1. Agent-Based Characterization of Web Regularities

Jiming Liu[1], Shiwu Zhang[1,2], and Yiming Ye[3]

[1]Department of Computer Science, Hong Kong Baptist University
Kowloon Tong, Hong Kong Email: jiming@comp.hkbu.edu.hk
[2]University of Science and Technology of China
[3]IBM T. J. Watson Research Center
Yorktown Heights, NY, USA Email: yiming@us.ibm.com

In Web Intelligence (WI) applications, it is a common practice to record Web log data. What remains a great challenge in Web log mining is how to characterize the underlying user behavior from the obtained data. In this chapter, we will focus on how to interpret strong regularities in Web surfing in terms of user decision-making patterns, and present an information foraging agent-based approach to characterizing user behavior. The experimental results based on information foraging agents enable us not only to capture the empirical regularities collected from a real-world Web site, but also to effectively unveil the underlying decision-making mechanisms in user surfing as well as how variables in such mechanisms can affect the empirically observed emergent regularities.

1.1 Introduction

One of the most important tasks in Web mining is called *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.

1.1.1 Empirical Regularities on the World Wide Web

Recently, researchers have identified a spectrum of 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 [1.27] or a power law. That is, the probability P of a variant taking value k is proportional to $k^{-\alpha}$, where α is from 0 to 2. The distribution presents a heavy tail if its upper tail declines like a power law [1.12].

Here is a subset 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 [1.8, 1.10, 1.13, 1.14].
2. The popularity of Web sites 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 [1.1, 1.10, 1.24].
3. The request inter-arrivals and Web latencies follow a heavy-tail distribution [1.9, 1.16, 1.26].
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 [1.5, 1.8, 1.9, 1.13].
5. The number of pages either across all Web sites or within a certain domain of the Web follows a power law [1.17].
6. The trace length of users within a proxy or a Web site, or across the Web follows a power law [1.3, 1.19, 1.21, 1.23].
7. The dynamical response of the Web to a Dirac-like perturbation follows a power law [1.20].
8. The distribution of links (both incoming and outgoing) among Web sites or pages follows a power law [1.2, 1.4, 1.6, 1.7, 1.11].

1.1.2 Regularity Characterization

Although researchers have empirically observed strong regularities on the Web, few of them have dealt with the issue of how such regularities are emerged. 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 [1.6], a random-network model with growth and preferential attachment factors is proposed that produces a power distribution of link number over Web sites or pages. Huberman [1.18] showed that the power-law distribution of page number over various Web sites can be characterized based on a stochastic multiplicative growth model coupled by the fact that Web sites 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 [1.19, 1.23]. Levene [1.21, 1.22] developed an absorbing Markov-chain model to simulate the power-law distribution of user navigation depth on the Web.

1.2 Problem Statements

The random-walk model [1.19, 1.23] and the Markov-chain model [1.22, 1.21] 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 World Wide Web. For instance, the random-walk model does not take into account the structure of the Web or the motivation of users. 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 [1.25]. 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.

1.3 An Overview of Foraging Agent-Based Web Characterization

We view users as *information foraging agents* living in the Web space. The Web space is essentially a collection of Web pages connected by hyperlinks. Each page contains certain information, and each hyperlink between two pages signifies certain content similarity between them. The content contained in a page can be characterized using a multi-dimensional *information vector* where each element corresponds to the relative information weight on a certain topic. Each agent foraging in the Web space has its special interests, varying from finding information in several topics to just wandering in the Web space. Agent intention can also be characterized with a multi-dimensional *interest vector*. When an agent finds certain Web pages in which the information is close to its interested topic(s), it will be more willing to *forage* to the pages at the next level, that is, it gets more motivational support to *surf* deeper. On the other hand, when an agent does not find any information after several foraging steps or it has found enough information satisfying its interests, it will stop navigating and exit the Web space. We can obtain certain statistical properties by recording individual agent foraging behavior.

1.4 Foraging Agents

1.4.1 Artificial Web Space

First, we construct an artificial Web space whose parameters can be assessed and altered. A common Web site usually contains several related topics linked by URLs. Each topic includes a certain number of pages, and the hyperlinks within the pages that belong to the same or different topics form the whole Web site unit. We may view a Web site as a graph consisting of nodes and links; nodes denote pages and links denote hyperlinks within pages. The information in a certain node can be represented as follows:

$$\mathtt{VP}_n = [vp_{n1}, vp_{n2} \ldots vp_{ni} \ldots vp_{nm}] \tag{1.1}$$

where
- $\mathtt{VP}_n$: the information vector of node (i.e., page) n,
- vp_{ni}: the relative information weight on topic i,
- m: the number of topics.

A hyperlink indicates that the content information in two nodes is very similar or relevant. We assume that there is a hyperlink within two nodes *only if* the two nodes are very similar or relevant, as reflected in the real-world Web sites. Here we describe the similarity between two nodes using their Euclidean distance, thus:

$$L(i,j) = D(\mathtt{VP}_i, \mathtt{VP}_j) = \left(\sum_{k=1}^{m}(vp_{ik} - vp_{jk})^2\right)^{1/2} \tag{1.2}$$

where
- $L(i,j)$: the similarity between nodes i and j,
- D: an Euclidean distance function.

With the above definition, the information organization at a Web site can be characterized quantitatively. We first create several groups of nodes, and each group corresponds to a single topic. The weights of the nodes belonging to group j are initialized as follows:

$$vp_i = \begin{cases} |randn(0, \sigma_p)|, & \text{if } i \neq j, \\ |T- \mid randn(0, \sigma_p) \mid \ |, & otherwise. \end{cases} \tag{1.3}$$

where
- $randn(0, \sigma_p)$: random number generator based on a normal distribution of mean 0 and variance σ_p,
- T: a constant.

We can adjust σ_p to get various topic distributions in the Web space; the smaller the σ_p is, the more focused the topic is. After acquiring the node vectors, it is important to build links inside these nodes. Since each link indicates certain similarity between two connected nodes, we will only build a link between two nodes if their similarity measure exceeds a predefined threshold, that is, the distance between the two nodes is less than a distance threshold, r. Threshold r is meaningful both in our artificial Web space and in the real-world Web; increasing the distance threshold, r, corresponds to increasing the number of links at a Web site as well as the site's average link distance, ALD, which implies that the average similarity between the contents of two nodes will decrease. In this respect, r reflects the connectivity of the Web. We refer it to as *accessibility*. The average link distance, ALD, is defined as follows:

$$\mathtt{ALD} = \frac{1}{\mathtt{NL}} \left(\sum_{i,j} L(i,j) \right) \tag{1.4}$$

where NL denotes the number of links.

Next, we assume that the agent starts its foraging from a home page that contains an index for relevant topics. In our experiment, the home page, p, is assigned as equal distance to each topic, that is:

$$vp_{pk} = T_p, \; k = 1 \ldots m \tag{1.5}$$

where vp_{pk} denotes the information contained in topic k of the Web page. T_p denotes the information weight on a topic.

The routine for constructing an artificial Web site is given as follows:

```
For each topic k
    Create node vectors
End
For each node i
    Initialize link list of node i
    For each node j
        If D(VP_i, VP_j) < r
            Add node j to link list of node i
            Add D(VP_i, VP_j) to link list of node i
        End
    End
End
```

Thus, by executing the above routine, an artificial Web site and its description can be obtained.

1.4.2 Interests of Foraging Agents

Due to the fact that agents visiting a Web site are diverse in many ways, including their interests in different topics, we define and initialize an agent interest vector, $\mathtt{VU}_n$, as follows:

$$\begin{aligned} \mathtt{VU}_n &= [vu_{n1}, vu_{n2} \ldots vu_{ni} \ldots vu_{nm}] \\ vu_{ni} &= |randn(0, \sigma_u)| \end{aligned} \tag{1.6}$$

where

$\mathtt{VU}_n$: an interest vector for agent n,
vu_{ni}: the agent interest weight on topic i.

Parameter σ_u corresponds to the interest variation of agents. Note that the interest vector of an agent can dynamically change during the course of navigating a Web site. When the agent realizes that it has retrieved useful information in a certain topic, its interest vector will be updated as the objective in searching other topics decays.

1.4.3 Motivational Support Aggregation

Another important parameter that characterizes agent foraging behavior is *motivational support*, S_t. Motivational support serves as the energy to stimulate an agent to forage further and deeper. If the agent reaches some useful information, it will get rewarded, and this support will go up. Otherwise, the support will go down. If the support exceeds a certain threshold, which implies that it has got enough information, the agent will end its foraging. On the other hand, if the support is too low, it will become impatient and quit foraging at the site. The motivational support is defined as follows:

$$S_{t+1} = S_t + \Delta M_t + \Delta R_t \tag{1.7}$$

where
S_t: motivational support at step t,
ΔM_t: motivation change at step t,
ΔR_t: reward received at step t.

In our foraging agent model, we define an initial support and let the maximum support threshold be directly proportional to the sum of its interest vector. Specifically, we assign the initial support and the maximum and minimum supports as follow:

$\mathtt{S}_{\mathtt{init}} = \frac{1}{2}(\sum_{i=1}^{m} vu_i)$,
$\mathtt{S}_{\mathtt{max}} = \sum_{i=1}^{m} vu_i$,
$\mathtt{S}_{\mathtt{min}} = 0$.

Other parameters describing agent foraging behavior include: agent step or foraging depth, agent location, and agent satisfaction factor.

1.4.4 Characterization of Foraging Decisions

It is empirically accepted that a hyperlink denotes the similarity between two pages, and the names of URLs usually describe the main issues of the pages that the URLs point to. An agent can, generally speaking, judge by looking at hyperlinks whether or not the pages are relevant to its interest, which means that the agent may predict the contents of the next level pages. However, since the prediction abilities of individual agents are different, the agents will be classified into several categories according to their prediction abilities.

Suppose that agent i, U_i, is in page j now. There are h hyperlinks in page j. We define and compute the probability of clicking on a hyperlink by each class of agents as follows:

1. *recurrent agent*: A recurrent agent is usually familiar with the Web structure, it can find the useful information right away. Each time when it decides to navigate further, it can find the page that is most relevant to its interest or objective. The probability of selecting pages at the next step is defined as follows:
$$\mathcal{P}_k = \begin{cases} 1, \; if D(\mathtt{VU}_i, \mathtt{VP}_k) = \min(D(\mathtt{VU}_i, \mathtt{VP}_j)), \\ 0, \; otherwise. \end{cases} \tag{1.8}$$
where $j = 1 \ldots h$.
2. *rational agent*: An agent in this category is new to the Web site. It knows clearly what it wants and selects a direction based on the information of hyperlinks. But, it cannot select a direction whose information vector is most close to its interest vector, because it is unfamiliar with the structure of the Web. The probability of reaching the next level node k, $\mathcal{P}_k$, can be derived as follows:
$$\mathcal{P}_k = \frac{D(\mathtt{VU}_i, \mathtt{VP}_k)^{-1}}{\sum_{j=1}^{h} D(\mathtt{VU}_i, \mathtt{VP}_j)^{-1}} \tag{1.9}$$
where $j = 1 \ldots h$.
In our experiments, we find that it is hard to select the next level nodes well with this method, especially in situations where there are many outgoing links. In order to emphasize the prediction ability, we define the probability in our implementation as follows:
$$\mathcal{Q}_j = D(\mathtt{VU}_i, \mathtt{VP}_j) - \mathrm{mean}(D(\mathtt{VU}_i, \mathtt{VP}_j)), \qquad j = 1 \ldots h \tag{1.10}$$
$$\mathcal{Q}_j = 0 \qquad if \; \mathcal{Q}_j < 0 \tag{1.11}$$
$$\mathcal{P}_k = \frac{\mathcal{Q}_k^{-1}}{\sum_{j=1}^{h} \mathcal{Q}_j^{-1}}, \qquad \mathcal{Q}_k \neq 0 \tag{1.12}$$
3. *random agent*: A random agent has no strong intention to get something and just wanders among pages. It randomly selects a direction to surf. The probability of reaching node k at the next step, $\mathcal{P}_k$, is computed as follows:
$$\mathcal{P}_k = \frac{1}{h}, \qquad k = 1 \ldots h \tag{1.13}$$

It should be mentioned that agent interest vector varies in the process of foraging. When an agent reaches and finishes reading page i, its interest will change according to this page's information vector. The updating mechanism is defined as follows:

$$\begin{aligned} &\mathtt{VU}_t = \mathtt{VU}_{t-1} - \lambda \cdot \mathtt{VP}_i, \\ &vu_{kt} = 0, \quad for \quad vu_{kt} < 0, k = 1 \ldots m \end{aligned} \tag{1.14}$$

where λ is an absorbing factor, which reflects how much information is accepted by the agent on average. $\mathtt{VU}_t$ and $\mathtt{VU}_{t-1}$ denote the agent interest vector after and before accessing information in page i, respectively.

1.4.5 Motivational Support

Motivational support is the key factor that determines agent behavior as to whether or not it will surf pages at the next level. There are three main items in Eq. 1.7. The first item, S_t, denotes the influence of past steps and initial support. ΔM denotes agent patience to navigate in order to find information; it will change along with the latency (waiting time to get information). ΔR denotes the reward of finding relevant information.

Specifically, ΔM can be characterized as follows:

$$\Delta M = -(const(M) + var(M)) \tag{1.15}$$

where $const(M)$ is constant at each step, and $var(M)$ is a variable factor that is different at each step. We may adopt different mechanisms to get $var(M)$, as listed below:

1. Agent persistence in foraging will normally decay along with the time or steps spent on foraging. Thus, we can adopt a motivation updating mechanism based on an exponential function:

$$var(M) = \alpha \cdot e^{\gamma \cdot step} \tag{1.16}$$

   where α and γ are positive constant factors.
2. Research indicates that the distribution of agent waiting time to get Web pages is a log-normal distribution [1.16]. Because the major problem that agents encounter in Web navigation is the speed to access information [1.15], it is reasonable to assume that the distribution of motivation should be a log-normal function as well:

$$\log(var(M)) \sim \mathrm{normal}(\mu, \sigma_t) \tag{1.17}$$

The reward for finding some desirable information that an agent will receive at each step is considered as being proportional to how much information that the agent accepts. In our foraging agent model, since the changing of the agent interest vector corresponds to the amount of information that the agent accepts, the reward function can be defined as follows:

$$\Delta R = \sum_{i=1}^{M} (vu_i(t-1) - vu_i(t)) \tag{1.18}$$

As the reward (ΔR) for the agent is always greater or equal to zero, it provides agents the energy to surf on the Web. The lost motivation (ΔM) in the agent is always negative, and hence it discourages the agent to surf on the Web further.

1.5 An Outline of the Algorithm

With the above definitions, we can view a user as an *information foraging agent* living in the Web space. The information contained in the Web pages will be the resource that the agent attempts to find and utilize. If the agent cannot find enough resource, it will abort its foraging behavior. Or, if it finds sufficient information resource, it will stop exploring further. All autonomous agents start to forage from a certain page, which is connected to other pages on various topics. The behavior of agent foraging can be described with the following algorithm:

Initialize network
Initialize agent profile
For each agent i
    While agent support $S < \mathtt{S_{max}}$ and $S > \mathtt{S_{min}}$
        Find the links within node k that agent locates
        Select direction (goal node) of next step
        Surf goal node
        Update agent motivational support based on Eq. 1.7
        Update agent interest vector based on Eq. 1.14
    End
    If agent support $S > \mathtt{S_{max}}$
        Agent i is satisfied
    Else
        Agent i is unsatisfied
    End
End

1.6 Experiments

Next, let us take a look at some experimental results generated from this information foraging model and compare the experimental data with real-world observed data.

After all the artificial agents end their foraging on the Web, their foraging depth/step distribution and the rank of link utility according to the link utilizable frequency are assessed based on the foraging path records. The link utilizable frequency is defined as the times a link is passed through by information foraging agents, which is also called link *clicks*.

Experiment 1 is conducted with $5,000$ artificial agents foraging according to the above-described model. The key experimental parameters are given in Table 1.1.

Parameter	Value
number of topics:	10
number of nodes:	215
Web page information vector σ_p :	0.5
agent categories:	recurrent/rational
number of agents:	5000
agent interest vector σ_u :	0.25
accessibility r :	0.7
absorbing factor λ :	0.5
$cons(M)$:	0.2
threshold T :	1
threshold T_p :	0.05
motivation updating:	1st method
α :	0.03
γ :	0.2

Table 1.1. Parameters of Experiment 1.

Figures 1.1 and 1.2 present the click distributions recorded in Experiment 1, with respect to agent categories *recurrent* and *rational*, respectively.

It should be pointed out that the lines in Figures 1.1 and 1.2 are all obtained by applying a weighted linear regression method. In the weighted linear regression, we assign each link frequency/click a weight, according to the magnitude of the link frequency/click occurrence. This implies that the link clicks with a larger agent population will have a higher weight. The objective of this method is to discount the influence of the links that are close to the home page. From Figures 1.1 and 1.2, we can realize that there do exist strong regularities in agent foraging behavior. The distribution of link clicks in agent foraging falls into a power law. This result agrees well

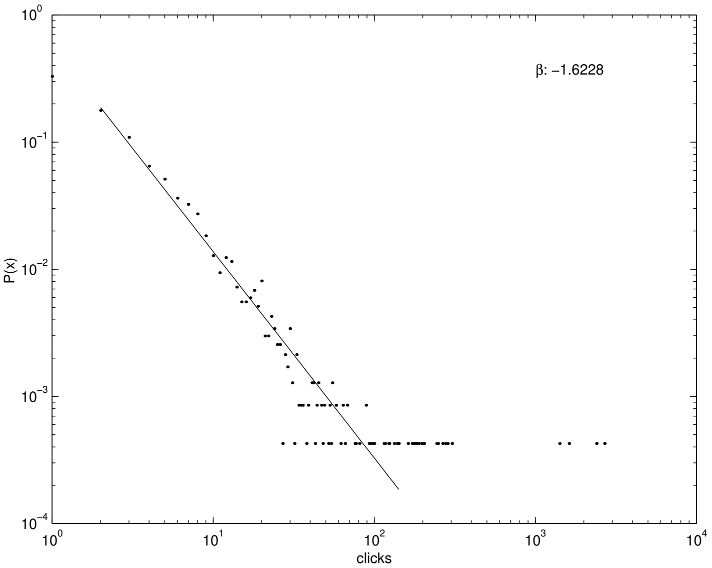


Fig. 1.1. Probability distribution of foraging intensity with *recurrent agents* obtained from Experiment 1 (*i.e.*, the distribution of link clicks – the total number of times that agents pass through a link). The tail is a power law, with power $\beta = -1.6228$. This result is obtained by applying a weighted linear regression method.

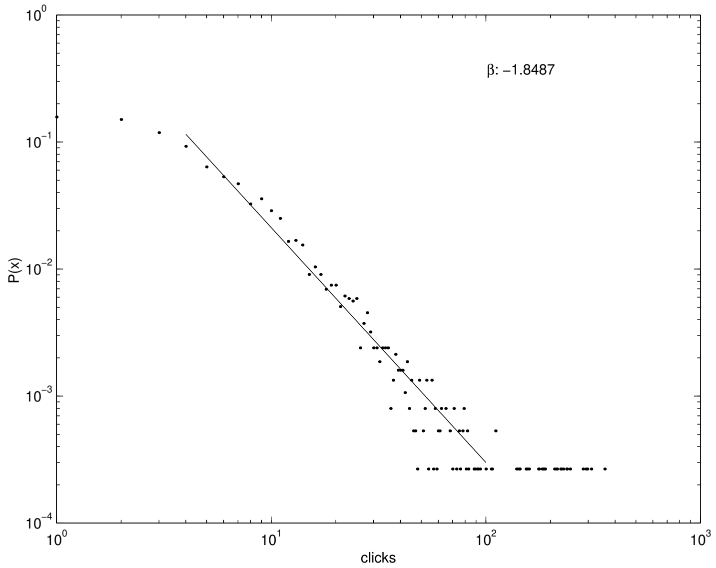


Fig. 1.2. Probability distribution of foraging intensity with *rational agents* obtained from Experiment 1 (*i.e.*, the distribution of link clicks – the total number of times that agents pass through a link). The tail is a power law, with power $\beta = -1.8487$. The result is derived based on a weighted linear regression method.

with the research findings in [1.18], which reveals that the distribution of site occurrence by popularity falls into a power law.

1.6.1 A Comparison with Real-World Log Data

In order to validate our foraging model, we also collect some real-world log data sets from the Web. One of the data sets is *Microsoft* Web log data file recorded from the Microsoft Corporate Web site. This data set contains the log of randomly-selected anonymous users visiting *www.microsoft.com* during a one-week period in February 1998. There were 32,711 users with average surfing steps equal to 3 and a total number of areas equal to 294. The total number of surfing routes is 6,336, and hence the average route number for each area is 6,336/294, which is equal to 21.55.

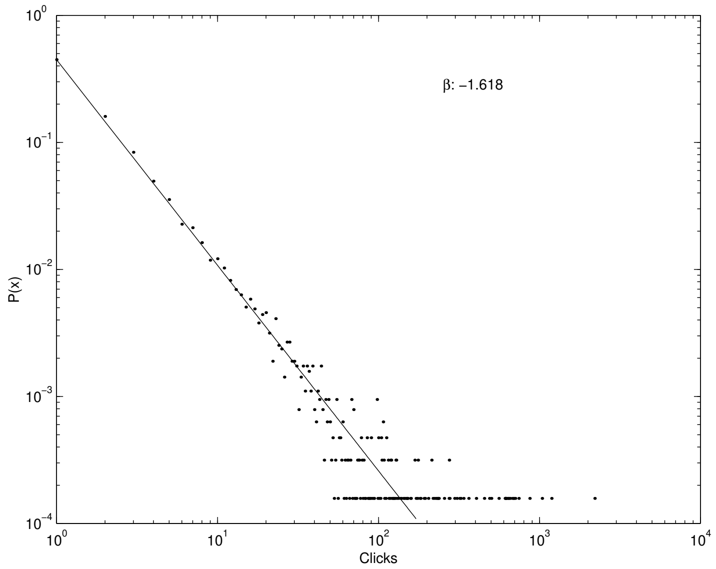


Fig. 1.3. Real-world *Microsoft* Web site log data. The distribution of link clicks agrees well with a power law, with power $\beta = -1.618$. This result is obtained by using a weighted linear regression method.

The click distribution shown in Figure 1.3 is obtained using the observed *Microsoft* Web site data set. Comparing Figures 1.1 and 1.2 with Figure 1.3, we can find that the results are very similar, in both the shape and the parameters of fitted functions. This validates that our information foraging agent model is correct, which reflects and explains the dynamics of user navigation behavior in the Web space.

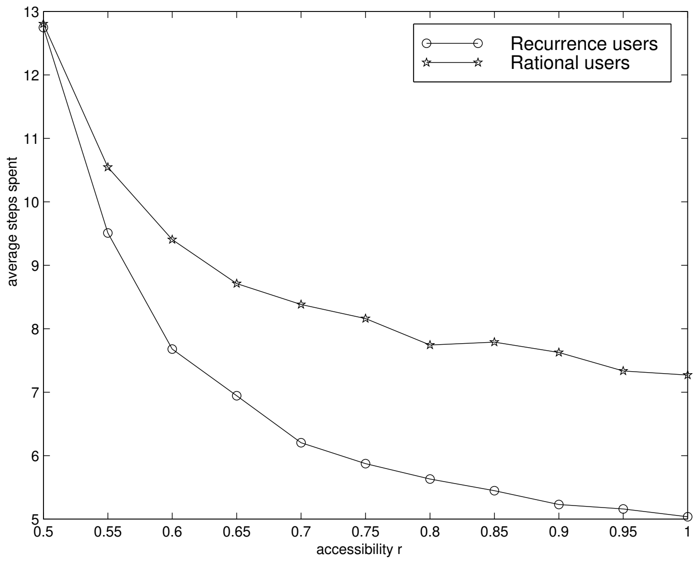
1.6.2 Further Remarks

After validating the foraging agent model, we have also designed several other experiments to verify the influence of various parameters. As mentioned earlier, increasing r will lead to the increase in link number, simultaneously the average distance of links, `ALD`, will decrease. The corresponding meaning in the actual Web structure for r is the similarity (distance) of various linked pages. Large distance usually implies that there are direct links among pages belonging to different topics and more links per page. Agents with multiple interests will jump from one topic to another easily, while agents with a single interest will feel obsessive to select a direction from many hyperlinks within a page. So we design Experiment 2 to verify the influence of various value of accessibility r. In Experiment 2, we alter r while keeping other parameters unchanged and observe the result of the experiment.

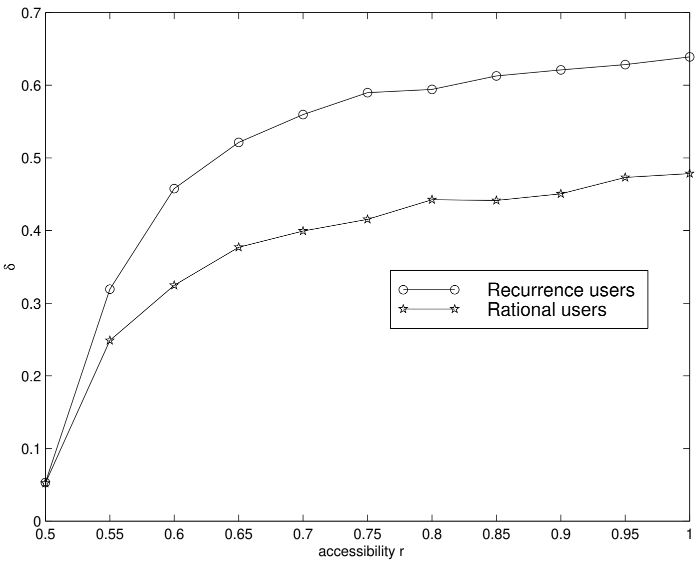
Figures 1.4 and 1.5 show the dependences of the agent average step and agent satisfaction rate on accessibility r, respectively. In Figure 1.4 *a*, the average step by rational agents is lager than that by recurrent agents. The reason is that the ability of finding useful information by rational agents is weaker than that of recurrent agents, so the rational agents must pass through more pages to satisfy themselves, their satisfaction rate will be lower than recurrent agents, as shown in Figure 1.4 *b*. The designer of the Web site usually wishes that a user could spend more time or surf deeper to get more information and satisfy the user interest, simultaneously. Figure 1.5 *a* shows the compound effect of agent foraging depth and agent satisfaction rate. From the analysis of Figure 1.5 *b*, we can obtain that in order to get the best effect, the value of accessibility r should be $0.65 \sim 0.75$. Then the average link number per node is about 19.2, which is close to the real *Microsoft* Web data, 21.5.

It is important for the designer of a Web site to identify the difference among the navigation patterns of various classes of users, Experiment 3 attempts to further reveal the navigation patterns of random agents. Because the random agents have no strong interest in the Web site, σ_u is set to 0.25. Other parameters are the same as those in Experiment 1. We have noted that the distribution of random foraging depth is very similar to the other two data sets. However, the distribution of link clicks does not agree well with the observed *Microsoft* Web data and the recurrent/rational agent data, as shown in Figure 1.6. In fact, random and recurrent agents are two extreme cases. Most Web users will be rational agents, whose prediction ability for the next level (step) is between those of random and recurrent agents.

In the above experiments, we suppose that agent interest value σ_u and Web page's information value σ_p are all constant. In fact, we do not know exactly the scales of σ_u and σ_p. Here we alter the value of σ_u and observe the effect. We focus on rational agents and adopt the second motivation updating mechanism as mentioned earlier. Accessibility r is set to 0.7. Figure 1.7 shows

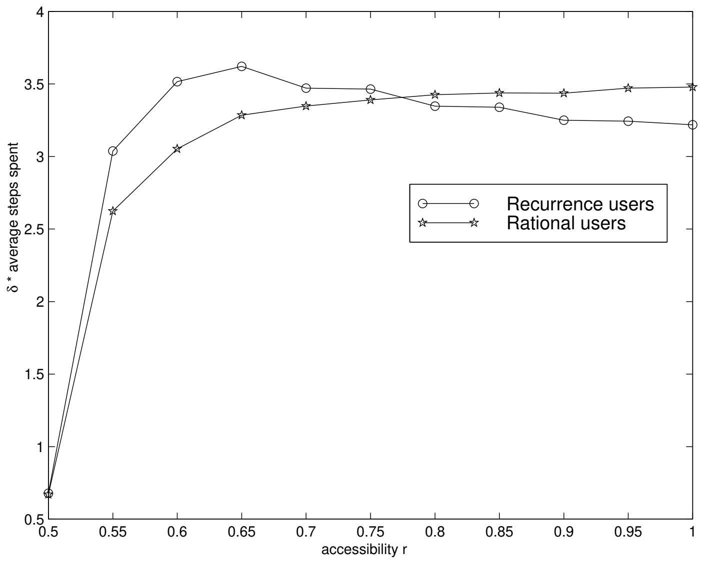


(a)

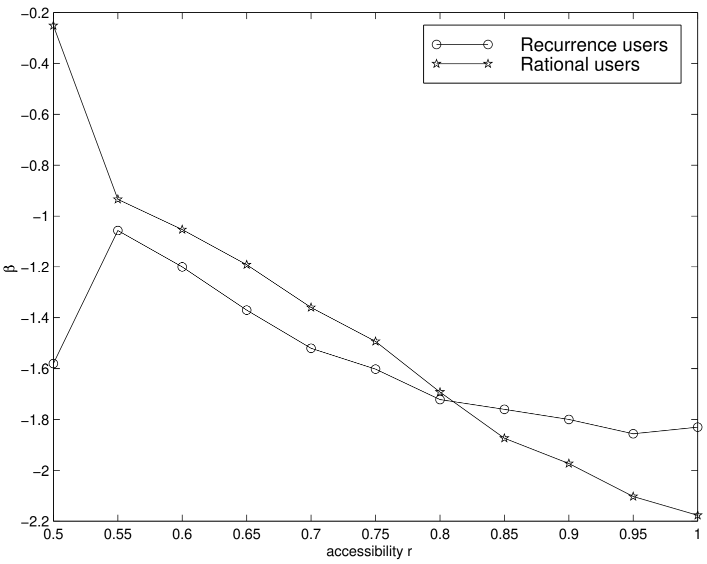


(b)

Fig. 1.4. Experiment 2: Performance with respect to the Web structure. (a): Average foraging depth against accessibility r where $\star$ denotes rational agents and $\circ$ denotes recurrent agents. (b): The satisfaction rates δ for different agent classes with respect to accessibility r.



(a)



(b)

Fig. 1.5. Experiment 2: Performance with respect to the Web structure. (a): Compound effect - Multiplication of average agent steps and satisfaction rates of different agent classes. (b): The power β of tail in distribution of link clicks against accessibility r.

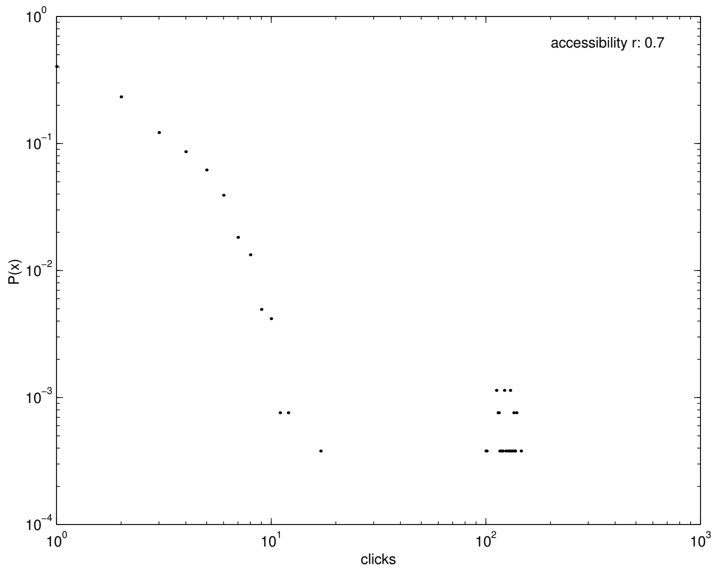
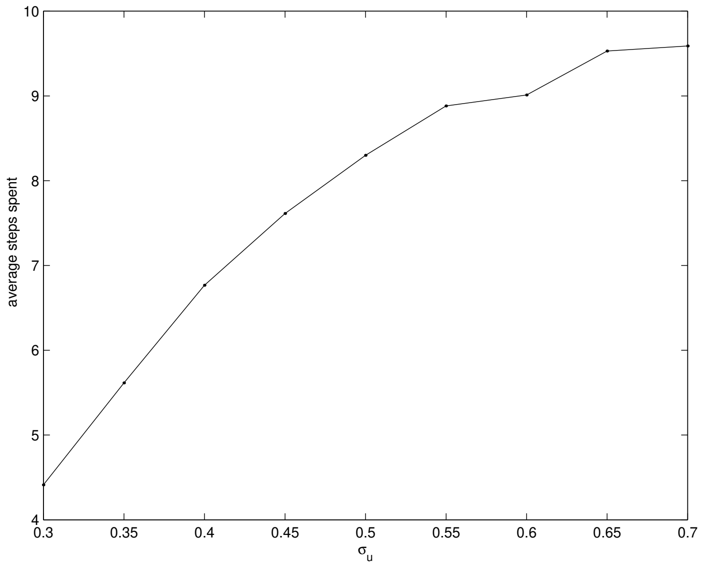


Fig. 1.6. Experiment 3: Probability distribution of foraging intensity (link click frequency) with *random agents*.

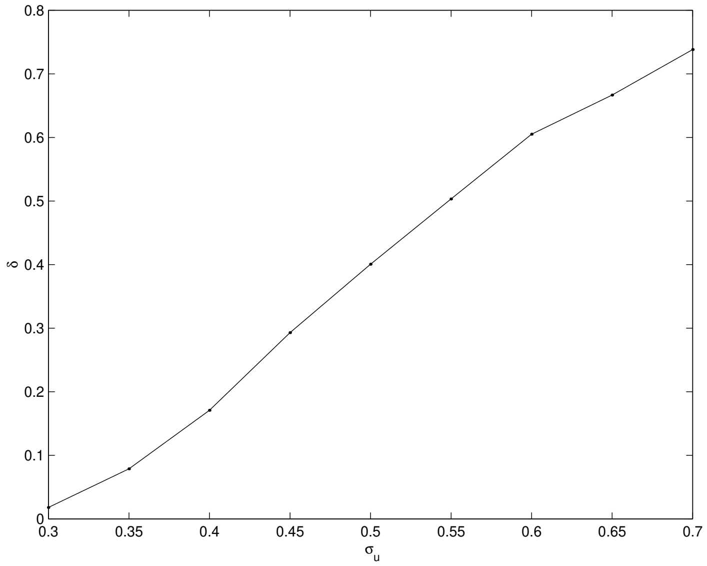
the result of this experiment. From Figure 1.7, we find that when an agent has a strong interest in the sites, it will stay there for a long time and examine more pages to find useful information. The probability of being satisfied will increase as well.

1.7 Conclusions

In Web Intelligence (WI) applications, it is a common practice to record Web log data. What remains a great challenge in Web log mining today is how to characterize the underlying user behavior from the obtained data. In this chapter, we have shown how to interpret strong regularities in Web surfing in terms of user decision-making patterns, and presented an information foraging agent-based approach to characterizing user surfing behavior. Our experimental results have successfully captured the regularities as empirically observed from a real-world Web site, and at the same time, effectively unveiled the underlying decision mechanisms in user surfing as well as the role of such mechanisms in producing empirical Web regularities. The results as well as the agent-based approach presented here are very important for our understanding of how to develop and improve Web contents and organizations as well as how to explain and induce certain user behavioral patterns.



(a)



(b)

Fig. 1.7. Influence of agent interest σ_u. (a): Average steps spent with respect to σ_u. (b): Agent satisfaction rate δ with respect to σ_u.

References

1.1 Adamic, L. A. and Huberman, B. A.: The Nature of Markets in the World Wide Web. http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.
1.2 Adamic, L. A. and Huberman, B. A. (1999): Technical comment to 'Emergence of scaling in random networks'. Science **286**, 509–512. http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.
1.3 Adar, E. and Huberman, B. A.: The Economics of Surfing. http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.
1.4 Albert, R., Jeong, H., and Barabasi, A.-L. (1999): Diameter of World-Wide Web. *Nature*, **410**:130–131, September 9.
1.5 Arlitt, M. F. and Williamson, C. L. (1996): Web Server Workload Characterization: The Search for Invariants. *Proceedings of the ACM SIGMETRICS'96 Conference*, 126–137.
1.6 Barabasi, A.-L. and Albert, R. (1999): Emergence of Scaling in Random Networks. *Science*, **286**:509–512, October 15.
1.7 Barabasi, A.-L., Albert, R., and Jeong, H. (2000): Scale-free Characteristics of Random Networks: The Topology of the World Wide Web. *Physica A*, **281**:69–77.
1.8 Barford, P., Bestavros, A., Bradley, A., and Crovella, M. (1999): Changes in Web Client Access Patterns: Characteristics and Caching Implications. *World Wide Web, Special Issue on Characterization and Performance Evaluation*, **2**:15–28.
1.9 Barford, P. and Crovella, M. (1998): Generating Representative Web Workloads for Network and Server Performance Evaluation. *Measurement and Modeling of Computer Systems: Proceedings of the ACM SIGMETRICS Conference*, 151–160.
1.10 Breslau, L., Cao, ., Fan, L., Phillips, G., and Shenker, S. (1998): Web Caching and Zipf-like Distributions: Evidence and Implications. Technical Report 1371, Computer Sciences Department, University of Wisconsin-Madison.
1.11 Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000): Graph Structure in the Web. *Proceedings of the Ninth World Wide Web Conference (WWW9)*.
1.12 Crovella, M. E. and Taqqu, M. S. (1999): Estimating the Heavy Tail Index from Scaling Properties. *Methodology and Computing in Applied Probability*, **1**(1):55–79.
1.13 Cunha, C. R., Bestavros, A., and Crovella, M. E. (1995): Characteristics of WWW Client Based Traces. Technical Report BU-CS-95-010, Computer Science Department, Boston University.
1.14 Glassman, S. (1994): A Caching Relay for the World Wide Web. *Computer Networks and ISDN Systems*, **27**(2):165–173.
1.15 GVU: WWW User Surveys. http://www.gvu.gatech.edu/user_surveys.
1.16 Helbing, D., Huberman, B. A., and Maurer, S. M. (2000): Optimizing Traffic in Virtual and Real Space.
1.17 Huberman, B. A. and Adamic, L. A.: Evolutionary Dynamics of the World Wide Web. http://www.parc.xerox.com/istl/groups/iea/www/growth.html.
1.18 Huberman, B. A. and Adamic, L. A. (1999): Growth Dynamics of the World-Wide Web. *Nature*, **410**:131, September 9.
1.19 Huberman, B. A., Pirolli, P. L., Pitkow, J. E., and Lukose, R. M. (1997): Strong Regularities in World Wide Web Surfing. Science **280**, 96–97.
1.20 Johansen, A. and Sornette, D. (1999): Download Relaxation Dynamics on the WWW Following Newspapers Publication of URL. *IMA 'Hot Topics' Workshop: Scaling Phenomena in Communication Networks*.

1.21 Levene, M., Borges, J., and Loizou, G. (2001): Zipf's Law for Web Surfers. *Knowledge and Information Systems*, **3**:120–129.
1.22 Levene, M. and Loizou, G. (1999): Computing the Entropy of User Navigation in the Web. Research Note RN/99/42, Department of Computer Science, University College London.
1.23 Lukose, R. M. and Huberman, B. A. (1998): Surfing as a Real Option. *Proceedings of the International Conference on Information and Computation Economics*, 45–51.
1.24 Maurer, S. M. and Huberman, B. A.: Competitive Dynamics of Web Sites. http://www.parc.xerox.com/spl/groups/dynamics/new.shtml.
1.25 Thatcher, A. (1999): Determining Interests and Motives in WWW Navigation. *Proceedings of the Second International Cyberspace Conference on Ergonomics (CybErg1999)*.
1.26 Yan, T. W., Jacobsen, M., Garcia-Molina, H., and Dayal, U. (1996): From User Access Patterns to Dynamic Hypertext Linking. *Proceedings of the Fifth World Wide Web Conference (WWW5)*, 1007–1014.
1.27 Zipf, G. K. (1949): *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, MA.