularities will not be affected by the motivation update mechanism. From Figures 16 and 17, we also find that the distribution of unsatisfied agents has a heavier tail (*i.e.* higher values) than that of satisfied agents. Figure 18 presents the parameter distributions in Experiment 6, with respect to the Web structure parameter, degree-of-coupling, r.



(a)



(b)

Fig. 9. Distribtuions of link-click-frequency in Experiment 2. (a) Distribution of link-click-frequency for *recurrent* agents. The distribution tail is approximately a power law with power $\beta_l = -1.83$, as obtained by weighted linear regression. (b) Distribution of link-click-frequency for *rational* agents. The distribution is approximately a power law with power $\beta_l = -1.37$, as obtained by weighted linear regression.

F. Remarks on Simulation and Validation

F.1 Parameters in Simulation

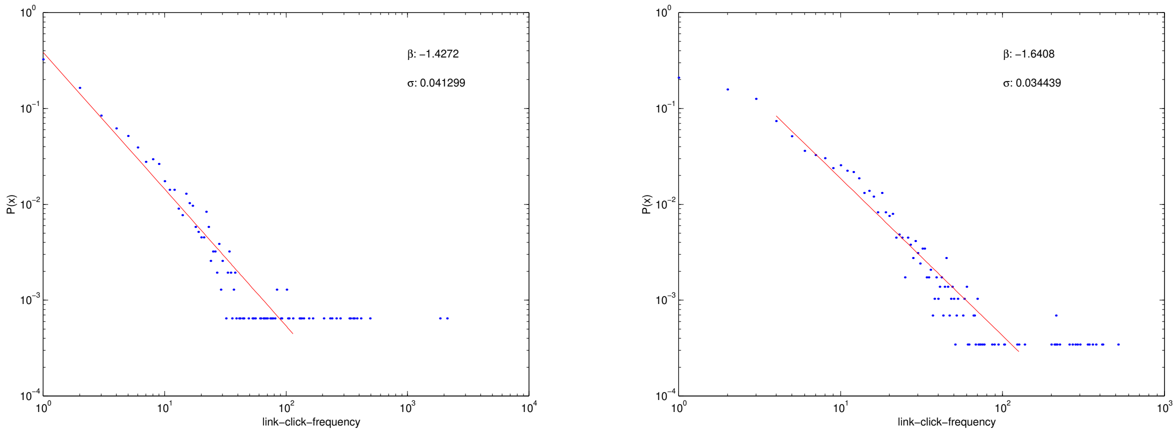
In our present studies, we have selected the number of nodes and the number of topics in an artificial Web server based on the following general considerations:

1. the simulation experiments should be computationally manageable to obtain results;
2. the order of magnitude should be somewhat comparable to those of the empirical datasets used for validation.

There are also other adjustable parameters in the above-mentioned simulation. In order to test the sensitivity of the parameters, we have conducted other experiments to examine the possible effects on distribution regularities while changing the number of agents, the number of domains or topics, and the parameters for the distribution of agent Interest Profiles and for the Content Distribution in the Web server. The results of our experiments reveal that altering these parameters will not change the regularities of power-law or exponential distributions as mentioned above, but alter the shape parameters for the distributions. This further indicates that the distribution regularities emerged from agent foraging behavior is stable and ubiquitous.

F.2 The Web Server Traces for Validation

In this work, we have used real-world Web traces to experimentally validate the regularities observed from our model. The advantages of using such traces are:



(a) (b)

Fig. 10. Distribtuions of link-click-frequency in Experiment 3. (a) Distribution of link-click-frequency for *recurrent* agents. The distribution tail is approximately a power law with power $\beta_l = -1.64$, as obtained by weighted linear regression. (b) Distribution of link-click-frequency for *rational* agents. The distribution is approximately a power law with power $\beta_l = -1.43$, as obtained by weighted linear regression.

1. it is relatively easier to compare with related work on characterizing strong Web regularities, *e.g.*, those reported in [3];
2. the regularities empirically observed from the server traces are less biased towards particular user segments;
3. in Web servers such as the NASA Kennedy Space Center Web server, most of the Web pages were centrally maintained and linked with only few out-going links.

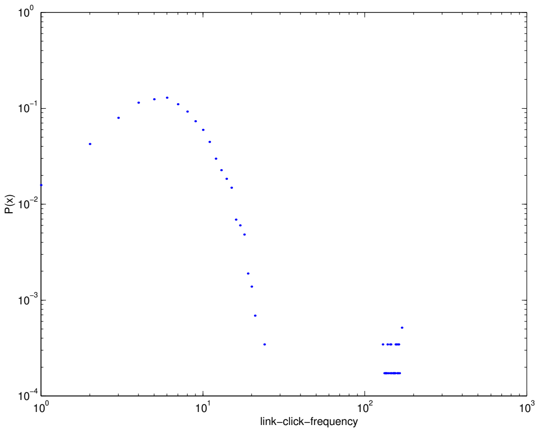
However, it should be mentioned that generally speaking, due to the possible linkages with other Web servers, only a Web server log might not be sufficiently capture complete user access sessions[1]. In addition, one html request may not correspond to one click in some cases.

As client-based traces of Web traffic become widely available and less biased, it would be interesting to conduct a further validation of our simulation-based characterization study using such traces. From the client-based trace characteristics as highlighted in [11], [51], we conjecture that the results would be similar to those presented in this paper.

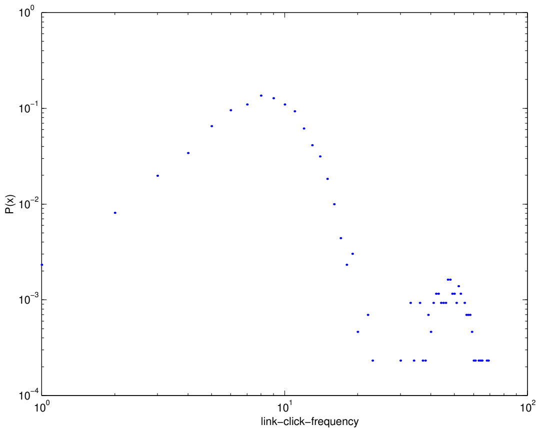
VII. Conclusion

In this paper, we have demonstrated an agent-based modeling approach to characterizing empirical Web usage regularities. In particular, we have formulated an information foraging agent-based model and validated this model against some empirical Web log datasets. We have found that:

[1] In our simulation experiments, this issue was in part taken into account by truncating a foraging session that reaches the *boundary* of the artificial Web server.



(a)



(b)

Fig. 11. Distributions of link-click-frequency for *random* agents. (a) Random agents in Experiment 1. (b) Random agents in Experiment 2.

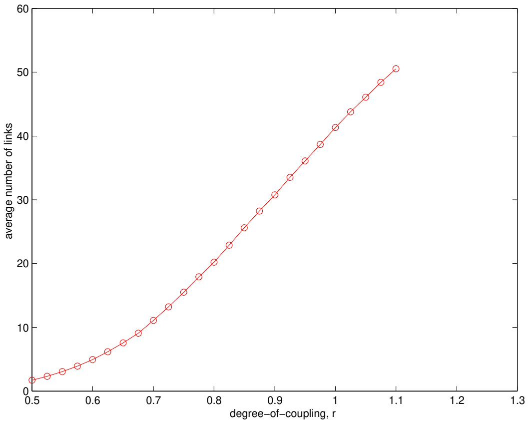
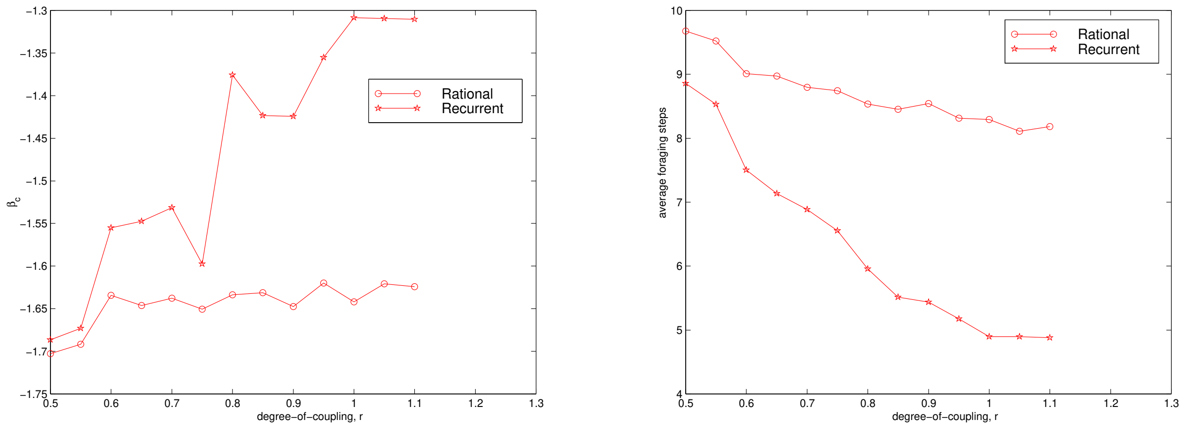


Fig. 12. The average number of links with respect to degree-of-coupling, r, in Experiment 4.

1. Our white-box model, incorporating the behavioral characteristics (*i.e.*, motivation aggregation) of Web users with measurable and adjustable factors, does exhibit the regularities as found in empirical Web data. The foraging operations in the model correspond to the surfing operations in the real-world Web server.
2. Different navigation strategies can lead to different emergent regularities. From random agents to recurrent agents, the *power law* in link-click-frequency distribution will become more and more obvious. The only distinction among the different categories of agents in our information foraging model is their ability to predict which of linked next-level pages may contain more interesting contents.



(a) (b)

Fig. 13. (a) The *power* values, β_c, in the observed power-law distributions of foraging depth and (b) the *average foraging steps*, with respect to degree-of-coupling, r, in Experiment 4. '$\circ$' corresponds to *rational* agents and '$\star$' corresponds to *recurrent* agents.

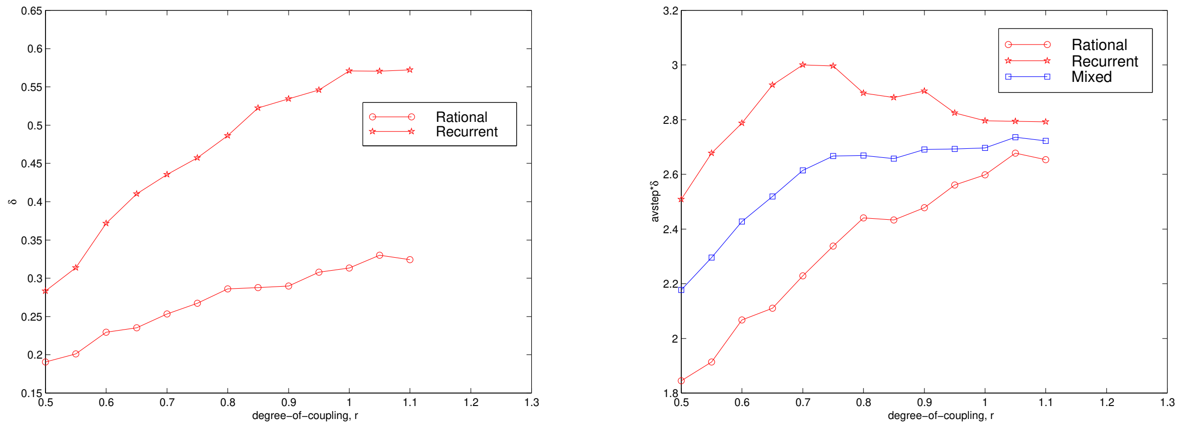
3. As far as the distribution of user Interest Profiles underlying emergent regularities is concerned, the regularities of power-law distributions observed in agent foraging depth in accessing Web pages are largely resulted from the power-law distribution of user interests in various topics.
4. The regularity of agent foraging depth in accessing Web pages may not be affected by how Web contents are distributed among Websites.
5. Further, if we separately record users who can successfully find relevant information and those who fail to do so, we can still observe those regularities.

In summary, our work offers a means for explaining strong Web regularities with respect to user Interest Profiles, Web Content Distribution and coupling, and user navigation strategies. It enables us to predict the effects on emergent usage regularities if certain aspects of Web servers or user foraging behaviors are changed.

While presenting an interesting and promising research direction, we should point out that one of the useful extensions for future work would be to show how the quantitative representations or constructs as used in modeling Web contents and user interest profiles are manifested in the real-world Web.

REFERENCES

[1] R. Albert, H. Jeong, and A.-L. Barabasi, "Diameter of World-Wide Web," *Nature*, vol. 410, pp. 130–131, September 9, 1999.
[2] S. Lawrence and C. L. Giles, "Accessibility of information on the Web," *Nature*, vol. 400, pp. 107–109, 1999.
[3] B. A. Huberman, P. L.T. Pirolli, J. E. Pitkow, and R. M. Lukose, "Strong regularities in World Wide Web surfing," *Science*, vol. 280, pp. 96–97, April 3, 1997.
[4] Graphics, Visualization, and Usability Center, "GVU's WWW user surveys," http://www.gvu.gatech.edu/user_surveys.



(a) (b)

Fig. 14. (a) The satisfaction rate, δ, and (b) the combined measure of agent foraging depth and satisfaction rate, with respect to degree-of-coupling, r, in Experiment 4. '$\circ$' corresponds to *rational* agents and '$\star$' corresponds to *recurrent* agents.

[5] S. Cbakrabarti, B. E. Dom, D. Gibson, and J. Kleinberg, "Mining the Web's link structure," *IEEE Computer*, vol. 32, no. 8, pp. 60–67, August 1999.

[6] D. Gibson, J. Kleinberg, and P. Raghavan, "Inferring Web communities from link topology," in *Proceedings of the Ninth ACM Conference on Hypertext and Hypermedia*, 1998, pp. 225–234.

[7] B. Mobasher, N. Jain, E. Han, and J. Srivastava, "Web mining: Pattern discovery from World Wide Web transactions," Tech. Rep. TR-96050, Department of Computer Science, University of Minnesota, 1996.

[8] J. E. Pitkow, "Summary of WWW characterizations," *Computer Networks and ISDN Systems*, vol. 30, no. 1–7, pp. 551–558, 1998.

[9] R. Cooley, J. Srivastava, and B. Mobasher, "Web mining: Information and pattern discovery on the World Wide Web," in *Proceedings of the NiththIEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November 1997, pp. 558–567.

[10] L. D. Catledge and J. E. Pitkow, "Characterizing browsing strategies in the World-Wide Web," *Computer Networks and ISDN Systems*, vol. 26, no. 6, pp. 1065–1073, 1995.

[11] C. R. Cuhna, A. Bestavros, and M. E. Crovella, "Characteristics of WWW client based traces," Tech. Rep. BU-CS-95-010, Computer Science Department, Boston University, 1995.

[12] R. Cooley, P.-N. Tan, and J. Srivastava, "Discovery of interesting usage patterns from Web data," in *Proceedings of WE-BKDD'99*. August 1999, pp. 163–182, Springer-Verlag.

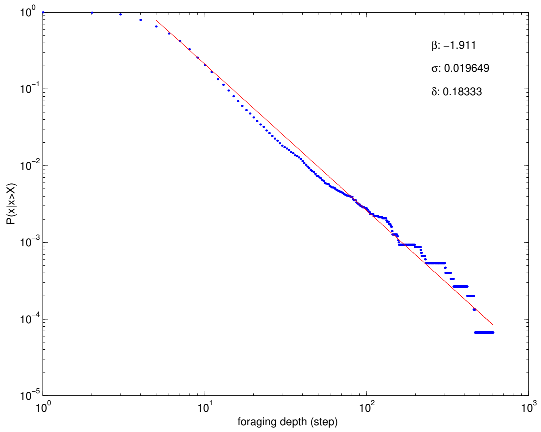
[13] J. Pei, J. Han, B. Mortazavi-asl, and H. Zhu, "Mining access patterns efficiently from Web logs," in *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2000)*, April 2000.

[14] M. Spiliopoulou, "The laborious way from data mining to Web log mining," *International Journal of Computer Systems Science and Engineering: Special Issue on Semantics of the Web*, vol. 14, pp. 113–126, March 1999.

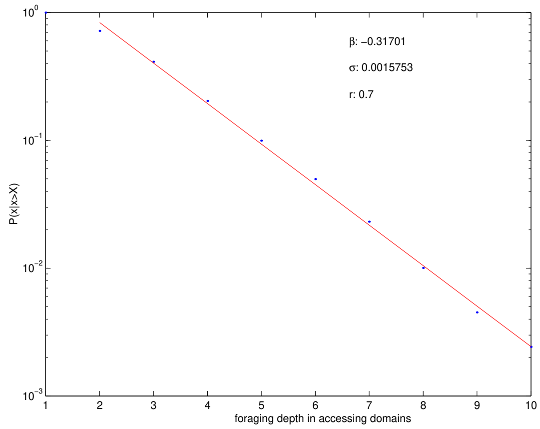
[15] M. Spiliopoulou, C. Pohle, and L. Faulstich, "Improving the effectiveness of a Web site with Web usage mining," in *Proceedings of WEBKDD99*. August 1999, Springer-Verlag.

[16] O. R. Zaane, M. Xin, and J. Han, "Discovering Web access patterns and trends by applying OLAP and data mining technology on Web logs," in *Proceedings of Advances in Digital Libraries (ADL'98)*, April 1998, pp. 19–29.

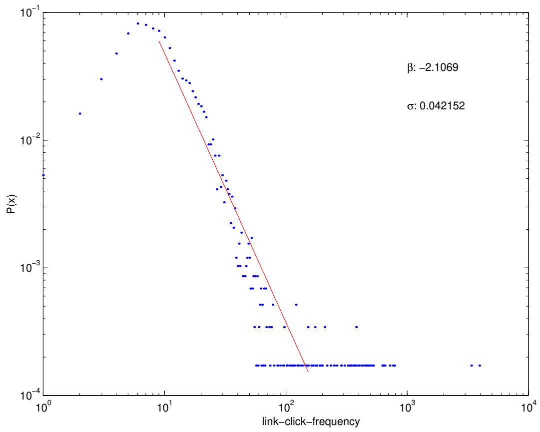
[17] A. Joshi and R. Krishnapuram, "On mining Web access logs," in *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2000, pp. 63–69.



(a)



(b)



(c)

Fig. 15. A *mixed population* of recurrent, rational, and random agents in Experiment 5. (a) Cumulative distribution of foraging depth, which has a power tail. (b) Cumulative distribution of foraging depth in accessing domains. (c) Distribution of link-click-frequency.

[18] O. Nasraoui, H. Frigui, A. Joshi, and R. Krishnapuram, "Mining Web access logs using relational competitive fuzzy clustering," in *Proceedings of the Eighth International Fuzzy Systems Association World Congress (IFSA'99)*, August 1999.

[19] T. W. Yan, M. Jacobsen, H. Garcia-Molina, and U. Dayal, "From user access patterns to dynamic hypertext linking," in *Proceedings of the Fifth World Wide Web Conference (WWW5)*, May 1996, pp. 1007–1014.

[20] P. Barford, A. Bestavros, A. Bradley, and M. Crovella, "Changes in Web client access patterns: Characteristics and caching implications," *World Wide Web, Special Issue on Characterization and Performance Evaluation*, vol. 2, pp. 15–28, 1999.

[21] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," Tech. Rep. 1371, Computer Sciences Department, University of Wisconsin-Madison, April 1998.

[22] S. Glassman, "A caching relay for the World Wide Web," *Computer Networks and ISDN Systems*, vol. 27, no. 2, pp. 165–173, 1994.

[23] V. Padmanabhan and J. Mogul, "Using predictive prefetching to improve World-Wide Web latency," in *Proceedings of the SIGCOMM'96 Conference*, 1996.

Degree-of-coupling, r	0.5∼1.1
Number of agents	5,000
Number of nodes	254
Number of topics	10
T_c	0.1
ΔM_t^c	0.2
ΔM_t^v	2nd
α_m	0.03
γ_m	0.2
ϕ	1.2
λ	0.5
μ_t	1.0
σ_p	0.25
σ_t	0.2
σ_u	0.5
θ	1

TABLE IV
THE PARAMETERS FOR EXPERIMENT 6.

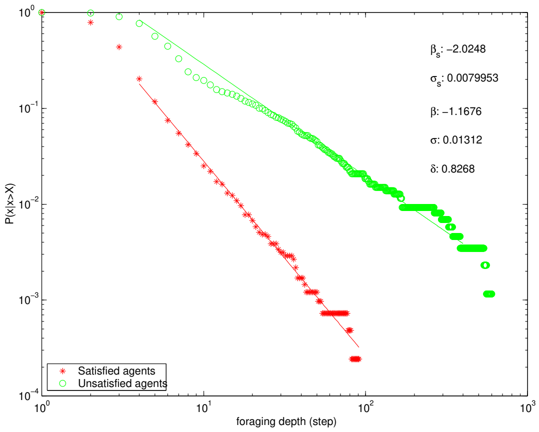


Fig. 16. Cumulative distribution of foraging depth for *recurrent* agents in Experiment 6, with r=0.7. '$\circ$' corresponds to unsatisfied agents and '$\star$' corresponds to satisfied agents.

[24] C. R. Cuhna and C. Jaccoud, "Determining www user's next access and its application to pre-fetching," in *Proceedings of the Second IEEE Symposium on Computers and Communications (ISCC'97)*, July 1997.

[25] M. F. Arlitt and C. L. Williamson, "Web server workload characterization: The search for invariants," in *Proceedings of the ACM SIGMETRICS'96 Conference*, Philadelphia, PA, April 1996, pp. 126–137.

[26] P. Barford and M. Crovella, "Generating representative Web workloads for network and server performance evaluation," in *Measurement and Modeling of Computer Systems: Proceedings of the ACM SIGMETRICS Conference*, July 1998, pp. 151–160.

[27] J. Mogul, "Network behavior of a busy Web server and its clients," Tech. Rep. TR-95.5, Digital Western Research Laboratory, CA., 1995.

[28] S. Madria, S. S. Bhowmick, W.-K. NG, and R. P. Lim, "Research issues in Web data mining," http://www.sinokdd.163.net/paper/webmining.html.

[29] G. K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge, MA, 1949.

[30] M. E. Crovella and M. S. Taqqu, "Estimating the heavy tail index from scaling properties," *Methodology and Computing in*

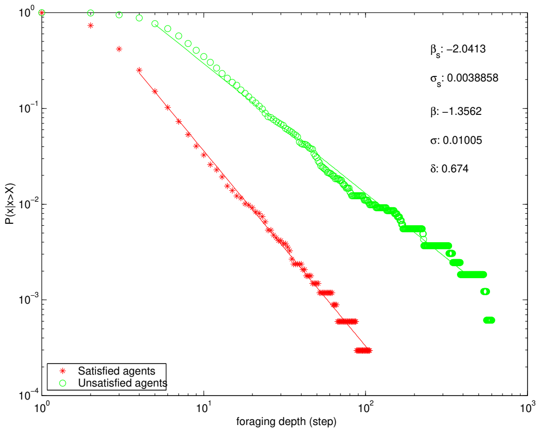
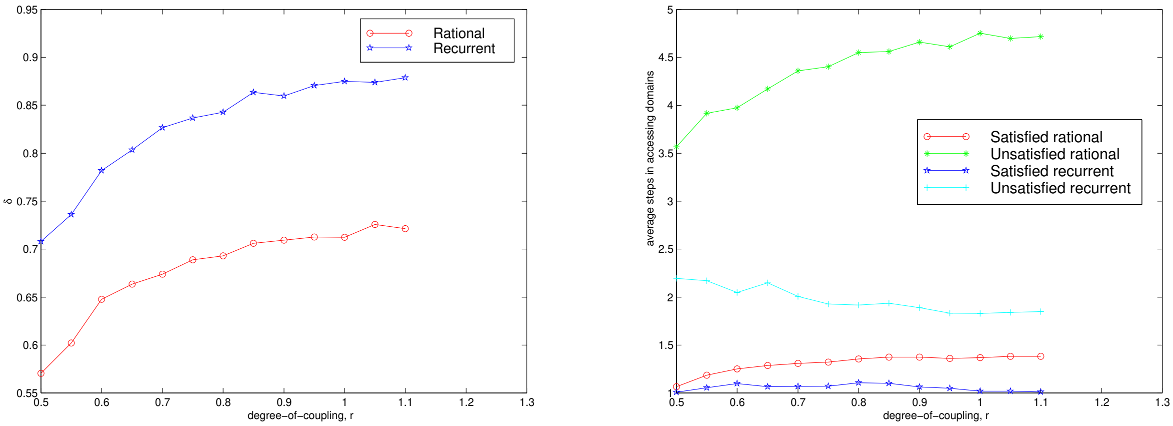


Fig. 17. Cumulative distribution of foraging depth for *rational* agents in Experiment 6, with r=0.7. '$\circ$' corresponds to unsatisfied agents and '$\star$' corresponds to satisfied agents.



(a) (b)

Fig. 18. Parameter distributions in Experiment 6, with r changing from 0.5 to 1.1. (a) The satisfaction rate, δ, and (b) the average steps in accessing domains, with respect to r.

Applied Probability, vol. 1, no. 1, pp. 55–79, 1999.

[31] L. A. Adamic and B. A. Huberman, "The nature of markets in the World Wide Web," http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.

[32] S. M. Maurer and B. A. Huberman, "Competitive dynamics of Web sites," http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.

[33] D. Helbing, B. A. Huberman, and S. M. Maurer, "Optimizing traffic in virtual and real space," in *Traffic and Granular Flow '99: Social, Traffic, and Granular Dynamics*, D. Helbing, H. J. Herrmann, M. Schreckenberg, and D. E. Wolf, Eds. Springer, Berlin, 2000.

[34] B. A. Huberman and L. A. Adamic, "Evolutionary dynamics of the World Wide Web," http://www.parc.xerox.com/istl/groups/iea/www/growth.html.

[35] E. Adar and B. A. Huberman, "The economics of surfing," http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.

[36] M. Levene, J. Borges, and G. Loizou, "Zipf's law for Web surfers," *Knowledge and Information Systems*, vol. 3, pp. 120–129, 2001.

[37] R. M. Lukose and B. A. Huberman, "Surfing as a real option," in *Proceedings of the International Conference on Information and Computation Economics*, October 1998, pp. 45–51.

[38] A. Johansen and D. Sornette, "Download relaxation dynamics on the WWW following newspapers publication of URL," in *IMA 'Hot Topics' Workshop: Scaling Phenomena in Communication Networks*, October 1999.

[39] L. A. Adamic and B. A. Huberman, "Technical comment to 'Emergence of scaling in random networks', vol. 286 15 science october 1999, pp.509-512," http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.

[40] A.-L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, October 15, 1999.

[41] A.-L. Barabasi, R. Albert, and H. Jeong, "Scale-free characteristics of random networks: The topology of the World Wide Web," *Physica A*, vol. 281, pp. 69–77, 2000.

[42] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, "Graph structure in the Web," in *Proceedings of the Ninth World Wide Web Conference (WWW9)*, May 2000.

[43] B. A. Huberman and L. A. Adamic, "Growth dynamics of the World-Wide Web," *Nature*, vol. 410, pp. 131, September 9, 1999.

[44] M. Levene and G. Loizou, "Computing the entropy of user navigation in the Web," Research Note RN/99/42, Department of Computer Science, University College London, 1999.

[45] A. Thatcher, "Determining interests and motives in WWW navigation," in *Proceedings of the Second International Cyberspace Conference on Ergonomics (CybErg1999)*, 1999.

[46] F. Menczer, "Mapping the sementics of Web text and links," *IEEE Journal on Selected Areas in Communications*, , no. submitted, 2002.

[47] F. Menczer, "Lexical and semantic clustering by Web links," *IEEE Transactions on Knowledge and Data Engineering*, , no. submitted, 2002.

[48] G. W. Flake, S. Lawrence, C. L. Giles, and F. Coetzee, "Self-organization of the Web and identification of communities," *IEEE Computer*, vol. 35, no. 3, pp. 66–71, 2002.

[49] J. F. Cove and B.C. Walsh, "Online text retrieval via browsing," *Information Processing and Management*, vol. 24, no. 1, pp. 31–37, 1988.

[50] D. E. Kieras, D. E. Meyer, S. T. Mueller, and T. L. Seymour, "Insights into working memory from the perspective of The EPIC Architecture for modeling skilled perceptual-motor and cognitive human performance," in *Models of Working Memory*, A. Miyaki and P. Shah, Eds. Cambridge University Press, New York, 1999.

[51] J. E. Pitkow, "Summary of WWW characteristics," *The World Wide Web Journal*, vol. 2, no. 2, pp. 2–13, 1999.