A Collaborative Multi-Agent Approach to Web Information Foraging

Yassine Drias and Gabriella Pasi

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

Abstract. In this paper the task of Information Foraging (IF) is considered as a useful paradigm to address Exploratory Search. In the context of IF, a Web navigation strategy is introduced and formalized, and a multi-agent based model is proposed to exploit a collaborative approach to Information Foraging. A system based on this model has been developed, and its preliminary evaluations on the ACM and DBLP repositories are reported. The results are promising and show the ability of the proposed Web Information Foraging system to find relevant Web pages.

Keywords: Web Information Foraging, Collaborative Search, Multi-Agent Systems

1 Introduction

Nowadays, the human ability to explore the huge amount of data on the Web is relatively limited [1] [14]. The usual way in which users address the task of locating information relevant to their needs is by means of Web Search Engines, which support the task of Information Retrieval (IR).

In [2], the author pointed out that Web queries can be classified in three categories based on the user intent. They can be *navigational*, when users aim to reach a particular Web page; *informational*, when the goal is to get information on a certain topic, or *transactional* if the purpose is to reach Web pages where further interaction would take place. More recently, the Exploratory Search paradigm has been introduced, to address situations in which users intend not only to find relevant documents, but they are also interested in learning, discovering and understanding complex and new topics [5]. Besides querying, Exploratory Search systems are also based on Web browsing strategies. The idea is to go beyond returning a set of ranked results as an answer to a query [4, 13].

A recent paradigm related to Exploratory Search, and which aims at discovering paths leading to relevant information on the Web is Information Foraging. The idea of Information Foraging is grounded on the *Optimal Foraging Theory* [12] developed by anthropologists to model animals' behavior while foraging food. The *Information Foraging Theory* was first developed in [9]. The authors established their study on the similarity between the animal food foraging behavior and the behavior of humans while seeking information on the Web. 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.

The task of Web Information Foraging consists in browsing the Web to collect information related to specific user's needs. An automatic Information Foraging process simulates the user's behavior while surfing on the Web with the aim of finding relevant information. It starts with a topical user interest, then by using an information scent measure [10] it aims to find the surfing paths that lead to Web pages relevant to the user. Recent works are focused on defining Information Foraging systems that are able to discover in an automatic way the surfing paths that could be useful to users in order to help them to learn, discover and augment their knowledge on a specific topic [7, 8].

In this paper, we propose a new collaborative Web Information Foraging model; to make Web Information Foraging more effective and efficient, we propose to address it in a collaborative way. In fact, as outlined in [6], a collaborative approach to the task of Web Search allows to achieve a higher recall and has the potential to improve users' search skills. The collaborative aspect of our model relies on creating several aliases of a user seeking information on the Web by means of a multi-agent system. In this paper we also report some preliminary evaluations conducted on large datasets including the *ACM* and the *DBLP* repositories, which show the effectiveness of the proposed approach.

2 A Collaborative Approach to Web Information Foraging

The Web can be represented as a directed graph $G(P, L)$, where the vertices P correspond to Web pages, and the edges L represent the connections between pages. A directed edge connects a page p_i to a page p_j if on page p_i there exists a link referring to page p_j. Let us consider a user with a topical interest, who starts surfing on the Web from an initial Web page to reach a page containing relevant information to her/his interest. The sequence of visited Web pages during this process is called a *surfing path*. The transition choices that the user makes when surfing define her/his surfing strategy, which affects the way s/he selects the next Web page to visit, based on the information scent s/he perceives of each accessible page. The information scent of a given page is an estimation of how much useful information the user is likely to get on this page. Web surfing strategies are related to users' behavior, which may depend on the users' familiarity with both the Web and computer science technologies, and also on their purpose behind Web surfing.

In Section 2.1 we propose and formalize a Web surfing strategy, and we exploit it by defining and implementing a collaborative (multi-agent) system for Information Foraging.

2.1 The employed Web surfing strategy

In their Web navigation paths, users aim at increasing the information scent they perceive when moving from one Web page to another. If at time t a user is on a page p_i, her/his motivation to move to a new page p_j relies on the increase of information scent s/he would obtain in moving from the former to the latter page. The information scent increase at time *t+1* with respect to time t can be modeled by a quantity proportional to the similarity between the content of the new selected page p_j and the user's interest, which consist in a vector of keywords representing the user's information need and her/his motivation behind Web surfing. We propose Formula (1) to estimate the information scent a user would get from a given Web page p_j.

$$InfoScent(p_j) = InfoScent(p_i) + S \tag{1}$$

where S is the similarity between the user's interest and the content of page p_j. If we suppose that p_i is the initial Web page from which the foraging starts, then the information scent at page p_i would be equal to the similarity between its content and the user's interest.
In [3] different Web surfing strategies have been analysed but not formally modeled; among the proposed strategies the *rational surfing behavior* is related to users who start Web surfing with specific goals, but who might incur in some random choices due to a lack of knowledge on certain topics or unfamiliarity with the Web.

We propose a formalization of the rational surfing behavior in Formula (2), which defines the probability $P(p_i, p_j)$ to move from a page p_i to one of its linked pages p_j.

$$P(p_i, p_j) = \frac{InfoScent(p_j)^{\alpha} \times sim(p_i, p_j)^{\beta}}{\Sigma_{l \in N_i}(InfoScent(p_l)^{\alpha} \times sim(p_i, p_l)^{\beta})} \tag{2}$$

where α and β are parameters that control the relative weight of the *InfoScent* and the content-based similarity of pages p_i and p_j. N_i is the set of outgoing pages from page p_i.

The outcome of a Web information foraging process consists in a surfing path containing all the Web pages visited by the user during the Web browsing. The path starts from an initial page and ends with a target page which contains information relevant to the user's interests. The number of Web pages contained in the surfing path is called the surfing depth. A relevance score can be associated with a surfing path, by assessing the similarity of the last page in the path and the formal representation of the user's interest (we do this by formally representing the above objects as vectors of terms and by applying the *Cosine Similarity*).

In order to make the foraging process operational and efficient, and to the aim of defining effective systems based on this paradigm, it is unrealistic to provide just a path as a potential answer to an information need. To tackle the above issues, we propose a collaborative model that simulates the user's surfing activity starting from several distinct initial pages; the proposed approach is presented in Section 2.2.

2.2 Web Information Foraging using a Multi-agent System

The collaborative approach to information foraging proposed in this Section is based on the Web surfing strategy described in section 2.1. By this approach the foraging activity motivated by an information need is carried out by several agents that act as *aliases* of a specific Web user. The Web is the environment where the agents operate. We adopt a graph-based representation of the Web as previously explained, and we assume that each agent browses a part of the Web graph with the aim of finding relevant information. Starting with the same user interest, the agents forage information simultaneously in the Web with the goal of reaching relevant Web pages. The collaboration feature is ensured through the blackboard communication model, which offers real-time communication and results sharing between the agents as suggested in [6]. The system is parallelized and decentralized, and it is composed of two kinds of agents that can work simultaneously:

- A group of foraging agents that have the task of simulating the behavior of a specific Web user when searching for information on the Web;
- A blackboard-handler agent that takes in charge the management of the blackboard system and the sorting of the partial solutions proposed by the foraging agents.

Figure 1 illustrates the architecture of the multi-agent system that is composed by the following modules:

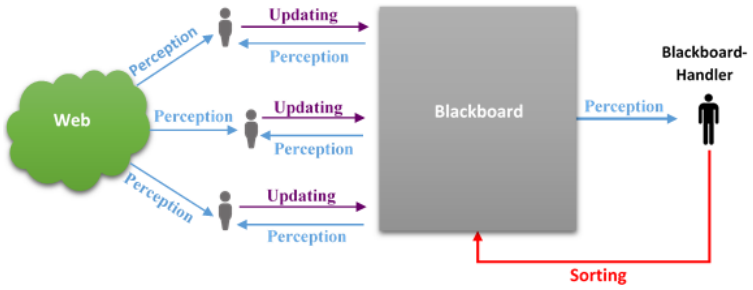

Fig. 1. Architecture of the Web Information Foraging Multi-Agent System.

Foraging Agents Each foraging agent has the task of simulating the rational behavior of a single Web user when searching for information on the Web. More concretely, the agent must determine at every click corresponding to a page p_i the best move to another page p_j.

At the beginning of the process, several foraging agents are launched at the same time and each of them starts its foraging from an initial Web page randomly selected. Each time a foraging agent makes a move towards a new Web page

using Formula (2), it evaluates its new partial solution (actual surfing path) and communicates it to the other agents by writing it on the blackboard. The blackboard-handler agent will then sort the solutions according to their relevance to the user interest and will share the top solutions on the blackboard. If the proposed solution by the foraging agent is a top solution then the agent continues surfing on the same path, else the agent abandons its current surfing path and starts surfing again from a new Web page defined randomly. Procedure 1 presents the pseudo-code of the foraging agents' behavior.

Procedure 1: Foraging Agent

1. Select a Web page at random ;
2. Perceive the environment and check the neighborhood of the current Web page ;
3. Move to a new Web page using Formula 2 ;
4. Add the new Web page to the surfing path ;
5. Communicate the surfing path to the other agents by writing it on the blackboard ;
6. Wait for the blackboard-handler agent to sort the solutions ;
7. Check the blackboard ;

if *8. the surfing path belongs to the top solutions* **then**
  go to 2 ;
else
  9. Abandon the current surfing path and Go to 1 ;

Blackboard System The blackboard system is a communication model that allows information sharing between the agents in our system. More concretely, it is an area of shared memory accessible in read and write mode by all the agents. The blackboard facilitates the collaboration between the agents as it offers them the possibility to exchange knowledge and share their partial solutions and suggestions for the sake of finding the best information sources that give answers to the user's interest.

Blackboard-Handler Agent The main function of this agent is to sort the solutions reported by the foraging agents in the blackboard according to their relevance to the user interest. Each time a *foraging agent* reports a new solution in the blackboard, the blackboard-handler agent inserts it in the right position in order to keep the top solutions sorted. Procedure 2 shows the pseudo-code of the blackboard-handler agent.

3 Preliminary Experiments

In this section we report the preliminary experiments that we have undertaken on both the DBLP and ACM repositories. Both repositories give access to sci-

Procedure 2: Blackboard-Handler Agent

1. Check the blackboard ;
2. Sort the solutions proposed by the foraging agents according to their relevance ;
3. As soon as a new solution is reported by a foraging agent insert it in the right position in the sorted list containing the top solutions ;
4. Update the top solutions list on the blackboard ;
5. Go to 3;

entific publications through their websites where each scientific publication has a corresponding Web page. Based on experiments conducted on the considered datasets, we fixed the values of the empirical parameters of Formula (2) as follows: $\alpha = 2, \beta = 1$. Furthermore, We used 80 *foraging agents* and fixed the size of the top solutions list to 25.

3.1 Datasets and Examples of Results

We built our first dataset using the XML version of the *DBLP* website available on http://dblp.uni-trier.de/xml. We constructed a partial graph of the website considering 65,336 articles that represent the nodes of the graph. After that, we linked all the articles that share at least one author in common. The obtained graph contains 65,336 nodes, it has a size of 640,810 and a density equal to $0.15 * 10^{-3}$.

As a second Dataset, we used the *ACM-Citation-network V8* [11], which contains 2,381,688 papers extracted from *ACM* along with 10,476,564 citation relationships. Each paper is associated with an abstract, a list of authors, a year, a venue, and a title.

Table 1 shows the top five results on the DBLP dataset for the user's interest described in the first column. The second column exhibits the last page of the surfing path found by the system. The third and fourth columns present respectively the score of the surfing path and the surfing depth.

Table 2 presents the outcomes yielded by our approach on the ACM citation network. For each user interest we give the full surfing path starting from the first visited page. We notice that the last page in the surfing path is the most relevant to the user. However, the other pages are also related to the user's interest, and they may be interesting to her/him.

3.2 Comparative Evaluations

In order to validate our approach, we compared the results produced by our system to those returned by the *DBLP* and the *ACM* local search engines. In fact, both repositories are equipped with local search engines (http://dblp.uni-trier.de/search/, http://dl.acm.org/), which give the Web users the opportunity to perform a keyword-based search for scientific publications. We selected about

User's Interest	Last Web page in the Surfing Path	Score	Surfing Depth
Approximate, Personalized, Information, Retrieval, Database, Management, Systems	Approximate Retrieval: A Comparison of Information Retrieval and Database Management Systems	0.86	4
	Exploiting the Functionality of Object-Oriented Database Management Systems for Information Retrieval	0.67	3
	Models for Integrated Information Retrieval and Database Systems	0.62	3
	Personalized Information Retrieval System	0.57	3
	Enriching Information Retrieval	0.44	2

Table 1. Top 5 ranked results for a user interest on DBLP

20 user's interests representing information needs related to scientific domains and related research fields. We introduced the selected user's interests separately as inputs to our system, and then we performed a search with those same user's interests by using the local search engines of each website. The results produced by our approach were similar to those of the *DBLP* and *ACM* search engines, however there are some differences it is worth to point out. The first difference is related to the fact that although the datasets we used do not cover the complete *DBLP* and *ACM* repositories, this did not affect the results quality returned by our approach regarding the tested user's interests. The second difference resides in the way of presenting the search results to the user. Both *DBLP* and *ACM* employ Information Retrieval techniques, for each search a set of results is returned. The results are ranked according to their relevance to the considered query, and each of them direct to a Web page dedicated to a scientific publication. On the other hand, our approach returns a set of surfing paths also ranked by their relevance. Each surfing path contains one or more Web pages with the last one being the most relevant. Our way of presenting the search results using Information Foraging takes into consideration the relations between the different scientific publications/Web pages, unlike the classical way used by the majority of search engines. For instance, the knowledge about the co-authorship relation in our *DBLP* dataset and the knowledge about the citation relation in the *ACM-Citation-network* enhance the search process by allowing the system to benefit from those relations, thus not being only dependent on the keywords describing the user's interest.

3.3 Users' Feedback

We asked a group of 10 Master's and Ph.D. students to give us 3 topics related to their research area on which they want to get scientific articles. At the end we got 30 different user interests, each one of them is modeled by a set of keywords that describe the information need of the user. After that, we have run our program for each user interest and presented the outcomes to the corresponding

User's Interests		Surfing Path	Surfing Depth
Ant, System, Optimization	1	The cloud-based framework for ant colony optimization	2
	2	**Ant system: optimization by a colony of cooperating agents**	
Neural, Networks	1	Automatic detection of photoresist residual layer in lithography using a neural classification approach	2
	2	**Neural Networks**	
Satisfiability, Sat, Optimization	1	Counting solutions of CSPs: a structural approach	3
	2	CNF satisfiability test by counting and polynomial average time	
	3	**Global Optimization for Satisfiability (SAT) Problem**	
Information, Foraging	1	Automated content transformation with adjustment for visual presentation related to terminal types	3
	2	The Web Book and the Web Forager: an information workspace for the World-Wide Web	
	3	**Information foraging in information access environments**	
Information, Retrieval, Algorithms	1	How far we've come: Impact of 20 years of multimedia	4
	2	Content-Based Image Retrieval at the End of the Early Years	
	3	Text databases and information retrieval	
	4	**Information retrieval: data structures and algorithms**	

Table 2. Results for different users' interests on the *ACM-Citation-network V8* dataset

user. We then asked each user to determine the position of the most relevant result among the list of ranked results.

Figure 2 details the feedback we got from the users. The values on the horizontal axis represent the position of the surfing path that is the most relevant to the user while the values on the vertical axis correspond to the rate of the satisfied user interests.

4 Conclusion

We proposed in this paper a new collaborative multi-agent approach to Web Information Foraging where the foraging agents cooperate to find relevant Web pages according to a user's interest. We defined a Web surfing strategy inspired from real Web users' behavior and gave the pseudo code of the software agents. In order to obtain a preliminary evaluation of our approach and compare it to classical IR, we considered two datasets with different sizes.

The results we achieved show the ability of our approach to find relevant Web pages in an effective way based on a user interest. Our approach is also capable of exploiting relations related to the Web structure. For instance, it can use the citations and co-author relations in order to access scientific publications.

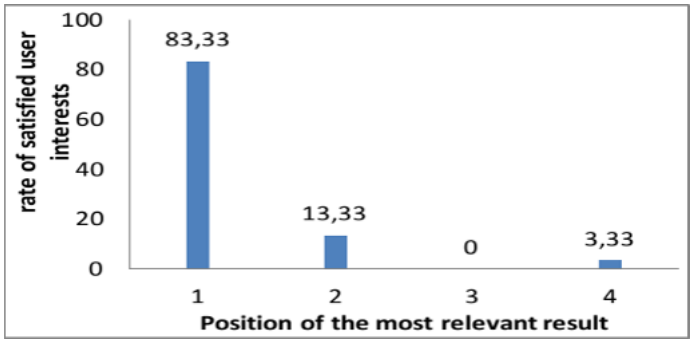

Fig. 2. Outcomes of a user study involving 10 students.

References

1. Berman, F.: From teragrid to knowledge grid. Communications of the ACM 44, 27–28 (2001)
2. Broder, A.Z.: A taxonomy of web search. sigir forum. SIGIR Forum 36(2), 3–10 (2002)
3. J., L., S.W, Z.: Characterizing web usage regularities with information foraging agents. IEEE Trans. Knowl. Data Eng 40, 74787491 (June 2004)
4. Marchionini, G.: Exploratory search: from finding to understanding. Commun. ACM 49(4), 41–46 (2006), http://doi.acm.org/10.1145/1121949.1121979
5. Mirizzi, R., Noia, T.D.: From exploratory search to web search and back. In: Proceedings of the Third Ph.D. Workshop on Information and Knowledge Management, PIKM 2010, Toronto, Ontario, Canada, October 30, 2010. pp. 39–46 (2010), http://doi.acm.org/10.1145/1871902.1871910
6. Morris, M.R.: Collaborative search revisited. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. pp. 1181–1192. CSCW '13, ACM, New York, NY, USA (2013), http://doi.acm.org/10.1145/2441776.2441910
7. Paik, J., Pirolli, P.: ACT-R models of information foraging in geospatial intelligence tasks. Computational & Mathematical Organization Theory 21(3), 274–295 (2015), http://dx.doi.org/10.1007/s10588-015-9185-x
8. Piorkowski, D., Fleming, S., Scaffidi, C., Bogart, C., Burnett, M., John, B., Bellamy, R., Swart, C.: Reactive information foraging: An empirical investigation of theory-based recommender systems for programmers. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 1471–1480. CHI '12, ACM, New York, NY, USA (2012), http://doi.acm.org/10.1145/2207676.2208608
9. Pirolli, P., Card, S.K.: Information foraging. Psychological Review 106(4), 643–675 (1999)
10. 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)
11. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: KDD'08. pp. 990–998 (2008)
12. 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)

13. White, R.W., Roth, R.A.: Exploratory Search: Beyond the Query-Response Paradigm. Synthesis Lectures on Information Concepts, Retrieval, and Services, Morgan & Claypool Publishers (2009), http://dx.doi.org/10.2200/S00174ED1V01Y200901ICR003
14. Zhu, Y., Zhong, N., Xiong, Y.: Data explosion, data nature and dataology. Proceedings of the 2009 International Conference on Brain Informatics, BI'09, Springer-Verlag (2009)