Swarm Intelligent Surfing in the Web

Jie Wu and Karl Aberer

School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPF), Lausanne
1015 Lausanne, Switzerland
{jie.wu, karl.aberer}@epfl.ch

Abstract. Traditional ranking models used in Web search engines rely on a static snapshot of the Web graph, basically the link structure of the Web documents. However, visitors' browsing activities indicate the importance of a document. In the traditional static models, the information on document importance conveyed by interactive browsing is neglected. The nowadays Web server/surfer model lacks the ability to take advantage of user interaction for document ranking. We enhance the ordinary Web server/surfer model with a mechanism inspired by swarm intelligence to make it possible for the Web servers to interact with Web surfers and thus obtain a proper local ranking of Web documents. The proof-of-concept implementation of our idea demonstrates the potential of our model. The mechanism can be used directly in deployed Web servers which enable on-the-fly creation of rankings for Web documents local to a Web site. The local rankings can also be used[1] as input for the generation of global Web rankings in a decentralized way.

Keywords: swarm intelligence, information retrieval, search engine, ranking, filtering

1 Introduction

Information seeking is social behavior, and it is shaped by factors that operate at the cognitive, affective, and situational levels [5]. No matter whether the seeking occurs on the Web or in a traditional library, it is always guided by perceptions of information quality, accessibility, conditions and requirements of a particular situation. Ranking according to specific aspects of the documents, e.g. the degree of importance, and the profiles and requirements of the users is thus extremely helpful in the process of seeking information.

1.1 Motivation

Traditional ranking models rely on a static snapshot of the Web graph, basically the link structure among the Web pages. While the ranking models of Web search rely on the static snapshot of the Web graph, including the structure and the

[1] Please refer to our companion paper [3] on our Ranking Algebra.

J.M. Cueva Lovelle et al. (Eds.): ICWE 2003, LNCS 2722, pp. 431–440, 2003.

document content, the information conveyed in the dynamic interactive browsing activities is simply neglected and wasted. However, the relevance judgements collected at browsing-time by Web surfers should be stronger than those obtained from static Web links authored by single Web authors. Basically, the current research on the Web information seeking has the following problems:

1. Methodology: the relationship among the Web documents is studied in a rather static way. The point is the Web is actually a social system with the clients roaming in it and interacting with the Web servers. Thus browsing as natural relevance feedback is ignored. It should be treated in a dynamic way.
2. Model: given the dynamic nature of the Web social system, the existing static document model does not work for high performance information seeking. Thus the relevance implied by the dynamic surfing sessions is not used to modify the weight of the information contained in a document. A new dynamic and intelligent model has to be developed to describe the dynamic interactions.
3. Technical limitation: the nowadays Web server/surfer model lacks an interaction mechanism between the server and the surfer. The Web server plays no active role in the surfer's decision for choosing highly ranked pages for continuing surfing. Although research was done on customized Web proxies, they mainly deal with the collection of user profiles, accessing patterns, etc.. We expect the Web servers to do more than passively answering file requests in helping surfers identify which links or documents are more important and hence valuable in Web browsing.

Thus, we enhance the ordinary Web server/surfer interaction model with a mechanism inspired by swarm intelligence [1] to make it possible for the Web servers to participate actively in the interactive surfing activities and in obtaining a proper ranking of Web documents according to general importance for them. This is done by adding a swarm intelligent module to the Web server architecture to obtain the necessary information and do corresponding analysis on the fly. Thus our approach not only makes use of the traditional criteria such as link structure but also captures the ranking information provided through the dynamic interactive Web browsing activities of the Web surfers. We believe that it is worthwhile to consider a dynamic server/surfer model instead of one of static links for the following reasons:

1. To verify that the group intelligence in a self-organized system at the collective level is higher than that of any individual.
2. To obtain more precise rankings by exploiting surfer's knowledge embedded in the surfing sessions. In short, we want to infer local ranking information from the interactive surfing activities, which is not possible with current Web server implementation and modern search engines.

We first implemented a little game of *quest for treasure* to show how the idea works in the Web in analogy to the ants' hunting for food. Then we developed

a module for the Apache Web server version 2. Preliminary experiments show that it is a very interesting way to get the local document ranking, and this proof-of-concept implementation of our idea demonstrates the potential of our new model which can be extended to a distributed architecture in the future.

1.2 Our Contributions

We particularly study a information seeking problem in the Web, i.e. how to intelligently rank Web documents according to the general importance of them. Here are our main contributions:

1. Raising the idea of using dynamic importance information embedded in the interaction between a Web server and surfing users. We argue that the traditional analysis of static Web links does not satisfy the requirements of precise ranking of ever-changing Web documents;
2. Providing a model inspired by Swarm Intelligence to analyze the emergent behavior of the Web activities. Individual surfing activities can collectively demonstrate some kind of group intelligence that is higher than any single one in a self-organized fashion;
3. Applying the idea in the computation of Web documents ranking for search engines. The emergent result of the interactions may also lead to more recent hence valuable ranking than the results produced by traditional algorithms relying on static link analysis;
4. Having developed a module for the successful Apache Web server to realize our idea. The module can be loaded into the popular Web server after simple configuration. After this module is enabled, the Web server can automatically collect and analyze the dynamic interactions between the server and the surfers and generate on-the-fly document rankings on the Web.

2 Web Surfing and Swarm Intelligence

In many situations human beings are not so different from insects. Some aspects in the analogy are: both the food and information resources are unevenly distributed in the nature or on the Web; the resource procurement is dynamically characterized by all sorts of uncertainty and fluctuation; and both searchers have limited time and chance to explore and exploit the resources. Only those who adopt through learning the strategies that maximize their harvest rates can become winners. Thus, in order to forage information on the Web people may consider making use of the collective intelligence also existing in social insects.

2.1 What's Swarm Intelligence All about?

The social insect metaphor has been studied to solve problems for a long time and in many fields [1]. People find that the approaches are good for dealing with problems distributed, direct or indirect interactions among relatively simple agents, flexibility and robustness. For the purpose of clarification, we include

a very short summary of swarm intelligence here. More details of Swarm Intelligence can be found in [1] and the website [2].

The scenario is that some ants are looking for food. Suddenly an obstacle appears and interrupts the ants' moving. Now an important step occurs. The ants choose randomly in front of the obstacle to turn right or left at the beginning, but trails of biological pheromones are left on the road when they move on to the destination or go back to the nest. The more ants pass by a road, the more biological trails are left on it. On the other hand, the higher pheromone density a road has, the higher possibility it is chosen by the ants in the future because these insects instinctively follow the pheromones. With more and more ants' hitting the right spot and successfully getting the food back to the nest, on the shortest path from the nest to the food the pheromones will accumulate with the highest density. Thus the shortest optimal path is found finally, and actually formed without any instructions, in a totally self-organizing way.

Swarm intelligence has been reported to have applications in many fields, such as combinatorial optimization, communication networks, robotics, etc.. As far as we know, nobody else or other research groups have tried to apply this idea for Web surfing to obtain ranking of documents for the purpose of information retrieval.

2.2 Analogies of Insects and Web Servers

Let us make the analogies between our model of Web surfing and the biological society more precise. We call the counterpart of biological pheromone in the Web *Web pheromone*.

Social Insects Society	Web System
Ant	Web client/surfer
Food	information on Web pages
Hunting for food	Web browsing
Biological pheromone trail	recorded Web pheromone maintained by the Web server
Interaction	request from the client and reply from the server
Pheromone density	popularity/importance of a Web page

It is easy to imagine that an important page, which will potentially be ranked by ranking algorithms, needs continuous repeated hits to maintain its level of importance. In a decentralized system, self-organization relies on the following four important components:

1. positive feedback (amplification) recruitment to a food source, high probability that many accesses mean good quality of information;
2. negative feedback (counterbalances positive feedback) food source exhaustion, satiation, crowding at food source;
3. fluctuations, random walks, errors;
4. multiple interactions trail-following events interact with trail-laying actions, moreover individuals should make use of their own activities as well as of others' activities.

With respect to Web surfing we can establish these components as follows:

	Source	Provokes
Positive feedback	Results from search engines link recommendations advertisements ranking system	Frequent visits, long term value exhaustive page exploration recommendations to others crowding, server slowdown
Negative feedback	unreadable content overcrowding, slow server no interest	abandon server relief no recommendations
Fluctuations	curiosity surf for fun individual intelligence	discover new pages discover alternative pages optimized paths
Multiple interactions	collectively used information each click leaves trails follow others	adds to collective knowledge profit of other users feedback give and get at the same time

The Web surfers here are the natural agents in a self-organizing system. Surely surfers are not non-intelligent, but as human have only really limited insight on the Web and most of the time can only follow the hyper links created by someone else without any knowledge of the structure of the Web graph, we can safely take the surfers as the primitive agents that abide by simple operation rules.

Furthermore, the Web server here is not only a passive listener to the requests for Web documents, but also an active participant of the self-organizing system by assuming the role of an arbiter who assures the rules are carried on during the interactive interactions between the requesting visitors and the requested Web documents which form together the ecological environment for the self-organizing system.

3 A Swarm Intelligent Web Server

Surfers with certain intelligence usually lack the knowledge of how to wander efficiently around the Web to satisfy their information needs. In the sense of information foraging, we assume the content of Web pages being the persistent prey that the surfers want to forage and every time a surfer visits a page she leaves some Web pheromone, which is maintained by the environment, i.e. the Web server, to it. Upcoming visits to the same page will enhance the density of its Web pheromone if its content is appreciated by the surfers. If the page is not so useful for the surfers, people will certainly not come back and thus the density of the page's Web pheromone will decrease by means of natural evaporation. By integrating a such dynamic self-organizing mechanism into a Web server, we make the server swarm intelligent. We call a Web server that is equipped with our Swarm Intelligent module a *Swarm Intelligent* Web server.

3.1 Interaction Rules

Interaction is one of the most important features in our dynamic server/surfer model. It is different from the traditional collaborative filtering techniques where a user has to, if not forced to, manually choose or input a feedback value usually in numerical range, e.g. 1 to 10, so as to let the server know how the Web page is evaluated by the visitor. On the contrary, in our model, all the feedbacks are collected automatically without any artificial intervention.

We identify the following feedbacks from the visitors to the server. They are kept in the pheromone data structure.

1. Visitors' evaluation over the general importance of a Web documents. This is maintained through the dynamic change of the density variable *phero* of the Web pheromone tuple after the interactions between the intelligent server and surfers.
2. Time of visit. Visits are made at different arrival time, in different frequencies, from different referrer, with different sojourn length. This information can be inferred from a time stamp variable kept as part of the Web pheromone. All these components of a visit pattern bear information of a visitor's evaluation of the visited page.

Worthy of notice, we do not perform any user tracking. We are not interested in the behavior of individual users, rather in the collective phenomena of this social system which is realized by our dynamic server/surfer model.

We also identify the following feedbacks from the server to the visitors:

1. Content enhancement: colorize (or apply other user interface techniques here) according to the pheromone density of the link. Direct real time feedback inside the Web page helps the surfers to make optimal decisions while looking for specific content.
2. Dynamic page headers can be inserted at the top of each served page on the server, for example, 10 pages of a site that have the highest pheromone density.
3. Dynamic link suggestion: pages with similar amount of pheromone; the pages that visitors who bring much pheromone here also go to; etc.

3.2 Pheromone Representation

The most important attribute of Web pheromone is the density which records the trails, and hence the endorsement of Web surfers on the general importance of a document. As the density is always changing (because of evaporation, accumulation, spreading, etc.), a time stamp of the last access is kept together with it to enable the next round of update computation. Thus we store both the density and the time stamp for each document to keep track of its Web pheromone information. Here is our model:

Given the document set D of a Website, the Web pheromones associated with the documents are defined as:

$$p : D \to V \times T$$

where $V = [0, \infty)$, is the real value representing the density of the Web pheromone; and $T = [0, \infty)$, is the last access time, i.e., the update time of the density variable. For convenience, we will use $p_V(d)$ and $p_T(d)$ to denote respectively the density value and the last access time of the Web pheromone associated with document d.

The rules of pheromone update are as follows:

1. Positive update increasing the density. This may happen due to two kinds of interactions in our model: firstly, a surfer comes to the page and brings more Web pheromone, we call this effect *accumulation*; secondly, the Web pheromone of other Web pages is diffused and arrives here, we call this effect *spreading*.
2. Negative update decreasing the density. This happens due to the natural diffusion of the Web pheromone. We call this effect *evaporation*.

Accordingly, the update strategy is composed of three regulations. Worthy of notice is that variations of the policies are possible, and are subject of further study. However we expect that the results would be quite robust as the situation shown in the study of IR models.

Pheromone Accumulation. The Web pheromone changes when a surfer changes in his surfing session from one document to another by following the link between them. In this case the newly accessed page's Web pheromone is increased.

In the current initial experiments we increase the Web pheromone of a page $d \in D$ accessed at time $p_T(d) := t$ by 1,

$$p_V(d) := p_V(d) + 1$$

We might change the policy of how much to increase in future experiments, e.g. based on the origin of the access (if the referrer is an important Web site to combine the concept of Hub and Authority sites), or the user profile (the weight assigned to a visitor according to her trust records). This leaves space for further study.

Pheromone Evaporation. Evaporation is the mechanism for realizing the negative feedback in the environment.

The evaporation property of pheromone information is very important. The simplest but yet most powerful approach to model it, is the use of a half-life time function. A given half-life time value determines the amount of time during which the pheromone density is halved. These parameters can be set heuristically for example to the mean time gap between two hits to a Web page.

Evaporation leads to the following update of pheromone:

$$p_V(d) := p_V(d) \cdot (\frac{1}{2})^{\frac{t - p_T(d)}{\delta}}$$

where $p_T(d) := t$ is the time of access, δ is the half-life time value. This update is done before the accumulation is computed.

For example, suppose the $\delta = 1\ day = 24\ hours$ and at the time $p_T(d) = Jan.\ 25,\ 2003,\ 12:40, p_V(d) = 14.0452$. If there is no new visitor in the following 30 hours, then at $t = Jan.\ 26,\ 2003,\ 18:40$, the new $p_V(d)$ is diminished drastically to 5.9053.

Pheromone Spreading. As described in Kleinberg's paper on hubs and authorities [8], an important (authoritative) page is usually pointed by many other pages and a directory (hub) page usually points to lots of other pages. This situation fits well the natural spreading of biological pheromone. The Web pheromone is spread across the Web links which mimics the conveyance of importance from one page to another.

Assuming $P_1, \cdots, P_k$ where $|P_i| = N_i, i = 1, \cdots, k$ are the k sets of pages that can be reached from document d by following Web links $i = 1, \cdots, k$ times, then we define the pheromone of the document d with spreading as:

$$p_V(d) := p_V(d) + \sum_{i=1}^{k} \sum_{j=1}^{N_i} P_i(j) \times f^i$$

where $P_i(j)$ is the j'th page in the set P_i; and $f \in [0, 1)$ is a fading factor for the pheromone spreading. $p_T(d) := t$ is the time when d is accessed, and the update is done after Pheromone Evaporation but before Accumulation.

Spreading of Web pheromones against the direction of links makes the more relevant documents more "attractive" since it will more likely attract surfers, this also parallels the situation of the natural swarm intelligence where insects return to the net by following and leaving more pheromones on the same path.

We define a maximal spreading radius to avoid cycling. For the purpose of simplicity in our proof-of-concept implementation we set the radius to 1.

Both the fading factor and the spreading radius affect the problem of cycle detection and avoidance. However it is a well studied problem in graph theory and its applications for example the Internet routing algorithms, so we don't further investigate it in our model here.

4 Evaluations

Our evaluations include 2 parts. One is called Quest for Treasure. The idea was to start a quest for a treasure in order to see, if in a self-organised system, changes to the environment will result in a collective optimisation of navigation.

We had 12 rooms where the visitors could navigate through. In two of these rooms were treasure chests, which had to be explored. Above each button were the actual pheromone density of the underlying link and the plain visit counter. Visitors had to use for the navigation the density which was computed and shown by the server module. Red numbers remind people that the pheromone there has very high density. After a certain time we could just follow the red links and we found the treasure. The next morning we removed one of the treasure chests,

and we could observe that during the next few hours the colors had changed. On the way to the room where we removed the chest the density of pheromone was decreased and the red links again led now to the one and only chest. So we could observe a very simple form of self-organisation by collectively using the Web pheromone information from the server module.

Hence, we demostrate that the swarm of internet surfers is indeed more intelligent than a single surfer.

The other evaluation we made is the comparison between our method and two other popular approaches in the computation of Web document rankings: plain Web counter and PageRank [7]. We call the ranking of Web documents generated by our *Swarm Intelligent* Web server an *Intelligent Ranking* of the documents. We did a small-scale experiment with a lab Web site which has about 200 Web pages in order to explore this possibility. We pre-computed the static PageRank ranking. We installed the module and requested volunteers to surf the site. Basically the visitors are the professors, assistants and students of the university. The experiment lasted for 2 days (because the module is not quite stable and its crash from time to time influences the Web server's functioning).

The results show that the *Intelligent Ranking* is quite different from the PageRank which shows our dynamic model in the social Web plays an important role in generating potentially more precise and recent Web document rankings.

5 Related Works

Related works have been carried on at different levels. The term *information diet* used earlier in this paper is more at the level of social science. Researchers suggest that [6] the optimal selection of Web pages to satisfy a user's information needs is a kind of optimal information diet problem. Because of these social aspects of the Web information, search engines and information extraction from unstructured sources become extremely important. Two prominent algorithms, HITS [8], and the PageRank [7] are found important. But both of them have their limitations and are not suitable for doing search in a decentralized environment. [3]

At a more technical level, collaborative filtering and relevance feedback techniques are studied. Collaborative filtering is basically a technology for distribution of opinions and ideas in society and facilitating contacts between people with similar interests [4]. It is in fact not directly related to information retrieval, rather a mate of IR, i.e. a way to help people look for information of desire.

Relevance Feedback is tightly related with Collaborative Filtering. Visitors are provided a form to input their evaluation over the Web document in a range of 1 to 10 usually. This sort of system often works as an extension to a Web browser. For example, in the Ant World system [9], which also uses the term pheromone in its Digital Information Pheromones (DIP), when a user starting her quest, she formulates the goals of the quest in a short description, similar to a search engine query, and an optional long description, resembling a TREC query. As the user browses the Web, a small "console window" hovers on top of the screen, soliciting the user's opinion on the usefulness of the pages she visits for the goals of her quest. But this kind of system is quite inconvenient for the end users. It bothers the users to input manually evaluation values or descriptions

for every Web page she visits. Its practical significance is limited for this reason. On the other hand our approach does not impose any extra requirements on the surfers in their surfing sessions.

6 Conclusion

In this paper, we propose the idea of using swarm intelligence in the context of learning and mining of the Web. We define a model of users' interaction with a Web server with regard to the accesses to Web documents. We explained the concepts, and the strategies we adopt in the proof-of-concept implementation. We verify our idea in two aspects: one to show that collective intelligence is feasible to obtain and it is better than any single one in a self-organized system; the other to obtain meaningful ranking for Web documents on a Web server. We expect more promising results in the future to demonstrate the significance of social intelligence in the field of Web.

References

1. Eric Bonabeau, Marco Dorigo, Guy Theraulaz, "Swarm Intelligence: From Natural to Artificial Systems", Oxford University Press, 1999.
2. http://iridia.ulb.ac.be/ ants/
3. Karl Aberer, Jie Wu, "A Framework for Decentralized Ranking in Web Information Retrieval", The Fifth Asia Pacific Web Conference, APWeb 2003, Xi'an, China, 23–25 April 2003.
4. Alexander Chislenko, "Automated Collaborative Filtering and Semantic Transports", http://www.lucifer.com/ sasha/articles/ACF.html, 1997.
5. Chun Wei Choo, Brian Detlor, Don Turnbull, "Web Work: Information Seeking and Knowledge Work on the World Wide Web", Kluwer Academic Publishers, 2000.
6. Peter Pirolli, James Pitkow, Ramana Rao, "Silk from a Sow's Ear: Extracting Usable Structures from the Web", Conference on Human Factors in Computing Systems, CHI-96, Vancouver, British Columbia, Canada, 13–18 April 1996.
7. Larry Page, Sergey Brin, Rajeev Motwani, Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", Technical report, Stanford, Jan. 1998.
8. Jon Kleinberg, "Authoritative Sources in a Hyperlinked Environment", in *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 1998.
9. Paul Kantor, Endre Boros, Ben Melamed, Dave Neu, Vladimir Menkov, Qin Shi, Myung-Ho Kim, "Ant World", in *Proceedings of SIGIR'99 (22nd International Conference on Research and Development in Information Retrieval)*, 1999.