An Endogenous Fitness Paradigm for Adaptive Information Agents

Filippo Menczer
Cognitive Computer Science Research Group
Computer Science and Engineering Department
University of California, San Diego
La Jolla, CA 92093-0114 USA
fil@cs.ucsd.edu

Wolfram Willuhn
Communication Technology Lab
Image Science Group
ETH-Zentrum, ETZ F86
8092 Zurich, Switzerland
wolfram@vision.ee.ethz.ch

Richard K. Belew
Cognitive Computer Science Research Group
Computer Science and Engineering Department
University of California, San Diego
La Jolla, CA 92093-0114 USA
rik@cs.ucsd.edu

Abstract. We propose a model, inspired by recent artificial life theory, applied to the problem of retrieving information from a large collection of documents such as the World Wide Web. A population of agents is evolved under density dependent selection for the task of locating information for the user. The energy necessary for survival is obtained from both environment and user in exchange for relevant information. By competing for energy, the agents robustly adapt to their environment and are allocated to efficiently exploit their shared resources.

1 Introduction

The World Wide Web (WWW) is an information environment made of a very large distributed database of heterogeneous documents, using a wide-area network and a client-server protocol. The structure of this environment is that of a graph, where the nodes (documents) are connected by hyperlinks. The typical strategy for accessing information on the WWW is by navigating across documents through hyperlinks, retrieving the information of interest along the way. However, due to the *dynamic* and *distributed* nature of the environment, retrieving specific information is a hard task [De Bra and Post 1994, McBryan 1994]. On the other hand, the WWW represents an ideal environment in which to apply techniques recently matured in the field of artificial life (ALife). In particular, *adaptive* and *distributed* algorithms seem to appropriately capture the complexity of such an environment. In this paper we propose a model inspired by the *endogenous fitness* metaphor to search for relevant information through a population of intelligent agents evolving in the WWW environment.

2 Background

The traditional solution to the problem of information retrieval (IR) on the WWW is to build a database index of all the documents, on which to use traditional search and retrieval techniques. Such virtual libraries are built and updated off-line, either in a user-driven fashion or by automated exhaustive programs, called *spiders* or *robots.* A well-known robot is the WWW Worm [McBryan 1994]. This approach has several drawbacks: first, the size of the WWW makes it increasingly difficult to update virtual libraries without inefficient use of network resources. Moreover, any database has to abstract away important information, concerning both the content of documents and their structure [Belew 1985]. Finally, the information gathering and retrieval processes are independent and therefore feedback from the latter cannot be used to adaptively improve efficiency nor quality of the former.

To remedy some of these problems, the client-based Fish Search algorithm has been proposed [De Bra and Post 1994]. This approach uses the metaphor of a school of fish, where agents in a population survive, reproduce, and die based on the energy gained from their performance in the retrieval task. This approach, however, stops short of applying the endogenous fitness paradigm to solve the remaining problems: no mutation occurs at reproduction, and each agent in the population follows a non-adaptive, exhaustive depth-first search algorithm. While some heuristics are used to order the graph traversal, the lack of intelligent cutting of search branches results in slow speed and high network consumption (caching being proposed as a remedy).

The idea of adaptive IR is not new [Belew 1989]. More recently, learning agents and traditional GAs have been successfully applied to information filtering [Maes and Kozierok 1993, Sheth and Maes 1994]. Work in progress indicates that learning is sped up by extending such models to collaborative multi-agent systems [Lashkari, Metral, and Maes 1994]. Using a distributed population of cooperating best-first search agents has been recently proposed in the WebAnts project [Mauldin and Leavitt 1994] to overcome the single-server and single-client bottlenecks.

3 Information search by endogenous fitness

Endogenous fitness models are becoming an increasingly appreciated and well-understood paradigm in the ALife community [Mitchell and Forrest 1994]. While a thorough discussion of the subject is out of the scope of this paper [see, eg, Menczer and Belew 1994], we outline in this section the main aspects that make this paradigm useful for the problem of IR in the WWW.

A population of agents becomes evolutionary adapted in a dynamic environment by a steady-state GA. *Energy* is the single currency by which agents survive, reproduce, and die, and it must be positively correlated with some performance measure for the task defined by the environment. Agents asynchronously go through a simple cycle in which they receive input from the environment as well as internal state, perform some computation, and execute actions. Actions have an energy cost but may result in energy intake. Energy is used up and accumulated internally throughout an agent's life; its internal level automatically determines reproduction and death, events in which energy is conserved.

Agents that perform the task better than average reproduce more and colonize the population. Indirect interaction among agents occurs without the need of expensive communication, via competition for the shared, finite environmental resources. Mutations afford the changes necessary for the evolution of dynamically adapted agents. This paradigm enforces density-dependent selection: the expected population size is determined by the carrying capacity of the environment. Associating high energy costs with expensive actions intrinsically limits the computational load.

In the heterogeneous environment of the WWW, it is hard to associate a fitness measure with a strategy in general, but it is easy to estimate the results of a strategy applied to a particular query. Different information search and retrieval strategies may be optimal for different queries, just as different behaviors may be optimal for different environments. Only the end results of a search (the retrieved documents) can be evaluated, and the agents identify the relevant information from its correlation with energy [Menczer 1994]. Therefore populations will quickly evolve strategies effective in the different environments specified by both information space and queries. Adaptation means for agents to concentrate in high energy areas of the Web, where many documents are relevant. Each agent's survival will be ensured by exchanging an adequate flow of information for energy.

4 An implementation

We have implemented the endogenous fitness model in a simple prototype of IR system for the WWW. The user provides a query consisting of a set of keywords. A population of agents is initialized with some energy, some random strategy, and some distribution in the Web (eg, uniformly in a default set of starting home pages).

INPUT: n, m, e, s, t

1 initialize population of n agents with m random weights each

2 while *number of agents* > 0 do for all agents:

2.1 get energy $\Delta E = e*R$ where current document's relevance $R =$ *number of occurrences of query words in document*

2.2 if $R > s$ get extra energy from user

2.3 sort new links from document according to estimate $L =$ *sum of occurrences of each query word in document, weighted inversely to the number of headers between the positions of the link and the word in the document*

2.4 take first m links, multiply their estimates by their respective weights, normalize to 1

2.5 choose random link and use it for next document

2.6 subtract 1 energy unit for movement

2.7 if $E < 0$ destroy agent

2.8 if $E > t$ clone agent, split parent's energy with offspring, mutate a random offspring weight by additive noise

Table 1: Algorithm for adaptive information agents.

In order to keep search strategies simple while allowing adaptivity, stochastic selection is used to navigate across hyperlinks. For each cycle, each agent estimates the hyperlinks from the current node to decide which node to visit next. The estimate is based on a fixed matching function of the current document and the user keywords. The links are then sorted accordingly, and each estimate is multiplied by a weight. The weighted estimates are normalized and used as probabilities by a stochastic selector. The weights represent the adaptive part of the agent's search strategy. They evolve by selection, reproduction, and mutation. Different weight vectors can implement search strategies as different as best-first (1,0,...,0), biased random walk (1,...,1), or any middle course.

When a link is selected, the agent traverses it and finds itself in a new document. (To prevent agents from sitting at a favorable place without searching, no agent is allowed to return to a previously visited document.) Any traversal incurs an energy cost. The amount of energy an agent receives by finding a document is determined by its relevance, which in turn is estimated by a precision measure from standard IR theory.

Each agent can also decide to present any document to the user hoping to get bonus energy. This decision can use state information (the current level of internal energy) along with document relevance. The user optionally provides feedback by increasing or decreasing the agent's energy. When energy exceeds a fixed threshold, the agent produces an offspring by cloning with random weight mutations; when energy is reduced below zero, the agent dies.

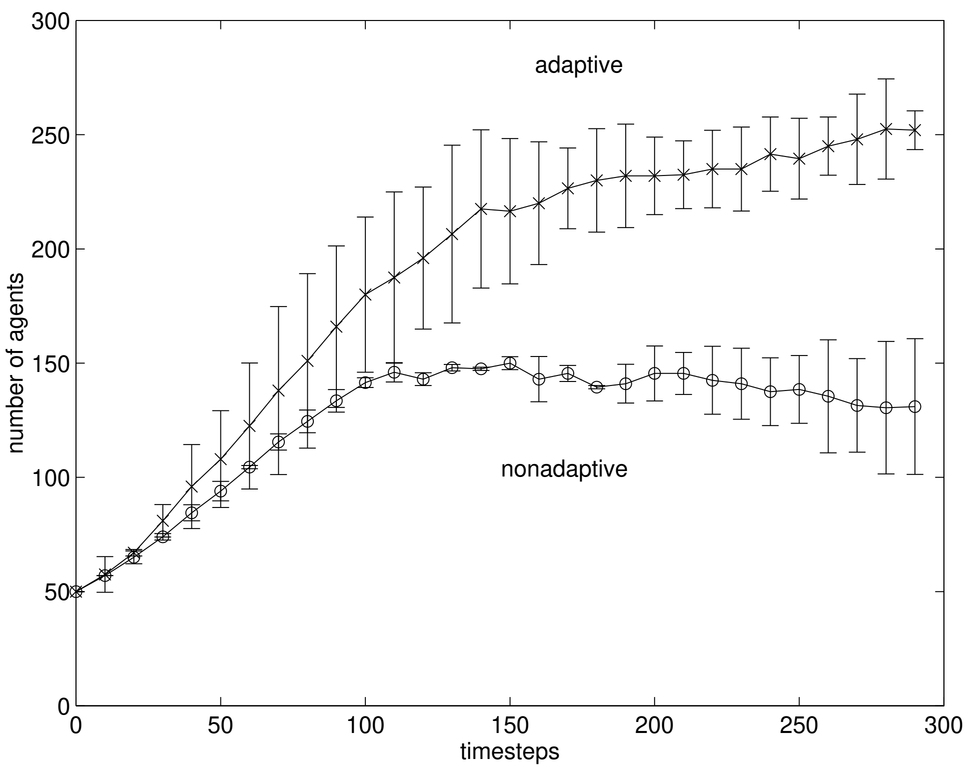

Figure 1: Population dynamics for information agents. Adaptive agents are capable of sustaining a larger population by harvesting more energy from the same environment. The parameters of the simulation are *n=50, m=5, e=3,* $s=\infty$, *t=50.*

At steady-state, the user receives a flow of documents in the form of a list of Web nodes that is updated on-line. The search ends when the population gets extinct, converges according to some measure, or is terminated by the user. The algorithm is illustrated in Table 1.

5 Preliminary results

Preliminary experiments of our system have been carried out on a limited test bed represented by a collection of 116 relatively short documents describing the WWW project: agents can move to any node whose URL starts with http://info.cern.ch/hypertext/WWW/. The total number of links is 178, while 26 of the documents contain a total of 149 query words occurrences.

Given an impossible query (no words in documents match those in query), the environment kills the agents and extinction occurs promptly as expected. Given good queries, the population quickly reaches a stable size determined by the environment's carrying capacity, and a steady-state document retrieval rate from information-rich areas.

We have compared the performance of two endogenous fitness populations of information agents: nonadaptive agents implement best-first search, while the strategy of adaptive agents evolves according to the selective pressures of the search environment. User feedback is not used yet. The results (see Figures 1 and 2) seem quite promising in showing feasibility, robustness, and good quality of retrieved documents.

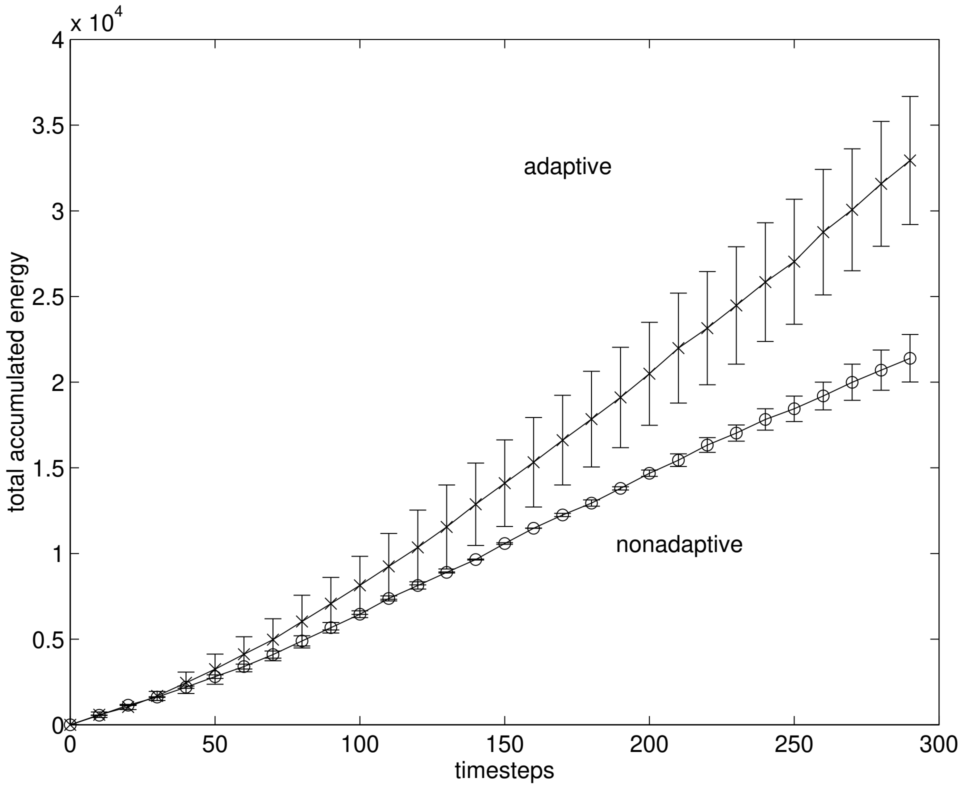

Figure 2: Cumulative energy collected by information agents. Adaptive agents can improve retrieval performance by adjusting their search strategy to the environment. The parameters of the simulation are *n=50, m=5, e=3,* $s=\infty$*, t=50.*

The main observations that we draw from performance measures at steady-state conditions are: (i) the adaptive population reaches a size 38% higher than the best-first population; (ii) energy and document retrieval rates for the adaptive population are as much as 46% higher than those for the best-first population. Since population size measures population fitness in endogenous fitness models [Menczer and Belew 1994], the overall relevance for documents retrieved by the adaptive population is significantly higher than for those of the best-first population.

6 Conclusion

We have illustrated the suitability of ALife modeling for an important real-world application such as intelligent IR in distributed, heterogeneous information environments. Endogenous fitness models, in particular, have been shown to be a natural paradigm within which to evolve populations of adaptive information agents.

The proposed approach overcomes many of the limitations found in existing systems. No index database is built, eliminating the problems of size, server load, and dynamic updating. The search process makes use of the WWW hyperstructure to carry out the search, and dynamically adapts to its changes as well as to the variability due to different users and queries. No overhead is incurred by explicit communication among agents, while competition for shared resources intrinsically enforces a balanced network load. Exhaustive search is overcome by more efficient, adaptive branch cutting in the search space. The selective pressure mechanism, removing agents from low energy zones and allocating new ones to information-rich areas, is supported by theoretical results about optimal on-line graph-search algorithms [Aldous and Vazirani 1990]. Current goals are evaluating how performance scales with search space size, as well as comparisons with alternative search methods.

Many extensions are possible for improved implementations of our model. Including server information in the computation of energy costs allows better exploitation of network resources. Smart convergence measures may speed up the location of information-rich areas in the search graph. Simple reinforcement learning during life may provide faster adaptive changes than evolution alone [Menczer 1994]. Finally, user feedback may drastically improve on-line performance by allowing to shape the adaptive landscape without detailed knowledge about its structure.

References

Aldous D and Vazirani U (1990) A Markovian extension of Valiant's learning model. Proc. 31st Annual Symposium on Foundations of Computer Science, 392-6. Los Alamitos, CA: IEEE Comput. Soc. Press

Belew RK (1989) Adaptive information retrieval: Using a connectionist representation to retrieve and learn about documents. Proc. SIGIR 89, 11-20, Cambridge, MA

Belew RK (1985) Evolutionary decision support systems: an architecture based on information structure. *Knowledge Representation for Decision Support Systems,* ed. by Methlie LB and Sprague RH, 147-160. Amsterdam: North-Holland

De Bra PME and Post RDJ (1994) Information retrieval in the World Wide Web: Making client-based searching feasible. Proc. 1st Intl.World Wide Web Conf., ed. by Nierstrasz O, Geneva: CERN

Lashkari Y, Metral M, and Maes P (1994) Collaborative Interface Agents. Technical report, MIT Media Lab

Maes P and Kozierok R (1993) Learning interface agents. Proc. 11th Natl. Conf. on Artificial Intelligence, 91-99. Los Angeles, CA: Morgan Kaufmann

Mauldin ML and Leavitt JRR (1994) Web agent related research at the Center for Machine Translation. Proc. ACM SIGNIDR 94

McBryan OA (1994) GENVL and WWWW: Tools for taming the Web. Proc. 1st Intl.World Wide Web Conf., ed. by Nierstrasz O, Geneva: CERN

Menczer F and Belew RK (1994) Latent Energy Environments. *Adapting Individuals in Evolving Populations: Models and Algorithms,* ed. by Belew RK and Mitchell M, SFI Studies in the Sciences of Complexity. Reading, MA: Addison Wesley

Menczer F (1994) Internalization of environmental messages in adaptive agents. Unpublished research proposal, Computer Science and Engineering Department, University of California, San Diego

Mitchell M and Forrest S (1994) Genetic algorithms and artificial life. *Artificial Life* 1:3

Sheth B and Maes P (1994) Learning Agents for Personalized Information Filtering. Technical report, MIT Media Lab