A Multi-Agent Simulation Framework for Spiders Traversing the Semantic Web

Christos Dimou, Alexandros Batzios, Andreas L. Symeonidis and Pericles A. Mitkas
Dept. of Electrical and Computer Engineering,
Aristotle University of Thessaloniki, Greece
Email: {cdimou, alex, asymeon}@issel.ee.auth.gr, mitkas@eng.auth.gr

Abstract

Although search engines traditionally use spiders for traversing and indexing the web, there has not yet been any methodological attempt to model, deploy and test learning spiders. The flourishing of the Semantic Web provides understandable information that may improve the accuracy of search engines. In this paper, we introduce BioSpider, an agent-based simulation framework for developing and testing autonomous, intelligent, semantically-focused web spiders. BioSpider assumes a direct analogy of the problem at hand with a multi-variate ecosystem, where each member is self-maintaining. The population of the ecosystem comprises cooperative spiders incorporating communication, mobility and learning skills, striving to improve efficiency. Genetic algorithms and classifier rules have been employed for spider adaptation and learning. A set of experiments has been performed in order to qualitatively test the efficacy and applicability of the proposed approach.

1. Introduction

Today's search engines use bots, commonly referred to as *spiders* or crawlers, in order to traverse and index the Web. While spiders are extensively used, there are no automated or semi-automated mechanisms for implementing, maintaining and testing the performance of intelligent Web Spiders. Within the context of the current work, we propose an agent-based simulation framework for designing and testing spider entities that operate in a semantically enhanced web environment. *BioSpider* adopts ecosystem rules that apply on a set of autonomous cooperating entities that collect semantic data from a simulated web in order to be incorporated in future semantically-aware search engines. This simulation framework is envisioned to help researchers apply various spidering techniques and study their effectiveness on semantically aware web content.

This paper is structured as follows: Section 2 reviews the background work; in Section 3 we outline BioSpider and its components; Section 4 illustrates an experimental evaluation; we conclude in Section 5 with remarks on future directions.

2. Background Work

Web spidering or web crawling is the automated and systematic traversing and indexing of web pages for subsequent search purposes. Since the early simple crawlers (e.g. [1]), there have been considerable advances, including scheduling and indexing integration into crawlers. The latest generation of crawlers focuses on the distribution of load to a population of cooperating spiders (e.g. [2]) and on the so-called 'focused crawling', the collection of pages related to given keywords, topics or other web pages (e.g. [5])

We approach the web spidering simulation problem, by adopting biology-inspired simulation techniques, that recently have drawn the attention of both biology and computer science researchers. Existing efforts in the literature vary from building individual based models that are focused on the environmental aspects of the system, to multi-species communities that consider both ecological and evolutionary dynamics and corresponding learning processes employed in society evolution ([3], [7]). Krebs and Bossel [4] developed the 'animat', a computational model of an artificial animal that lives and evolves in unknown environments. An improved version of this 'animat', has been introduced in the context of Biotope [8], a configurable tool for designing distributed computing simulation environments.

BioSpider employs Biotope's modelling infrastructure for designing a web spider-specific simulation environment. We argue that this work contributes in the following aspects:

1. Deployment of a configurable simulation framework for spiders traversing the Semantic Web.
2. Introduction of semantically enhanced simulated web content.
3. Study of evolutionary learning on cooperating web spiders in uncertain environments.

Proceedings of the 2006 IEEE/WIC/ACM International Conference
on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)
0-7695-2747-7/06 $20.00 © 2006 IEEE

4. Provision of a communication framework that assists spiders in balancing their exploration load.

3. Overview of BioSpider

BioSpider's overall architecture consists of a semantically enabled search engine employing a certain number of autonomous spiders that crawl the web by following links on web pages. Spiders send indexed information back to the search engine and exchange information with other spiders in order both to reduce workload and to avoid traps or sites with illegal or unwanted content. BioSpider adopts an ecosystem paradigm, according to which spiders resemble to living entities which are equipped with energy, vision, mobility, a knowledge model and communication abilities. Spiders are motivated to find useful content in their search so as to maintain a sufficient energy level, under a penalty-reward rationale.

3.1. Simulation Environment

The actual space that spiders move has been modelled as a $x \times y$ rectangular grid, each cell of which is considered to be a single web page. Transition paths are marked between cells, since each web page may contain a set of links pointing to other web pages. Following the terminology introduced in [8], each cell may be in one of the states that are summarized in Table 1.

3.2. Agent Model

As already stated, BioSpider follows the 'animat' paradigm. Spiders have therefore been equipped with certain abilities, namely: a) vision, b) movement, and c) reproduction ability.
a) Vision: We consider a $n \times n$ vision area for any spider. Vision models a 'peeking' procedure: a spider may peek $n \times n\text{-}1$ outgoing links each time in order to guess if they lead to trap or obstacle cells. The results of this procedure are stored in a vision vector. Finally, we introduce a p vision error probability, that models expected peeking errors and is further explained in Section 4.

Table 1. Cell content for BioSpider

Cell value	Cell content	BioSpider term
0	Vacant Space	No semantic content
1	Resource Enrichment	Semantic content
2	Resource Reduction	Link farm - spam page
3	Incompatibility	robots.txt
4	Agent	Spider

b) Movement: A spider is generally free to move in any direction of the grid, by following links stored in their vision vector. The decision for the motion path is determined on the basis of a set of rules; a shortest path algorithm, decides which movement rule is the strongest, while avoiding traps and obstacles.

c) Reproduction: Reproduction may occur, if spider energy exceeds a certain threshold. This natural procedure is adapted to BioSpider architecture for load balancing purposes, since the offspring inherits its parent's knowledge model, as well as an initial amount of energy. Factors such as the resulting energy levels of parents and offsprings and the percentage of the inherited classifier set adhere to the corresponding dispersal theory primitives.

3.3. Communication

Spiders need to communicate with each other in order to exploit the already recorded information about the content of the visited pages, as well as the optimal routes in their vision. In the case that both spiders decide to move on the same web page, the spider with the lower amount of energy reconsiders its route and visits another page, for load balancing purposes.

3.4. Knowledge Model

The most important feature of the BioSpider system is the ability of spiders to augment their intelligence. Spiders need to be able to evaluate their decisions and adapt their knowledge model in real time.

The learning mechanism that spiders use comprises three parts. First, there is a set of classifiers, forming the knowledge model. Classifiers are effective rules that agents use to make decisions. After any move, agents compare their current vision vector to the underlying rule base.

The classifiers are evaluated by a classifier evaluation mechanism, as spiders gain more experience about their environment. Classifiers are evaluated using a modified version of the bucket brigade algorithm. Every time a classifier leads the spider to change, the classifier's strength increases.

Finally, a genetic algorithm is employed to ensure that only the most appropriate classifiers survive. This increases the variety of classifiers in the knowledge base, by a "rejuvenation" of the spider mentality.

4. Experiments

The simulation environment for BioSpider has been implemented on the Java v1.5 language and the JADE environment, adhering to the FIPA specifications. Moreover, the Agent Academy framework [6] has been adopted in order to model and implement the knowledge model of the agents.

We embark our experimental testing of the proposed system by defining a series of *environmental* and *spider performance* assessment indicators that allow users to monitor the effect of changes in spider behavior.

4.1. Environmental indicators

1. *Resource Availability* is defined as the ratio of the total semantically enhanced web pages *wp* to the total harvesting distance d_{wp}.
2. *Environmental Variety* is defined as the ratio of distinct vision vectors *vv* to the occupied cells of the grid *oc*.
3. *Environmental Reliability* is defined as the complement of the vision error probability *p*.
4. *Content Value Dispersal* is a Gaussian random variable (GRV) that indicates the wealth of the content provided by the explored web pages. This indicator assigns a different energy unit reward for every available page, in contrast to [8] where all food cells are assigned with a static energy unit amount.

4.2. Spider performance indicators

1. *Energy*. The amount of energy for each spider, measured in ecosystem epochs, either increases on visiting useful web pages or decreases on visiting semantically unaware pages, spam sites or incompatible pages. When the energy reaches zero, the spider seizes to exist.
2. *Aging*. After reaching a certain age, the energy loss rate increases for any activity, as an ecosystemic way of balancing spider aging.
3. *Effectiveness, e*. A metric of effectiveness defined in terms of energy uptake rate *eur*, energy loss rate *elr* and energy availability rate *ear*: $1 + \frac{eur-elr}{ear}$
4. *Content exploration rate, cer*. The ratio of the total number of distinct semantically enhanced pages that a spider has explored $c_{explored}$ to the total number of moves of the spider *s*.
5. *Trap collision rate, tcr*. The ratio of the total number of traps that a spider has collided upon $t_{collided}$ to the total number of moves of the spider *s*.
6. *Unknown state rate, usr*. The ratio of the total number of unknown states *unknownState* to the total number of moves of the spider *s*. An unknown state is reached if no classifier rule matches the current vision vector of the spider.
7. *Reproduction rate, rr*. The ratio of the total number of a spider's *offspring* to the total number of moves of the spider *s*.
8. *s-Throughput*. The ratio of the total energy units that a spider has earned to the total number of moves of the spider *s*.

4.3. Results

The experimental phase of the proposed work focuses on the identification of the optimal parameters and their impact on the proposed assessment indicators. The following two different series of experiments were conducted, in order : a) to determine the near-optimal values for several environmental, communication or learning parameters with respect to the maximization of the effectiveness indicator, and b) to measure the effect of the above parameters to the overall throughput of semantically enhanced content to the parent search engine. For both configurations, a grid of 40×40 has been employed, with 10 agents at the initialization stage, each having a 5×5 vision vector.

4.3.1 *Maximizing effectiveness e*

A set of nine experiments were performed in this series. In Table 2, these experiments are denoted with E_{Ai}. *Genetic Algorithm Step* and *Communication Step* are expressed in ecosystem epoches (i.e. simulation steps), whereas *Knowledge Base Size* is expressed in number of stored rules for each spider. Table 3 illustrates the average indicator values for these experiments.

Some of the key observations include the following: a) Increasing the communication rate leads to overall better effectiveness ($E_{A1} - E_{A2}$). However, if the communication is too frequent (E_{A3}), no new significant rules can be produced. b) A larger knowledge base size enhances the effectiveness of spiders (E_{A3}, E_{A4}), even at higher levels of uncertainty (E_{A9}). c) Tuning the step of the genetic algorithm in uncertain environments critically increases the effectiveness of the spiders (E_{A6}, E_{A8}).

4.3.2 *Examining system performance*

In the second series, four experiments were performed. The corresponding parametrization for the experiments is denoted as *(*E_{Bi}*)* in Table 2. Having determined a diverse set of values for each experiment, we conduct a comparison over the proposed indicators. Experiments E_{B1} to E_{B4} were chosen as a testbed for comparing the actual efficacy of the spiders under highly volatile and uncertain environments.

Table 2. Parameters for the Experiments

Experiment	Genetic	Communication	Knowledge	Vision Error
	Algorithm Step	Step	Base size	Rate %
E_{A1}	250	1000	3000	5
E_{A2}	250	200	3000	5
E_{A3}	250	50	3000	5
E_{A4}	250	200	1000	5
E_{A5}	250	200	3000	5
E_{A6}	1000	1000	1500	30
E_{A7}	250	1000	1500	30
E_{A8}	50	1000	1500	30
E_{A9}	50	1000	1500	30
E_{B1}	250	1000	1500	5
E_{B2}	250	1000	1500	5
E_{B3}	250	500	3000	5
E_{B4}	50	500	1500	30

Table 3. Average Indicator Values

Experiment	*e*	*cer*	$trc \times 10^3$	*usr*	$rr \times 10^3$
E_{A1}	1.459	0.135	4.953	0.346	0.309
E_{A2}	1.641	0.132	5.107	0.418	0.288
E_{A3}	1.494	0.126	5.365	0.441	0.262
E_{A4}	1.632	0.139	6.280	0.399	0.298
E_{A5}	1.318	0.128	14.773	0.472	0.194
E_{A6}	1.247	0.115	9.935	0.460	0.162
E_{A7}	1.403	0.119	10.079	0.474	0.182
E_{A8}	1.500	0.113	9.673	0.559	0.225
E_{A9}	1.542	0.113	11.672	0.615	0.273

These experiments indicated that the Genetic Algorithm is perhaps the single most important element of BioSpider. As we increased the level of uncertainty and randomization in the system, in order to try to reflect the dynamic and continuously changing nature of the Web, we noticed that decreases in the Genetic Algorithm's step provided a significant improvement in the average throughput of Web spiders. Indeed, uncertainty and randomization reflect frequent changes in the environment, which can be better handled by newer and more frequent generations of Spiders. This, after all, is also the case in real life, where species that reproduce frequently almost always show increased adaptability compared to species with slower rates of reproduction.

5 Conclusions

Parsing and indexing the ever-changing Web in an efficient manner has always been a major challenge. Inspired by the similarities between the complex biological ecosystems and the nature of the Web, we developed an integrated infrastructure for creating and training Web Spiders for their quest for bandwidth and information exploitation. Most current approaches do not pay the proper attention to the parametrization of the communication framework and the abilities of self-organization and, thus, adaptability.

BioSpider enhances spider intelligence by incorporating elements from the fields of classifier systems, genetic algorithms and dispersal distance evolution. The series of experiments conducted with BioSpider prove the added-value of the framework in the areas of communication, self-organization and overall performance of a population.

References

[1] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1):107–117, 1998.

[2] J. Edwards, K. McCurley, and J. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. In *Proceedings of the Tenth Conference on World Wide Web*, pages 106–113, 2001.

[3] J. Haefner and T. Crist. Spatial model of movement and foraging in harvester ants (pogonomyrmex) (i): The roles of memory and communication. *Journal of Theoretical Biology*, 166:299–313, 1994.

[4] F. Krebs and H. Bossel. Emergent value orientation in self-organization of an animat. *Ecological Modelling*, pages 143–164, 1996.

[5] F. Menczer, G. Pant, P. Srinivasan, and M. Ruiz. Evaluating topic-driven web crawlers. In *Proceedings of the 24th conference of research and development in information retrieval*, pages 241–249. ACM Press, 2001.

[6] P. Mitkas, D. Kehagias, A. Symeonidis, and I. Athanasiadis. A framework for constructing multi-agent applications and training intelligent agents. *Agent Oriented Software Engineering IV, P. Giorgini, J. P. Mueller and J. Odell, (eds.), LNCS 2935: Springer-Verlag*, pages 96–109, 2003. Software available at: *http://sourceforge.net/projects/agentacademy*.

[7] S. Pecala. Neighborhood models of plant population dynamics. *Multispecies models of annuals*, 29:262–292, 1986.

[8] A. L. Symeonidis, V. Valtos, S. Seroglou, and P. Mitkas. Biotope: an integrated simulation tool for augmenting the intelligence of multi-agent communities residing in hostile environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part A, Special Issue on Self-organization in Distributed Systems Engineering*, 35(3):420–432, 2005.