Swarm Intelligence: Agents for Adaptive Web Search

Fabio Gasparetti and **Alessandro Micarelli** [1]

Abstract. In this paper, we present an adaptive and scalable Web search system, based on a multi-agent reactive architecture, which drew inspiration from biological researches on the ant foraging behavior. Its target is to search autonomously information on particular topics, in huge hypertextual collections, such as the Web, exploiting the outstanding properties of the agent architectures. The algorithm has proven to be robust against environmental alterations and adaptive to user's information need changes, discovering valuable evaluation results from standard Web collections.

1 INTRODUCTION

Search engines are the most conventional search tools Web users usually exploit to fulfil their information needs. Information Retrieval (IR) techniques are employed to retrieve interesting documents from large file collections built by means of high-performance crawlers, surfing endlessly new and updated resources. This approach arises several problems, such as the lack of freshness in the queries' results, and the low coverage of the Web. Furthermore, due to mirroring and aliasing, the resources can occur many times.

The aim of this paper is to present an adaptive and scalable Web search system, based on a reactive architecture, composed of a colony of cooperative distributed agents, which drew inspiration from researches on the ant foraging behavior [1]. The autonomy and the reactivity properties embodied in the architecture, yield the robustness against environmental alterations, and the adaptivity to the information need changes, reducing the exploration time, and increasing the quality of the retrieved information.

2 THE ADAPTIVE WEB SEARCH SYSTEM

The proposed adaptive Web search system is based on a innovative architecture which is inspired by the ant foraging behavior model and the Ant System computational paradigm. It is composed of a large number of agents trying collectively to satisfy user's requests. The intelligent behavior arises because of the agent's interaction with the environment, and with the other system's agents.

This work begins from an empirical observation: a link often represents the author's intention to connect the page in which it is located to another one regarding the same topic (phenomenon sometimes called *linkage locality* or *link-context conjecture*). So, if a page concerns a topic a user is interested in, and if he is currently visiting this page, he will probably look at the linked pages, because they could be probably related to the topic at issue. This observation has been employed in several algorithms which exploit the presence of the link structure among documents to guide the crawling on behalf of the users (e.g., [3, 5]).

Nevertheless, a typical similarity measure analysis between the page's content and the query is not enough to direct a crawl towards the interesting resources, as widely discussed in [4]. These observations have been taken into consideration in the developed agent architecture, and that is why it represents an interesting approach compared to the other architectures proposed in the literature.

The agent architecture is composed of a reactive agents colony living in a hypertextual environment, where it is possible to associate a similarity measure between the resources and the user' information needs.

Each agent corresponds to a virtual ant that has the chance to move itself from the hypertextual resource where it is currently located url_i, to another url_j, if there is a link in url_i that points to url_j. A sequence of links (i.e., pairs of urls) represents a possible agent's route, where the pheromone hormone trail could be released on at the end of each exploration. The available information for an agent when it is located in a certain resource is: the matching result of the resource with the user query, and the amount of pheromone hormone on the paths corresponding to the outgoing links.

The pheromone trails represent the mean that allows the ants to make better local decisions with limited local knowledge both on environment and group behavior. The ants employ them to communicate the exploration results one another: the more interesting resources an ant was able to find out, the more pheromone trail it leaves on the followed path. As long as a path carries relevant resources, the respective trail will be reinforced and the number of the attracted ants will increase.

The system execution is divided into cycles; in each of them, an ant makes a sequence of moves among the hypertextual resources. The allowable moves per cycle depend proportionally on the value of the latter. At the end of a cycle, the ants update the pheromone intensity values of the followed path as a function of the retrieved resource scores.

To select a particular link to follow, a generic ant located on the resource url_i at the cycle t, draws the *transition probability* value $p_{ij}(t)$ for every link contained in url_i that connects url_i to url_j. The $p_{ij}(t)$ is reckoned by the formula:

$$p_{ij}(t) = \frac{\tau_{ij}(t)}{\sum_{l:(i,l)\in E} \tau_{il}(t)} \qquad (1)$$

where $\tau_{ij}(t)$ corresponds to the pheromone trail between url_i and url_j, and $(i,l) \in E$ indicates the presence of a link from url_i to url_l (it has been used a Web directed graph representation (V, E), where V is the hypertextual page collection, and E is the link set).

To keep the ants from following circular paths, and to encourage the page exploration, each ant stores a L list containing the visited urls. A probability related to the path from url_i to url_j, if the url_j belongs to L, is 0. At the end of every cycle, the list is emptied out.

When the limit of moves per cycle is reached, the ants start the

[1] University of ROMA TRE, Rome, Italy email: gaspare@dia.uniroma3.it, micarel@dia.uniroma3.it

trail updating process. In this work we have evaluated two *updating rules*. In the first, the pheromone variation of the k-ant corresponds with the mean of the visited resource scores:

$$\Delta\tau^{(k)} = \frac{\sum_{j=1}^{|P^{(k)}|} score(P^{(k)}[j])}{|P^{(k)}|} \qquad (2)$$

where $P^{(k)}$ is the ordered set of pages visited by the k-ant, $P^{(k)}[i]$ is the i-th element of $P^{(k)}$, and $score(p)$ is the function that, for each resource p, returns the similarity measure with the current information needs: a $[0,1]$ value, where 1 is the highest similarity.

The process is completed with the τ value updates. The τ_{ij} trail of the generic path from url_i to url_j at the cycle $t+1$ is affected by the ant's pheromone updating process, through the computed $\Delta\tau^{(k)}$ values:

$$\tau_{ij}(t+1) = \rho \cdot \tau_{ij}(t) + \sum_{k=1}^{M} \Delta\tau^{(k)} \qquad (3)$$

where ρ is the trail evaporation coefficient. It must be set to a positive value less than 1 to avoid unlimited accumulation of substance caused by the repeated positive feedback. The summation widens to a subset of the N ants living in the environment. The definition of this subset will be soon discussed. At the beginning of the execution, all the $\tau_{ij}(0)$ values are set to a small constant τ_0.

An elitist strategy has been devised to speed up the exploration towards the most fruitful paths. At the end of the cycles, the subset of the N ants which have discovered the most interesting routes carry out the pheromone updating process, whereas the other exploration results are discarded. The subset's cardinality is M, which corresponds to the value in Eq. 3.

At every cycle the pheromone trails $\tau_{ij}(t)$ are updated according to the visited resource scores. This technique ensures the robustness against environmental alterations and the adaptivity to the user's information need changes. For instance, if a negative variation of the $\Delta\tau_{ij}^{(k)}$ value occurs at the cycle t (e.g., due to a query refinement, or a page alteration), the $\tau_{ij}(t+1)$ are subjected to a reduced feedback. For this reason, less ants will be attracted, and the $\Delta\tau_{ij}^{(k)}$ increments at the cycle $t+1$ are still further reduced, and so forth. In other words, each change in the environment causes a feedback that modifies the global system behavior in order to adapt itself to the new conditions.

3 EVALUATION

The experimental evaluation is based on the benchmark collection the .GOV, one of the Web document collections provided by the TREC community. The Vector space model has been used to assign a similarity measure to each document, which estimates how closely it matches a query (i.e., user's information need).

The start Web page set S was made up of random urls extracted from the set of documents satisfying the query. The explorations stopped after having visited 10000 pages.

The rate at which relevant pages are acquired is the most important evaluation factor for the intelligent crawler evaluation, as pointed out in [2]. For this scope, the score means of the last analyzed resources were estimated during each testing exploration. The observation that can be derived from the exploration results showed in Fig. 1, is the increase of the rate of relevant page acquisition when the proposed search system is employed (Fig. 1b), against a system based on a conventional unfocused crawler (Fig. 1a). In other words, the system was able to lead the exploration towards relevant pages, choosing the correct routes, reducing the time required to download and analyze the other resources.

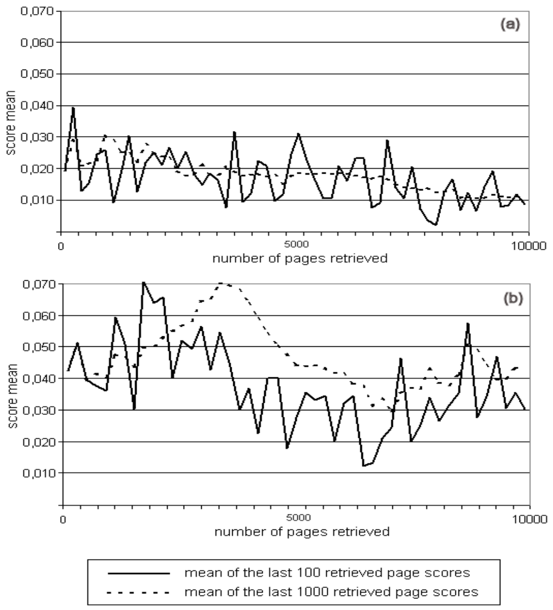

Figure 1. Rate at which relevant pages are acquired for the query '*employment statistics*', measured by the score mean of the last 100 and 1000 retrieved resources: with an unfocused crawler (a) and with the proposed search system (b).

4 CONCLUSION

We have proposed a reactive agent architecture which draw inspiration from the biological researches on the ant foraging behavior.

The result evaluations from a standard test collection demonstrate the architecture's effectiveness when it is exploited by an adaptive search system in the Web domain. The robustness against environmental alterations and the adaptivity to the user's information need changes, are other important features of this architecture. The autonomous search for user relevant resources is based on a common similarity measure, therefore the widespread IR techniques can be easily employed, obtaining personalized search systems, which provide help in the process of finding and organizing information.

Beside the achieved results, this work wills to prove how theories and architectures of agent and multi-agent systems can successfully be applied in huge and dynamic hypertextual domains, providing further tools for the information search task.

REFERENCES

[1] E. Bonabeau, M. Dorigo, and G. Theraulaz, 'Inspiration for optimization from social insect behavior', *Nature*, **406**, 39–42, (2000).

[2] S. Chakrabarti, M. Van Den Berg, and B. Dom, 'Focused crawling: a new approach to topic-specific Web resource discovery', *Computer Networks (Amsterdam, Netherlands: 1999)*, **31**(11–16), 1623–1640, (1999).

[3] F. Menczer and R. Belew, 'Adaptive retrieval agents: Internalizing local context and scaling up to the web', *Machine Learning*, **31**(11–16), 1653–1665, (2000).

[4] Y. Mizuuchi and K. Tajima, 'Finding context paths for web pages', in *Proc. of 10th ACM Conference on Hypertext and Hypermedia*, pp. 13–22, Darmstadt, Germany, (1999).

[5] C.C. Yang, J. Yen, and H. Chen, 'Intelligent internet searching agent based on hybrid simulated annealing', *Decision Support Systems*, **28**, 269–277, (2000).