Web Information Foraging

Yassine Drias and Gabriella Pasi

Università degli Studi di Milano-Bicocca, DiSCo
Viale Sarca 336, 20126 Milano
y.drias@campus.unimib.it , pasi@disco.unimib.it

Abstract. We present in this paper an approach to Web information foraging. We implemented a technique that helps Web users undertaking information foraging by simulating their behavior using a colony of artificial ants. Experiments were conducted on a website dedicated to the domain of Health. The results are promising and show the ability of our Web information foraging approach to find relevant Web pages.

Keywords: Information Foraging, Information Seeking, Ant Colony Optimization

1 Introduction

A recent paradigm related to accessing relevant information on the Web is Information Foraging. The task of information foraging consists in browsing the Web to collect information related to specific user needs under a time constraint. Recent work on information foraging are focalized to define algorithms that are able to discover in an automatic way the surfing paths that web users would follow while seeking information on the Web. The development of such systems may allow the users to spend less time locating the needed information; moreover this will also help them to easily identify the best sources containing that information.

The task of foraging is grounded on the *Optimal Foraging Theory* [10] developed by anthropologists to model the animal behavior while foraging food.The *Information Foraging Theory* was first developed in 1999 [7]. The authors established their study on the similarity between the animal food foraging behavior described in the Optimal Foraging Theory and the behavior of humans while seeking information online. The theory is based on the assumption that, when searching for information, users rely on their senses to perceive the *information scent* that helps them to reach their goal just like animals do when they follow the scent of their preys.

Information foraging consists in simulating the behavior of real users while seeking information on the Web. At the beginning, the information foraging process starts from a Web page and then according to the information need of the user, it decides which Web pages to visit in order to reach Web pages containing relevant information to the user. At the end of the process, a collection of surfing paths containing relevant Web pages is produced.

In this paper an approach to Web information foraging based on Ant Colony Optimization is proposed and evaluated. The artificial ants simulate the behavior of Web users while searching for information on the Web taking into consideration the complex structure of the Web and its volume. For this purpose, we propose a Web surfing model and implement a Web surfing strategy.

The rest of the paper is organized as follows. In section 2, we report the studies that are the most related to our concern. Web information foraging described is then described in section 3. In section 4, we present the contribution we developed for Web information foraging using Ant Colony Optimization. Section 5 summarizes the experiments we conducted on a medical website and the results we achieved.

2 Related Works

In the study presented in [8], the authors consider the Web as a semantic space and try to predict the navigational choices of Web users. The notion of information scent is introduced, which is measured as the mutual relevance between the user's goal and the Web pages' content. The authors tested their model on a database of selected tasks collected by a survey of more than 2000 web users. The results show that the measure of information scent is able to generate good estimations of web user interactions by predicting the links the users will click on and also when they decide to leave the website.

Strong regularities in Web surfing behavior were studied in [4] from a theoretical point of view. The authors proposed a model for studying surfing behaviors and the experiments they held showed common surfing behaviors. The study conducted in [3] shows that the Web pages are distributed over the sites according to a universal power law, which is an example of their strong regularities.

In [5] and [6] the authors consider Web topology, information distribution and interest profile in building a Wisdom information foraging agent. They found out that the unique distribution of agent interest leads for regularities in Web surfing. They also undertook an interesting study on three categories of users according to their interest and familiarity with the Web: A random user, a rational user and a recurrent user. The result is that independently from the type of users, the regularities of surfing are the same, which means that the user ability for predicting the surfing chain is predominant.

In [9] the author presents different interactive information retrieval models and shows the importance of developing such systems. Interactive information retrieval is based on human behavior and the fact that the feedback of Web users when performing a search on the Web can enhance the user experience and the performance of IR systems.

Link analysis has considerably improved the effectiveness of Web Search engines, thanks to the analysis of the hyperlink structure of the Web. According to [2], hyperlinks provide a valuable source of information for web information retrieval and a large number of links is created by independent individuals everyday.

In real life, web users generally do not get all the information they're looking for from the first Web page they visit. Our approach not only offers them the opportunity to have a direct access to the most relevant Web pages but also they can explore the whole surfing path which contains Web pages with complementary information that may interest them as well. This new way to access information is ensured thanks to the fact that information foraging is inspired from human Web navigation behavior. Our approach takes advantage from some related research areas such as interactive information retrieval and link analysis. It particularly takes into consideration the Web pages' information content, their distribution on the Web and the relation between them. The evaluation of our proposal was held on a real website, unlike previous works that were tested on log files or other forms of data.

3 Web Information Foraging

The Web is usually represented as a directed graph G(P, L). The vertices P correspond to the Web pages, where two Web pages p_i and p_j are connected via a directed edge if p_i contains a link referring to p_j. Considering a user with a specific interest, we may assume that he is interested in visiting a branch of the webgraph, one node at a time starting from an initial Web page and ending at a page containing some relevant information concerning his interest.

The transition choices that the user makes during the navigation define his surfing strategy. The set of visited Web pages while navigating is called a surfing path. The goal of information foraging is to determine the optimal paths to reach relevant Web pages for a given user interest. To this purpose, at each click on a page p_i , the question is to find the best move to another page p_j in order to build a surfing path.

In the literature it has been outlined that the surfing behaviors differ from one user to another, depending on the familiarity of the user with the Web environment, as well as his goal behind browsing the Web. In particular, three types of surfing behaviors have been identified in [5], but not formalized:

A *pseudo random strategy* which concerns users that are not familiar with the Web and are browsing it without having a strong interests in any specific topics; A *rational strategy* which is the closest to real-life Web navigation as most of Web users behave rationally. They usually have a specific goal behind Web surfing and they try to reach it by selecting the Web pages that seem the closest to their information need; A *recurrent strategy* concerning users who are familiar with the Web and have a well-defined goal behind browsing it. This kind of users always makes the best decision when they have to move from the current Web page they are into a new one by selecting the most relevant Web page among to the possible pages.

When surfing, Web users are guided by the information scent they get from Web pages. Starting with a weak amount, the information scent increases as the user gets closer to the Web pages that interest him. In order to model the

user's information need (interest), we consider a vector containing keywords that represent topics in which the user is interested.

4 Web Information Foraging using Ant Colony Optimization

As mentioned in the previous section, the majority of Web users behave rationally when seeking information on the Web [5]. The aim of our work is to develop algorithms to efficiently implement the rational surfing strategy. To this purpose we have applied Ant Colony Optimization (ACO). The main reason to choose ACO resides in the fact that it simulates the ants' food foraging behavior which is similar to the human information foraging behavior on the Web.

The ant system algorithm [1] includes several ant generations, each generation is composed of *NbAnts* ants. Where *NbAnts* is the population size of the ant colony. Each artificial ant starts building a solution from an initial state i generated randomly. Recall that a solution is a connection of states and it is constructed using a stochastic process. The ant chooses a new state j from the current state's neighborhood N_i, with a probability computed by Formula (1):

$$P(i,j) = \frac{phero[j]}{\sum_{l \in N_i} phero[l]} \tag{1}$$

To each state i is assigned an amount of pheromone denoted by *phero[i]*. The pheromone information is initialized with a very small value in order to simulate the fact that real ants deposit a very small amount of pheromone on the ground when starting their food foraging. Two structures are needed to compute the ant algorithm, a table named *Phero* to store the pheromone amount yielded by the ants each time they build a solution and a table called *sol* to save the best solution found by each ant. *phero[k]* corresponds to the pheromone amount associated with the solution found by ant k and *sol[k]* is the best solution determined by ant k. The tables are updated at each generation of ants. Besides, two variables namely *best* and *bestsol* are used to save respectively the best solution found during the current generation and the best solution computed since the beginning of the process. During the search, the pheromone amount, which represents the effectiveness of the solution, will be computed and associated with each solution found by the ants.

The strategies of updating pheromone simulate the evaporation of natural pheromone followed by a production of this chemical substance. The evaporation phenomenon gives rise to rule (2) where the empirical parameter ρ belongs to the interval [0, 1] and simulates the evaporation rate. Pheromone evaporation prevents from premature convergence. An online delayed update is performed at each generation of ants, the pheromone added is calculated for each state of the solution according to rule (3). It is a delayed update because the pheromone assigned to a state is updated once the ant determines a solution. For the offline update, rule (4) is applied. Recall that *bestsol* is the best solution found during

the previous iterations and *best* is the best solution of the current iteration. The added pheromone amount is proportional to the ratio of these values.

$$phero[i] = (1 - \rho) * phero[i] \quad (2)$$
$$phero[i] = phero[i] + \rho * f(s) \quad (3)$$
$$phero[k] = phero[k] + \rho * f(bestsol)/f(best) \quad (4)$$

4.1 Simulating the Rational Strategy using ACO

We present here the adaptation of the Ant Colony System algorithm to Web information foraging in order to simulate the rational surfing behavior.

As we previously mentioned in section 3, rational users are interested in specific topics and they forage in order to locate Web pages that contain information on those topics. When they reach a new Web page, they will try to decide whether or not the content sufficiently matches their interest and, if not, predict which Web page at the next level will become a more interesting one. In predicting the next-level contents, they will rely on the information scent they get from the titles or descriptions of various hyperlinks inside the current Web page. We notice that the information scent is analogous to the pheromone that guides the ants when they're seeking food. In our adaptation, the artificial ants simulate the behavior of Web users that have a rational surfing strategy.

Algorithm 1 ACO-WIF

Input: N (A part of the Web); user interest;
Output: bestsol, a surfing path ending with a relevant Web page;

```
 1: procedure ACO-WIF
 2:     for i=1 to NbAnts do phero[i]= 0.1;          ▷ pheromone initialization
 3:     end for
 4:     select at random a solution s from N;        ▷ a surfing path namely s
 5:     best := bestsol := s;
 6:     for i=1 to MaxIter do
 7:         for  k=1 to NbAnts do
 8:             sol[k] := build_Sol();
 9:             update the online pheromone using Formulas (2) and (3);
10:             if f(sol[k]) > f(best) then then best := sol[k];   ▷ f: fitness function
11:             end if
12:         end for
13:         if f(best) > f(bestsol) then bestsol := best;
14:         end if
15:         apply offline-update of pheromone using Formula (4);
16:     end for
17:     return (bestsol);
18: end procedure
```

Solutions encoding. The search space for the ant colony is the set of all possible surfing paths. In ACO ants build solutions which are surfing paths for

our case. A surfing path is composed by the set of all visited Web pages during the foraging process and contains at least one Web page. It starts with an initial page and ends with a target page which should incorporate relevant information. The number of Web pages contained in the surfing path is called the surfing depth. So ants will seek surfing paths that contain relevant Web pages. The adapted Ant Colony System for Web Information Foraging is outlined in Algorithm 1.

Building a solution. Each ant performs the task of exploring the best surfing path in a specific part of the Web. The ant builds a solution by selecting Web pages according to a probability P defined in Formula (5). In addition to the pheromone, we introduce a heuristic that brings a knowledge on our problem in order to help the ants to make better foraging moves. The heuristic we propose is measured as the similarity between potential next page p_j and the user's interest represented by the vector V. The neighborhood of a page p_i is a set of Web pages that are connected to p_i via a Web link. We introduce a noise parameter q_0 in order to simulate the fact that even Web users that behave rationally don't always make the best decision when surfing. This may be the result of multiple factors such as their unfamiliarity with the Web for example.

$$\begin{aligned} &\textit{if } q \leq q_0 \\ &\quad \textit{then } P(p_i, p_j) = \begin{cases} 1 \text{ if } p_j = argmax(phero[j]^{\alpha}(heur_{jv})^{\beta}) \quad \textit{for } p_j \in N_i \\ 0 \text{ else} \end{cases} \\ &\text{else} \qquad P(p_i, p_j) = \frac{phero[i]^{\alpha}\ (heur_{jv})^{\beta}}{\Sigma_{l \in N_i} phero[l]^{\alpha}\ (heur_{lv})^{\beta}} \end{aligned} \tag{5}$$

In other words, the ant decides stochastically to consider the most relevant Web page among the outgoing Web pages when $q \leq q_0$ and a Web page drawn at random otherwise.

The evaluation of a surfing path quality is performed by applying the fitness function f to the last Web page of the path. This function represents the similarity between the user interest and the description of the Web page which contains the title of the page and eventually some tags and keywords. It can be computed using one of the similarity functions defined in literature.

5 Experiments

5.1 Description of the real-world Benchmark

We performed our experiments on MedlinePlus, an online medical Website produced by the National Library of Medicine of the U.S. It provides information on over 900 diseases, health conditions and wellness issues. Our experiments deal with health topics described in an XML file that includes pages describing medical topics. The data for health topics is available at `http://www.nlm.nih.gov/medlineplus/xml.html`. We worked on the version of the 25th February 2016 where the number of Web pages was equal to 1946. Each topic is specified

by a title and contains the following elements: a URL, a unique identifier, the language of the topic (English or Spanish), the date of its creation, topic synonyms, translation to other languages, a full summary, a list of related topics, which are internal links to similar topics and external links.

5.2 Results

Different user interests were experimented for evaluating the rational surfing strategy. The results we focused on are: the last Web page on the surfing path (the most relevant one), its URL, its score, the surfing depth of the path and the surfing time in milliseconds. Table 1 exhibits the obtained results for the rational surfing strategy.

User interest	Relevant Page		Score	Surfing depth	Surfing time
	Title	URL			
Childhood, Leukemia	Childhood Leukemia	*/childhoodleukemia.html	1.0	2	115
Food, Poisoning	Foodborne Illness	*/Foodborneillness.html	1.0	3	167
Hypotension	Low Blood Pressure	*/lowbloodpressure.html	1.0	2	125
Skin, Allergies	Skin Conditions	*/skinconditions.html	0.5	2	992
Illness, Heat	Heat Illness	*/heatillness.html	1.0	4	175
Chest, Injury	Chest Injuries and Disorders	*/chestinjuriesanddisorders .html	0.66	5	338
Zika	Zika Virus	*/zikavirus.html	1.0	1	204

* : http://www.nlm.nih.gov/medlineplus

Table 1. Results for different user interests using the Rational Surfing Strategy

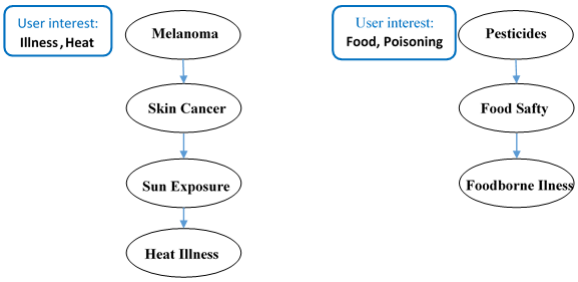

Fig. 1. Detailed surfing paths

From table 1, we can observe that our approach is able to find relevant Web pages to the user based on the vector describing his interest. The developed program also makes use of the health topics' synonyms provided by MedlinePlus in order to locate the most relevant Web pages to the user. For example, in the third instance of the table the user interest was *"Hypotension"* and the result returned by the program was a Web page entitled *"Low Blood Pressure"* which is a synonym of *Hypotension*.

Figure 1 shows more details about the results for the user interests *"illness, heat"* and *"food, poisoning"*. The surfing path for *"illness, heat"* starts from a Web page named *"Melanoma"* and ends with the Web page *"Heat Illness"*. The surfing depth in this case is equal to 4. We notice that the last Web page in the surfing path is the most relevant to the user. However, the other Web pages are also related to the user's interest and may be interesting for him.

6 Conclusion

In this work, we presented an approach to Web information foraging based on Ant Colony Optimization. We proposed a model for Web surfing in order to simulate the Web surfing behavior of real Web users. This idea was inspired from the information foraging theory which states that Web users and animals have similar behaviors when looking for information/food. Furthermore, a real website was used for the experiments instead of an artificial one or log files as it was performed in the literature. The results show the ability of our program to find relevant Web pages in a short time based on a user interest. The outcomes consist in a set of surfing paths ranked by relevance. This offers the user the possibility to go more in depth with getting information on a certain topic without spending too much time on visiting a lot of Web pages.

As a perspective, we intend to investigate Web information foraging in social and information sharing networks such as Twitter.

References

1. Dorigo, M., Caro, G.D.: Ant algorithms for discrete optimization. Artificial Life 5-3, 137–172 (1999)
2. Henzinger, M.R.: Link analysis in web information retrieval. IEEE Data Eng. Bull. 23(3), 3–8 (2000)
3. Huberman, B.A., Adamic, L.A.: Growth dynamics of the world-wide web. Nature 40, 7478–7491 (1999)
4. Huberman, B.A., Pirolli, P., Pitkow, J.E., Lukose, R.M.: Strong regularities in world wide web surfing. Science (1997)
5. Liu, J., Zhang, S.W.: Characterizing web usage regularities with information foraging agents. IEEE Transactions on Knowledge and Data Engineering 40, 7478–7491 (2004)
6. Liu, J., Zhong, N., Yao, Y., , Ras, Z.: The wisdom web: New challenges for web intelligence (wi). Expert System With Applications 40, 7478–7491 (2013)
7. Pirolli, P., Card, S.K.: Information foraging. Psychological Review 106(4), 643–675 (1999)
8. Pirolli, P., Fu, W.T.: Snif-act: A model of information foraging on the world wide web. User Modeling 2003, 9th International Conference 22-26, 45–54 (2003)
9. Robins, D.: Interactive information retrieval: Context and basic notions. InformingSciJ 3, 57–62 (2000)
10. Werner, E.E., Hall, D.J.: Optimal foraging and the size selection of prey by the bluegill sunfish (lepomis macrochirus). Ecology 55(5), 1042 (1974)