User Profiling for Web Search Based on Biological Fluctuation

Yuki Arase, Takahiro Hara, and Shojiro Nishio

Department of Multimedia Engineering Graduate School
of Information Science and Technology, Osaka University
1-5 Yamadaoka, Suita, Osaka 565-0871, Japan
{arase.yuki,hara,nishio}@ist.osaka-u.ac.jp

Abstract. Because of the information flood on the Web, it has become difficult to search necessary information. Although Web search engines assign authority values to Web pages and show ranked results, it is not enough to find information of interest easily, as users have to comb through reliable but out of the focus information. In this situation, personalization of Web search results is effective. To realize the personalization, a user profiling technique is essential, however, since the users' interests are not stable and are versatile, it should be flexible and tolerant to change of the environment. In this paper, we propose a user profiling method based on the model of the organisms' flexibility and environmental tolerance. We review the previous user profiling methods and discuss the adequacy of applying this model to user profiling.

Keywords: User profile, Web search, biological fluctuation.

1 Introduction

As our current life is always surrounded by Internet enabled devices, such as computers, cellular phones, PDAs and game consoles, the highly advanced information society allows us to collect information of concern far easily than the past. However, the larger the amount of information on the Web grows, the harder it becomes to find information of interest. According to a report of Google on July 2008, the number of unique URLs on the Web has already exceeded a trillion and the number of Web pages is practically uncountable. Furthermore, the number of Web pages is still rapidly growing every second.

Currently, people use Web search engines to find Web pages containing their information of interest. Most search engines use link structure of Web pages to decide authoritative Web pages, which is based on the idea that authoritative Web pages contain more reliable information than the other minors. When people query a search engine, these authoritative pages are ranked high as the search result. This criterion has been really effective to enable people to find reliable information without bothering by browsing hundreds of junk Web pages. However, the authority based ranking is not enough in the current situation of information flood, since the information on the Web has become too diverse in their semantic meaning to recommend based on

J.A. Jacko (Ed.): Human-Computer Interaction, Part III, HCII 2009, LNCS 5612, pp. 239–247, 2009.

only their reliability. As a result, people have to access many reliable but unnecessary Web pages to find pages that exactly match with their interests.

To solve this problem, it is effective to personalize Web search results based on users' interests in addition to the current authority based ranking. For this aim, user profiling is essential. However, since the users' interests are unstable and easy to change, user profiling is not easy task. It is apparent from the fact that although user profiling methods have been actively researched for decades, tracking users' versatile interests is still difficult. In this paper, we propose a novel approach to realize flexible and dynamic user profiling. We adopt a model called *Attractor selection* based on biological fluctuation to detect users' intentions. Since the Attractor selection is tolerant to the change of the assumed environment, it is suitable to model users' versatile interests.

Meanwhile, *ambient information environment* is a recent hot topic, where surrounding computers and embedded sensors in the environment detect users' situations and provide functions to satisfy users' needs without users' explicit declaration. In the precedent ubiquitous environment, users do access computers to satisfy their requirements. However, in the ambient environment, the environment does make actions to satisfy users' needs. The concept of ambient environment can be applicable to various fields, including user interface on which we are working. We define ambient user interface as interfaces that detect users' intentions and provide information according to them. We regard our personalization method as a realization of ambient interface for Web search.

The reminder of this paper is organized as follows. In section 2, we introduce previous user profiling methods and discuss the differences from our approach. In section 3, we briefly introduce the biological Attractor selection model. In section 4, we propose our user profiling method based on Attractor selection, and discuss its potential to solve problems that previous user profiling methods could not settle. Then, in section 5, we conclude this paper and describe our future work.

2 Related Work

There have been two directions for user profiling, as one makes use of explicit feedback from users, and the other does implicit feedback.

A popular method for the former approach is asking users to input their interest, which is adopted by some portal and news sites. Another method is to ask users to assign a score to their browsed Web pages according to the strength of their interest to contents of the pages. As represented by News Dude [1], a user can specify i) whether they think the content is interesting or not, ii) if they would like to see more similar information, or iii) if they have already seen the information previously.

An advantage of this approach is that extracted user profiles tend to be reliable, since users themselves input their interests. A disadvantage is that they trouble users to input their interests, and more, users have to change their profiles each time when their interests change.

The latter approach uses the users' browsing behaviors to extract user profiles. SUGGEST [2] adopts a two-level architecture composed by an offline creation of historical knowledge and an online engine that understands user's behavior. As the requests arrive at this system module, it incrementally updates a graph representation of the Web site based on the active user sessions and classifies the active session using a graph partitioning algorithm. Gasparetti et al. [3] proposed an algorithm with which the system could identify the users' interests by exploiting Web browsing histories. Claypool et al. [4] investigated different kinds of user behaviors, such as scrolling, mouse clicks, and time on page for extracting user profiles.

These methods are based on *learning* user interests, and thus, a large amount of training data is needed and they seem to take long time to converge to reasonable profiles.

Billsus et al. [1] proposed that users' interests can be classified into two, as long-term and short term interests. The long-term interests can be regarded as the users' intrinsic interests, and thus, seem to be stable over time. Therefore, they are easier to extract by directly asking users and using learning based methods. On the other hand, the short-term interests can be regarded as interim, reflecting users' current interests. This feature results in difficulty of tracking change of profiles because of their versatile nature in a short period. Especially for Web search, users' interests can be classified to the short-term interests, since users usually search information which they need, concern, and get interested in, at that time. Therefore, a user profiling method of flexible and tolerant to environmental change is suitable. For this aim, we adopt the Attractor selection mechanism, which is based on the fluctuation of organisms and realizes flexible and environmental tolerant solutions, to track change of users' interests on Web search.

3 Attractor Selection

In this section we outline the principle of Attractor selection, which is a key component in our method. The original model for adaptive response by Attractor selection is given by Kashiwagi et al. [5].

Attractor selection defines each possible situation as *attractor*, and evaluates the current situation to select one of better attracters in a dynamic environment. The goodness of the attractor is estimated by the *activity* value. While the activity is high, the system keeps staying the current attracter. On the other hand, when the situation changes and the activity gets low, the system performs a random walk to find a more suitable attractor. Because of the random walk, the system performs fluctuation.

We can basically outline the attractor selection model as follows. Using a set of differential equations, we describe the dynamics of an M-dimensional system. Each differential equation has a stochastic influence from an inherent Gaussian noise term. Additionally, we introduce the activity α which changes the influences from the noise terms. For example, if α comes closer to 1, the system behaves rather deterministic and converges to attractor states defined by the structure of the differential equations. On the other hand, if α comes closer to 0, the noise term dominates the behavior of the system and essentially a random walk is performed. When the input values (nutrients) require the system to react to the modified environment conditions, activity α changes accordingly causing the system to search for a more suitable state. This can also involve that α causes the previously stable attractor to become unstable.

The random walk phase can be viewed as a random search for a new solution state and when it is found, α decreases and the system settles in this solution. This behavior is similar to the well known simulated annealing [6] optimization method, with the main difference that the temperature is not only cooled down, but also increased again when the environment changes.

The biological model describes two mutually inhibitory operons where m_1 and m_2 are the concentrations of the mRNA that react to certain changes of nutrient in a cell. The basic functional behavior is described by a system of differential equations, as the following equations show.

$$\frac{dm_1}{dt} = \frac{syn(\alpha)}{1+m_2^2} - \deg(\alpha)\, m_1 + \eta_1$$

$$\frac{dm_2}{dt} = \frac{syn(\alpha)}{1+m_1^2} - deg(\alpha)\, m_2 + \eta_2$$

The functions *syn(α)* and *deg(α)* are the rate coefficients of mRNA synthesis and degradation, respectively. They are both functions of α, which represents cell activity or vigor. The terms η_i are independent white noise inherent in gene expression. The dynamic behavior of the activity α is given as follows.

$$\frac{d\alpha}{dt} = \frac{prod}{Newtrient} - cons\ \alpha$$

$$Newtrient = \prod_{i=1}^{M} \left[\left(\frac{nutr\_thread_i}{m_i + nutrient_i} \right)^{n_i} + 1 \right]$$

Here, *prod* and *cons* are the rate coefficients of the production and consumption of α. The term $nutrient_i$ represents the external supplementation of nutrient i and $nutr\_thread_i$ and n_i are the threshold of the nutrient to the production of α and the sensitivity of nutrient i, respectively.

A crucial issue is the definition of the proper *syn(α)* and *deg(α)* functions. To have two different solutions, the ratio between *syn(α)* and *deg(α)* must be greater than 2 when there is a lack of one of the nutrients. When *syn(α) / deg(α)* = 2, there is only a single solution at $m_1 = m_2 = 1$. The functions *syn(α)* and *deg(α)* as given in [5] are as follows.

$$syn(\alpha) = \frac{6\alpha}{2+\alpha}$$

$$\deg(\alpha) = \alpha$$

The system reacts to changes in the environment in such a way that when it lacks a certain nutrient i, it compensates for this loss by increasing the corresponding m_i value. This is done by modifying the influence of the random term η_i through α, as Figure 1 shows. When α is near 1, the equation system operates in a deterministic fashion. However, when α approaches 0, the system is dominated by the random terms η_i and it performs a random walk.

In Figure 1, an example is given over 20000 time steps. We can see the following behavior. When both m_i values are equal, the activity is highest and α = 1. As soon as there is a lack of the first nutrient ($2000 \leq t < 8000$), m_i compensates this by increasing its level. When both nutrient terms are fully available again ($8000 < t \leq 10000$), the activity α becomes 1 again. An interesting feature of this method can be observed between $10000 < t < 13000$. Here, the random walk causes the system to search for a

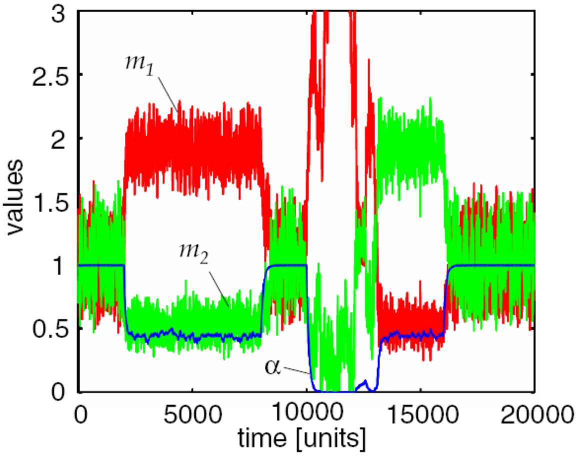

Fig. 1. Biological attractor selection model

new solution, however, it first follows a wrong "direction" causing α to become nearly 0 and the noise influence is highest. As soon as the system approaches the direction toward the correct solution again, α recovers and the system gets stable again. Such phases may always occur in the random search phase.

As we showed above, Attractor selection cannot always find the best answer. Instead of finding that, it tries to find a better attractor with quite a simple strategy. By this mechanism, Attractor selection ensures robustness in exchange for efficiency. Organisms can survive in such an unstable and dynamically changeable environment in the nature owing to inhering Attractor selection.

4 User Profiling Method Based on Attractor Selection

As we discussed in Section 2, the problem of previous user profiling methods is that they take long time to follow the change of users' interests. As Billsus et al. [1] proposed, users' interests can be classified into two, as long-term and short term interests. The long-term interests are easier to extract by directly asking users and using learning based methods owing to their stable nature. On the other hand, the short-term interests are difficult to track because of their versatile nature in a short period.

To detect the short-term interests, we adopt the Attractor selection scheme, which is suitable for finding solutions in a dynamically changeable environment.

4.1 Design of User Profiling Method

We model users' interests as attractors and their change as environmental change in the Attractor selection scheme. We define a user's profile as a ranking of pre-defined thirteen topics and detect the ranking using the Attractor selection scheme.

4.1.1 Definition of User's Profile

According to the definition of categories of Web sites at YAHOO! Japan [7], we decided the following thirteen topics as the users' interest candidates.

1) News
2) Entertainment
3) Sports
4) Art
5) Health
6) Politics
7) Economics
8) Life
9) Computer
10) Education
11) Technology
12) Local
13) Others

A user profile is a ranking of these thirteen topics, as *User profile* = {(1| *topic*$_1$),..., (13| *topic*$_{13}$)}, where *(rank*$_k$ | *topic*$_k$*)* represents that *rank*$_k$ is the rank of the topic and *topic*$_k$ is one of the candidate interest topics. For example, a profile of {(1| Technology), (2| computer),…,(13| Education)} means that this user is most interested in technology and computer related Web pages, while not interested in pages relating educational information.

Here, we assume that categories of Web pages are given. Since our definition of users' interest topics based on the Web sites' categories, we can expect to use Web sites' categories assigned by portal sites. Additionally, since there are many previous works conducting automatic Web page categorization, it is also possible to adopt these methods to categorize Web pages as a pre-processing step.

4.1.2 Definition of Activity

In Attractor selection, the value of activity α represents the goodness of the current attractor. In our case, α represents how the current user profile matches with the user's real interests. To evaluate it, we adopt the essence of DCG (discounted cumulative gain) to evaluate the current profile.

DCG is a measure of effectiveness of a Web search engine, often used in information retrieval. The premise of DCG is that highly relevant items appearing lower in a search result list should be penalized as the graded relevance value is reduced logarithmically proportional to the rank of the result. The DCG accumulated at a particular rank *p* is defined as follows:

$$DCG_p = rel_1 + \sum\nolimits_{i=2}^{p} \frac{rel_i}{log_2 i}$$

Where rel_i is a graded relevance of the result at rank *i*.

In our case, we cannot obtain explicitly graded values for each interest topic. However, it is reasonable to assume that users frequently browse Web pages of topics of interest and their browsing times will be longer compared to other topics. Therefore, we define rel_i using users' cumulative browsing times of topics when they browsed a certain number of Web pages.

The desired behavior of α is summarized as follows. If we have no information about a user, the candidate topics should be evaluated uniformly. A low α means that the current profile does not match with users' interests and a new one should be detected. We should keep the value of α as $0.0 < \alpha < 1.0$. The larger α is, the detected profile better matches with the user's interests.

As a whole, after browsing N pages, activity α is determined as follows.

$$\frac{d\alpha}{dt} = \delta\left[\left(\frac{E_{min}}{E}\right)^{\lambda} - \alpha\right] \tag{1}$$

$$E = \frac{f(n)}{DCG_p{}'} \tag{2}$$

$$DCG'_p = rel'_1 + \sum_{i=2}^{p} \frac{rel'_i}{log_2 i}$$

$$rel'_k = \sum_{j=1}^{N} browsing\_time_j \ (pagej \in topic\ k)$$

Here, δ and λ are constant parameters to adjust the adaption of α, E_{min}/E is the fraction of the evaluation value of previously found best matching profile over that of the current profile, *f(n)* is the normalization factor, *p* is the rank of a topic which should be considered. For example, if *p*=3, the method evaluates only the top three interests of the user profile instead of examining the whole profile, which results in stressing the topics of the best interest.

4.1.3 Calculation of Interest Ranking

As we showed in Section 0, the original form of Attractor selection is two dimensional. Leibnitz et al. extended the form to multi dimensional [8]. We adopt the multi dimensional form since we have to deal with the thirteen topic candidates.

For each topic, we decide its weight using Attractor selection, and rank the topics according to their weights. Specifically, we use the following multi dimensional form.

$$\frac{d}{dt}W_{i,j} = \frac{syn(\alpha)}{i + W_{max,j}^2 - W_{i,j}^2} - \deg(\alpha)\, W_{i,j} + \eta_{i,j} \tag{3}$$

$$syn(\alpha) = \alpha[\beta\alpha^{\gamma} + \phi^*]$$

$$\deg(\alpha) = \alpha$$

$$\phi^* = \frac{1}{\sqrt{2}}$$

Here, $W_{i,j}$ is the weight of topic i when being assigned rank j, $\eta_{i,j}$is a white noise, β and γ are constant parameters that adjust the effect of the noise term and activity α, respectively.

4.1.4 Flow of User Profiling Detection

We can now summarize the basic algorithm for detecting the user profile when a user browsed *N* Web pages:

1. Calculate activity α based on Equation (1).
2. Initialize the set of topics U_t that are already determined the rank as $U_t = \phi$.
3. Conduct the following process for each rank $j = 1, 2, \ldots, 13$:

   a. For each topic $i = 1, 2, \dots, 13$ $(\notin U_t)$, calculate the weight of the each topic $W_{i,j}$ based on Equation (3).
   b. Set max_i as i of the maxim value of $W_{i,j}$ $(i = 1, 2, \dots, 13(\notin U_t))$.
   c. Set j as the rank of the topic max_i and add the topic into U_t.

4. Calculate the feedback of the decided user profile based on Equation (2).
5. Update E_{min} if necessary.

4.2 Discussion

The Attractor selection scheme has been applied to some research fields [8][9][10]. The first application is for multi-path routing in overlay networks conducted by Leibnitz et al [8]. They showed that their method based on Attractor selection is noise-tolerant and capable of operating in a very robust manner under changing environment conditions. The authors also applied the Attractor selection scheme to routing problem in a mobile ad-hoc/sensor network environment [9]. They proved that their Attractor selection based method can operate entirely in a self-adaptive manner and that it can easily compensate for sudden changes in the topology of the network. On the other hand, Kitajima et al. applied the Attractor selection scheme to set parameters for filtering contents on data broadcasting services [10]. They assume an environment that broadcasting service providers broadcast enormous and various data to users and user clients have to filter out unnecessary data for users. By using the Attractor selection scheme to decide the order of filters, they can reduce the time for filtering in such a dynamic environment.

These three previous works show the robustness of the Attractor selection scheme to the change of the environment. In addition, another advantage of the Attractor selection scheme is that it operates without explicit rules and is simply implemented by numerical evaluation of the differential equations. In our case, users' interests are not stable in nature and versatile in a quite dynamic manner. Therefore, we can expect that our Attractor selection based method successfully track change of users' interests in a self-adaptive manner. Furthermore, because of its simplicity of implementation and requiring nothing to store users' histories to learn their profiles, we can implement and distribute our method as a light-weight plug-in to Web browsers, which means that users can receive a benefit of personalization very easily. It also has the advantage that our method is free from violating users' privacies, since it takes into account the current Web page the user browsed and does not need to store their browsing histories and behaviors.

5 Conclusion and Future Work

In this paper, we reviewed the works of user profiling and discussed their problems. Since most of the user profiling methods requires considerable amount of users' browsing histories as well as the information of their behaviors on the Web pages, it seems difficult to converge to the reasonable user profiles in a practical time, as the users' interests frequently change.

We briefly introduced the Attractor selection scheme that models the fluctuation inhering in organisms, and proposed the Attractor selection based method for user profiling in Web browsing. We defined a user's profile as the ranking of pre-defined topics and decide the ranking using the Attractor selection scheme.

As future work, we implement a practical application and conduct user experiments to examine the effectiveness of our method. Since the users' interests might be quite unstable and versatile, we should confirm the how quickly our method can converge to each attractor.

Acknowledgement

This research was partially supported by "Global COE (Centers of Excellence) Program" and Grant-in-Aid for Scientific Research on Priority Areas (18049050) of the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

1. Billsus, D., Pazzani, M.J.: A Personal News Agent that Talks, Learns and Explains. In: The Third Annual Conference on Autonomous Agents, Seattle, pp. 268–275 (1999)
2. Baraglia, R., Silvestri, F.: Dynamic Personalization of Web Sites Without User Intervention. Communication of the ACM 50(2), 63–67 (2007)
3. Gasparetti, F., Micarelli, A.: Exploiting Web Browsing Histories to Identify User Needs. In: International Conference on Intelligent User Interfaces (IUI 2007), Hawaii, pp. 28–31 (2007)
4. Claypool, M., Le, P., Waseda, M., Brown, D.: Implicit Interest Indicators. In: The Sixth International Conference on Intelligent User Interfaces (IUI 2001), USA, pp. 33–40 (2001)
5. Kashiwagi, A., Urabe, I., Kaneko, K., Yomo, T.: Adaptive Response of a Gene Network to Environmental Changes by Fitness-Induced Attractor Selection. PLos ONE 1(1), e49 (2006)
6. Aarts, E., Korst, J.: Simulated Annealing and Boltzmann Machines. Wiley, New York (1989)
7. Yahoo! Japan, `http://www.yahoo.co.jp/`
8. Leibnitz, K., Wakamiya, N., Murata, M.: Resilient Multi-Path Routing Based on a Biological Attractor-Selection Scheme. In: Ijspeert, A.J., Masuzawa, T., Kusumoto, S. (eds.) BioADIT 2006. LNCS, vol. 3853, pp. 48–63. Springer, Heidelberg (2006)
9. Leibnitz, K., Wakamiya, N., Murata, M.: Self-Adaptive Ad-Hoc/Sensor Network Routing with Attractor-Selection. In: IEEE GLOBECOM, San Francisco, pp. 1–5 (2006)
10. Kitajima, S., Hara, T., Terada, T., Nishio, S.: Filtering Order Adaptation Based on Attractor Selection for Data Broadcasting System. In: International Conference on Complex, Intelligent and Software Intensive Systems (CISIS 2009), Fukuoka (2009)