Artificial Life Applied to Adaptive Information Agents

Filippo Menczer **Richard K. Belew**

Computer Science and Engineering Department, 0114
University of California, San Diego
La Jolla, California 92093, USA
{fil, rik}@cs.ucsd.edu

Wolfram Willuhn

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

Abstract

We propose a model, inspired by recent artificial life theory, applied to the problem of retrieving information from a large, distributed 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 appropriate information. By competing for relevant documents, the agents robustly adapt to their information environment and are allocated to efficiently exploit shared resources. We illustrate the roles played by document locality, adaptive search strategies, and relevance feedback, in the information gathering process.

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 to navigate across documents through hyperlinks, retrieving the information of interest along the way. The *dynamic* and *distributed* nature of the environment, however, makes retrieving specific information a hard task [De Bra & 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. Traditional genetic algorithms [Mitchell & Forrest 1994] are characterized by exploitation of information, but the distributed information gathering problem requires adaptation rather than optimization. Therefore 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. This paper presents various extensions to the work recently reported in [Menczer, Willuhn & Belew 1994].

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]. The off-line 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 & 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 solving the

* Research partly supported by a fellowship from the Apple Advanced Technology Group

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 palliative).

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

Information Search by Endogenous Fitness

Endogenous fitness models are becoming an increasingly appreciated and well-understood paradigm in the ALife community [Mitchell & Forrest 1994]. While a thorough discussion of the subject is out of the scope of this paper [see, e.g., Menczer & Belew 1995], 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 genetic algorithm. *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 enforces a balanced network load by limiting inefficient uses of bandwidth.

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. Therefore populations will 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.

An Implementation

We have implemented the endogenous fitness model in a simple prototype of IR system for the WWW. The algorithm is illustrated in Table 1. 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. The ideal, zero-knowledge assumption is to start with a population at minimal distance from all nodes. Typical heuristics suggest to initialize the population with a uniform distribution in a default set of known starting points, or better yet in the documents returned by a preliminary call to a traditional search engine.

```
INPUT: $n>0, e>0, t>0$; query word(s)

1    initialize population of $n$ agents with random $\beta, \gamma, E$
2    while number of agents > 0 do for all agents:
2.1      compute for each link $i$ in current document the
         estimate $L_i$ = sum of occurrences of each query
         word in document, weighted inversely to the
         number of headers between the positions of i and
         the word in the document

         pick link $i$ to follow according to the probability
         distribution $P_i = \exp(\beta L_i) / \sum_j \exp(\beta L_j)$
2.2      $E \leftarrow E - N + eR$ where $N$ = server access cost
         measure and current document's relevance $R$ =
         number of occurrences of query words in
         document

         if $(R+1)\gamma > \frac{1}{2} \Rightarrow E \leftarrow E + F$ where $F$ = user
         feedback energy (optional)
2.3      if $E > t$ clone agent, split parent's energy with
         offspring, mutate offspring's $\beta, \gamma$

         else if $E < 0$ destroy agent
```

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 estimates, based on a fixed matching function of the current document and the user keywords, are scaled by a non-decreasing function to obtain a probability distribution that is in turn used by a stochastic selector. The slope of the non-linear scaling function is determined by the agent's genetic parameter β This trait represent the adaptive part of the agent's search strategy. It evolves by selection, reproduction, and mutation. Different β values can implement search strategies as different as best-first ($\beta=\infty$), random walk ($\beta=0$), 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. Ideally, the cost should be a function of the load imposed by the access on network resources, ultimately affecting search time. 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 is based on the relevance of the document and is biased by another genetic parameter, γ, determining the likelihood with which a document is considered interesting enough to be presented to the user. The extreme cases are when all documents are presented ($\gamma=1$) or none ($\gamma=0$). The user optionally provides feedback by increasing or decreasing the agent's energy. This is a natural model of relevance feedback, where the user can effectively modify the adaptive landscape with only incomplete knowledge of the search space.

When energy exceeds a fixed threshold, the agent produces an offspring by "local cloning." The genetic parameters undergo random mutations, and energy conservation is enforced by parents splitting their energy with offspring. When energy is reduced below zero, the agent dies. A possible variation of the algorithm in Table 1 is obtained if step 2.3 is moved ahead of 2.2: this is effectively equivalent to cloning with one-step "lookahead" [Menczer, Willuhn & Belew 1994]. We show in the next section that this variation actually results in a deterioration of performance, due to the loss of locality.

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.

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 query words. The graph corresponding to the test bed is shown in Figure 1. The fact that this collection is closed to the rest of the WWW is only one of its limitations.

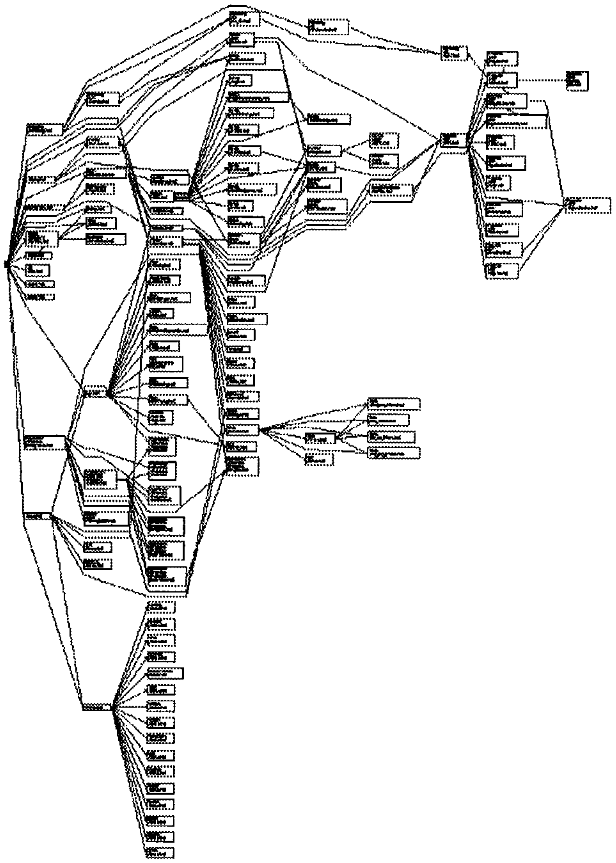

Figure 1: WWW subgraph used as a test bed in the experiments.

Given an impossible query (no words in documents match those in the query), the environment kills the agents and extinction occurs promptly, as expected. Otherwise, the population quickly reaches the environment's carrying capacity (determined by the distribution of query word occurrences in the collection) and a steady-state document retrieval rate from information-rich areas. These results hold over a wide range of simulations and seem quite promising in showing feasibility, robustness, and good quality of retrieved documents.

In the previous section we have mentioned two alternatives of our endogenous fitness algorithm, namely local cloning (cf. Table 1) and lookahead cloning. Figure 2 illustrates the superiority of the former. The increase of over 200% in the rate of collected energy demonstrates the importance of the search graph topology for effective information gathering.

In Figure 3 we have plotted the size of four populations of information agents to compare the performances of local vs. lookahead cloning as well as adaptive vs. nonadaptive (best-first) agents. Note that population size is an appropriate measure of population fitness in endogenous fitness models [Menczer & Belew 1995]. Once again local cloning results in a large performance improvement. The search strategy of adaptive agents can adjust according to the selective pressures of the information environment. However, adaptive populations score significantly better than nonadaptive ones only in the simulations with lookahead cloning. This suggests that local adaptation is particularly advantageous when less locality is preserved during the search process.

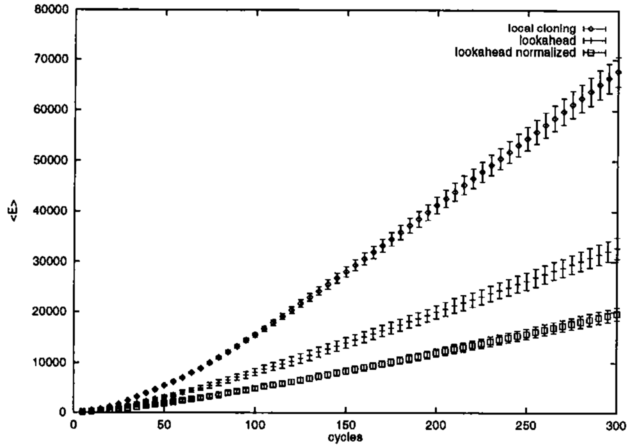

Figure 2: Cumulative energy collected by information agents. Given the same environment, agents with local cloning harvest more energy than those reproducing with one-step lookahead. The normalized curve is scaled to correct a difference in the e parameter across experiments. In this and the following plots, error bars correspond to ±1 standard deviation over repeated simulation runs.

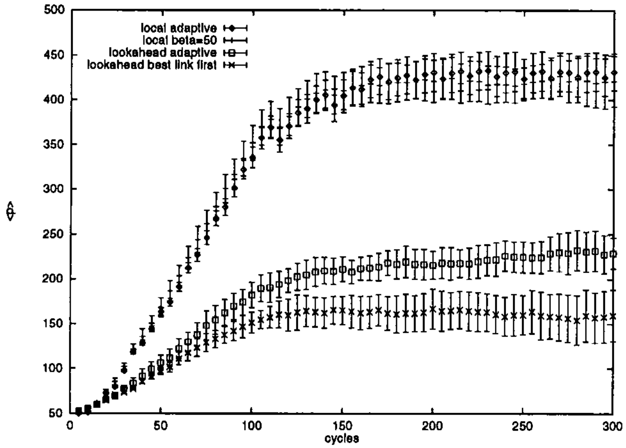

Figure 3: Population dynamics for adaptive and best-first information agents. The β parameter (cf. Table 1) evolves in the unit interval for adaptive populations, while the large fixed β=50 value implements a nonadaptive, pseudo-best-first strategy. Agents with local cloning can afford larger population sizes, but adaptive search exhibits a significant advantage over exploitative strategies only for agents with lookahead cloning.

Finally, user feedback is tested in simulations whose results are shown in Figure 4. Five particularly relevant documents are identified manually and assigned positive F values (cf. Table 1); the rest are given negative F so that our simulated user actually decreases the environment's carrying capacity, on average. Therefore the observed increase in population size confirms that the user, without any knowledge of information space topology, can use relevance feedback to alter the selective pressure and significantly improve performance.

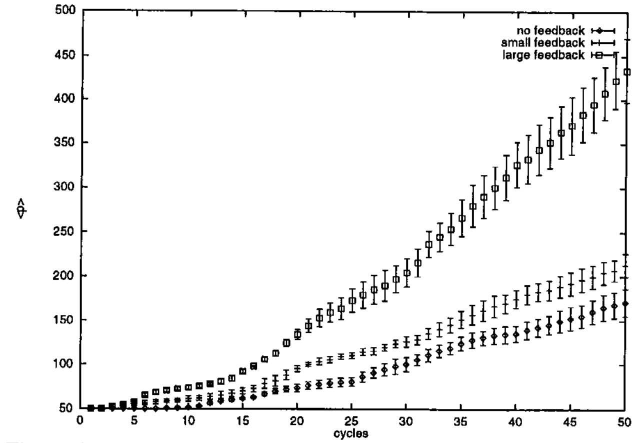

Figure 4: Population size for adaptive information agents with and without user relevance feedback. In the former cases, the parameter γ (cf. Table 1) is set to 0.2. The user interactively modifies the adaptive environment, thus accelerating the discovery of relevant documents.

Extensions

Many extensions are possible for improved implementations of our model. The experiments reported in the previous section simulate on-line access in order to avoid network load. Several options (AppleScript, Tcl, SodaBotL [Coen 1994]) are being considered for the implementation of actual on-line agents.

Better measures for document precision and link estimates, using for example cosine normalization, inverse frequency weighting, and word proximity, are under study. Local caching and back links are other additions likely to be incorporated in future implementations.

Simple reinforcement learning during life may provide faster adaptive changes than evolution alone [Menczer & Belew 1994]. Learning local characteristics about the search space should complement the slower process of genetic adaptation.

Another set of extensions depends on inter-agent communication. Smart convergence measures, crossover, shared caching, learning from other agents, and other forms of interaction, may all speed up the location of information-rich areas in the search graph. However, any advantages must be traded off with the incurred communication costs.

Including server information in the computation of energy costs allows a more efficient exploitation of resources and therefore decreases bandwidth waste. However, in the long run, we believe that network access will become a serious bottleneck for any client-based distributed search. The answer, of course, is to transfer agents from clients to servers. While many well-founded concerns make this solution unfeasible at present, we imagine that in the very busy network of a near future, the owner of a server might become willing to give up some CPU cycles on its machine in exchange for improved bandwidth. Transportable agents [Kotay & Kotz 1994] would then access the server documents locally, and only transfer relevant information back to the client.

Conclusions

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 approach we have proposed overcomes many of the limitations found in existing systems. No index database is built, eliminating the problems of size, server load, and dynamic updating.

We have shown locality to be important for distributed information gathering, and the use of the WWW hyperstructure has been made an essential feature of our model. The search process dynamically adapts to changes in the information environment, as well as to the variability due to different users and queries. Exhaustive search is overcome by more efficient, adaptive branch cutting in the search space. The model, making only minimal assumptions about the structure of the adaptive landscape, allows the user to easily improve on-line performance.

The selective pressure mechanism, removing agents from low energy zones and allocating new ones to information-rich areas, is connected to theoretical results about optimal on-line graph-search algorithms [Aldous & Vazirani 1994, Deng & Papadimitriou 1990]. We are currently working on a rigorous proof to link algorithmic complexity and expected performance.

Since communication is the bottleneck of any distributed algorithm (and even more so for client-based, on-line search), the problem addressed in this paper is well characterized by the need to limit communication among agents to its minimum. The endogenous fitness algorithm allows to achieve this goal in a natural way, because no overhead is incurred by explicit communication among agents. Density dependent selection occurs by way of competition for shared environmental resources; no ranking of the population is required. It should be noted, however, that certain "hidden" communication costs cannot be avoided. This is the case, for example, in order to make retrieved documents "disappear" and avoid redundant access. Caching and efficient data structures are being considered to minimize such hidden costs.

Many other directions remain open for further work. Present goals include analyzing evolved genetic parameters, evaluating how performance scales with search space size, and comparing our algorithm with existing search methods.

References

Aldous D and Vazirani U 1994. "Go With the Winners" Algorithms. Proc. 35th Annual Symposium on Foundations of Computer Science, 492-501. 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

Coen MH 1994. SodaBot: A Software Agent Environment and Construction System. Proc. CIKM'94 Workshop on Intelligent Information Agents. Gaithersburg, MD

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 Conference, ed. by Nierstrasz O. Geneva: CERN

Deng X and Papadimitriou CH 1990. Exploring an unknown graph. Proc. 31st Annual Symposium on Foundations of Computer Science, 355-361. Los Alamitos, CA: IEEE Comput. Soc. Press

Kotay KD and Kotz D 1994. Transportable Agents. Proc. CIKM'94 Workshop on Intelligent Information Agents. Gaithersburg, MD

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

Maes P and Kozierok R 1993. Learning interface agents. Proc. 11th AAAI Conference, 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 Conference, ed. by Nierstrasz O. Geneva: CERN

Menczer F and Belew RK 1995. Latent Energy Environments. *Adaptive Individuals in Evolving Populations: Models and Algorithms,* ed. by Belew RK and Mitchell M. Reading, MA: Addison Wesley

Menczer F and Belew RK 1994. Evolving sensors in environments of controlled complexity. *Artificial Life IV,* ed. by Brooks R and Maes P, 210-221. Cambridge, MA: MIT Press

Menczer F, Willuhn W, and Belew RK 1994. An Endogenous Fitness Paradigm for Adaptive Information Agents. Proc. CIKM'94 Workshop on Intelligent Information Agents. Gaithersburg, MD

Mitchell M and Forrest S 1994. Genetic algorithms and artificial life. *Artificial Life* 1(3):267-289

Sheth B and Maes P 1993. Evolving Agents for Personalized Information Filtering. Proc. 9th IEEE Conference on AI for Applications

Yang J and Korfhage RR 1993. Query Optimization in Information Retrieval Using Genetic Algorithms. Proc. 5th ICGA, 603-611. Urbana, IL