# Population Control for Multi-agent Based Topical Crawlers

Alban Mouton, Pierre-François Marteau

# Population control for multi-agents based topical crawlers

Alban Mouton
VALORIA
University of South Britanny
Vannes, France
Email: alban.mouton@univ-ubs.fr

Pierre-Francois Marteau
VALORIA
University of South Britanny
Vannes, France
Email: pierre-francois.marteau@univ-ubs.fr

***Abstract*—The use of multi-agents topical crawlers based on the model of endogenous fitness introduces the problem of population control. We propose to harness this point through an energy based model to balance the reproduction/life expectency of agents. Our goal is to ease the tuning of parameters and to optimize the use of available ressources for the crawling. We introduce a model based on energy designed to control the ratio number of agents over precision of the crawling. We present some experiments that show that the size of the population remains under control during the crawling.**

## I. Introduction

The interests of population based multi-agents models for focused crawling on the Web have already been introduced in previous works [1]. Most systems based on artificial life involve populations of agents which development is difficult to harness: the consequence is mainly an increase of necessary ressources for an efficient processing. Particularly in the topic of web crawling, Internet offers a very wide environment to explore: this last point drives the need of an efficient strategy of population control.

The aim of this article is to propose automatic regulation techniques to the endogenous fitness model applied to focused crawling. Firstly we re-define the notion of energy in the scope of the endogenous fitness model: We want to simplify the parameters of the model and to automatically preserve a viable population of agents. Secondly we introduce a notion of conservation of the energy on crawled pages that improves furthermore our control of the size of the population of agents and its activity. We then present some experiments that detail the behaviour of our model and shows its effectiveness.

## II. Related works

### A. Focused crawling and artificial life

The notion of *focused crawling*, also known as topical or topic-driven crawling, was first introduced by Chakrabarti & al. [2] to refer to Web crawlers specialized in topic-specific information discovery. Focused crawling raises many interesting topics. For example it generally leans heavily on the vast field of automatic text processing initiated by Salton [3] for topical judgments. Automatic extraction of link contexts in semi-structured documents is another recurrent problem addressed notably by Pant & al. [4]. The study of the topology of the topical graphs in the Web has consequences on focused crawling and free exploration of Web pages or exploitation of already discovered resources [5]. Interesting works attempt to formulate the numerous variants involved in the design of this kind of system and to propose suited evaluation frameworks and metrics to research teams working in the field [6] [7].

Quite early, some artificial life models were studied for developping efficient topical Web crawlers. Firstly, Menczer & al. [8] proposed to apply the endogenous fitness artificial life model to focused crawling. They also studied the interesting association of multi-agents models with contextual machine learning mechanisms and showed that adaptivity of agents could scale up to the size of the environment through cloning and deployment of the population [1]. This work led to the projects InfoSpiders and MySpiders [9]. Other approaches based on different models were proposed, some in particular were built upon the *ant colony paradigm* [10] [11]. Kushchu [12] proposed a survey of evolutionary and adaptive approaches applied to Information Retrieval on the Web.

### B. The core of the original Endogenous Fitness model for focused crawling

In this paper, we focus on the endogenous fitness model applied to focused crawling as described by Filippo Menczer [8]. The core of this model is presented in algorithm 1. In this multi-agents model agents independently visit Web pages and pick new links searching for relevant information. Agents clone themselves or die according to their success that is evaluated by converting relevant information into energy.

## III. New strategy for population control in the endogenous fitness model

### A. Energy balance

Control of the system through its energy parameters $C$ and $\rho$ is a quite difficult task to achieve. Depending on the values of these parameters the population of agents can undergo important variations of size that make resource consumption unpredictable. Excessive variations of the population of agents can also decrease the quality of the results because too many agents may imply a dispersed crawler that is insufficiently selective.

**Algorithm 1** Endogenous fitness based crawler

```
Initialize system with parameters:
- energy cost and gain rate : $C$ and $\rho$
- number of agents : $N$
- seeds pages : $S$
- query $Q$
Initialize $N$ agents with energy $E = 1$ and current document $D$ in $S$
loop
  For each agent alive :
  $D =$ Pick link in $D$
  $E = E - C + Similarity(Q, D) * G$
  Apply machine learning technique using $D$, $Q$ and $Similarity(Q, D)$ (optional)
  if $E > 1$ then
    Clone agent and share E between clones
  else if $E < 0$ then
    Death of agent
  end if
end loop
```

The idea we suggest is to dynamically measure relevance values of the crawled pages and try to establish a relation between relevance and energy to ensure a better stability of the population of agents.

We use a simple tf*idf similarity model in order to evaluate page relevance according to a topical request but indeed other and richer models can be used that are not in the scope of this paper. Independently from the used similarity model the endogenous fitness model requires relevance values to be converted into energy in order to control the activity of the crawler.

The original energy model [8] can be written as an equation:

$$\Delta e = f(r) = \rho \cdot r - C$$

$r$ is the relevance of the crawled page. $\Delta e$ is the energy update endorsed by the crawler.

$C$ is the cost in energy for the visit of a single page by an agent. It contributes to determine how selective the crawler is related to the success of its agents and how tolerant it is to the traversal of minimal relevance pages.

$\rho$ is the energy gain rate used to convert relevant information in energy. The range of $\rho$ depends on the order of the similarity metric used, and on the concrete relevance of collected pages.

In order to better regulate the population of agents we wish to control the reproduction rate and we choose to introduce a constant gain parameter that defines the energy gain of top relevant pages. Sigmoidal functions are often used in artificial life systems to model state transitions. An energy variation function inspired by sigmoid allows both cost and gain normalization for extreme values. Figure 1 shows how the energy varies according to our sigmoidal model.

$$\Delta e = f(r, R) = \left( (G + C) \cdot \frac{1}{1 + e^{-\lambda(r+R)}} \right) - C$$

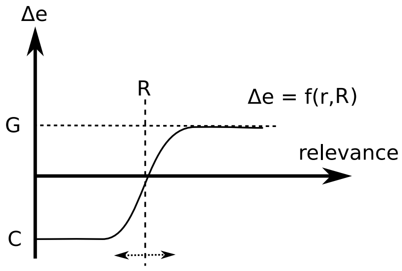

Fig. 1. Sigmoidal function between textual relevance of the documents and energy variations

Energy balance in the system when n pages have been collected is ensured if the following constraint is satisfied :

$$\sum_{i=0}^{n} f(r_i, R) = 0$$

where $r_i$ is the relevance of the $i^{th}$ visited page and $R$ is to be determined.

To calculate a satisfying $R$ we use an approximation of $f(r, R)$.

We define $g(r, R)$ as :

$$r \leq R \rightarrow g(r, R) = -C$$

$$r > R \rightarrow g(r, R) = G$$

For a high value of $\lambda$ the slope in the intermediate state of the sigmoid function is greater and over a large number of values $g(r, R)$ is a satisfying approximation of $f(r, R)$.

We can use a probabilistic interpretation to get the expected value of $\Delta e$.

$$\mathrm{E}(\Delta e) = \int_{r \leq R} f(r, R) \cdot p(r) \cdot dr + \int_{r > R} f(r, R) \cdot p(r) \cdot dr$$

Where $p(r)$ is the probability density that the collected page has a relevance of $r$.

When approximating $f$ by $g$ we get:

$$\mathrm{E}(\Delta e) \simeq \int_{r \leq R} -|C| \cdot p(r) \cdot dr + \int_{r > R} G \cdot p(r) \cdot dr$$

$$\mathrm{E}(\Delta e) \simeq -|C| \cdot Prob(r \leq R) + G \cdot Prob(r > R)$$

Where $Prob$ is the distribution associated to $p$. $Prob(r > R)$ is the probability that a page's relevance is higher than $R$.

We seek a null mathematical expectation for $\Delta e$:

$$\mathrm{E}(\Delta e) \simeq 0$$

Therefore:

$$\frac{Prob(r > R)}{Prob(r \leq R)} = \frac{G}{|C|}$$

To estimate a satisfying value for $R$ we sort the previously visited pages by their relevance values. Let $n$ be the total number of pages and $n^*$ be the rank of the page whose relevance is the closest to the ideal $R$.

We can use $n^*$ and $n$ to estimate $Prob(r \leq R)$ and $Prob(r > R)$:

$$Prob(r \leq R) \simeq \frac{n^*}{n}$$

$$Prob(r > R) \simeq \frac{n - n^*}{n}$$

$$\frac{Prob(r > R)}{Prob(r \leq R)} \simeq \frac{n^*}{n - n^*}$$

$$\frac{n^*}{n - n^*} \simeq \frac{G}{|C|}$$

$$n^* \simeq \frac{n}{C/G + 1}$$

In practice, we calculate $n^*$, on the basis of the history of the $n$ collected pages, then we set $R = r_{n^*}$

In order for the approximation of $f(r, R)$ by $g(r, R)$ to make sense we need to maintain a sufficiently large value for $\lambda$.

Through experience we defined $h(r_m, r_M, R)$, where $r_m$ is the smallest and $r_M$ the greatest relevance values respectively, such as :

$$\frac{8}{R - r_m} \leq r_M - R \rightarrow h(r_m, r_M, R) = \frac{8}{R - r_m}$$

$$\frac{8}{R - r_m} > r_M - R \rightarrow h(r_m, r_M, R) = \frac{8}{r_M - R}$$

We use $\lambda = h(r_m, r_M, R)$.

The user's parameters at this point are C and G. They tune the tolerance of the crawler to the traversal of low relevance zones, the reactivity of the agents to highly relevant pages and indirectly the selectivity of the crawler. In practice, a value of C close to 0 means that the crawler tolerates that its agents traverse many low relevance pages before dying. On the contrary, a value of C greater than 1 ensures that every agent who visited a low relevance page dies instantly. If G is set to 10 then an agent visiting a very relevant page creates about ten clones of himself. If C is set to 0.5 and G to 10 then about one page in twenty is considered relevant by the system and rewarded by an energy gain.

Through experimenting with the crawler, studying the graph of the Web, and more specifically topical sub-graphs of the Web, it is possible to define static values for these parameters that should offer satisfying results in all use cases. This would provide an auto-adaptive system and let the crawling process in an unsupervised way.

### *B. Energy conservation*

In order for the crawler to be highly selective and reactive to the discovery of relevant information it is necessary to use quite high values of the energy parameter G. Even if our global energy balance computation provides the main stability of the system, it doesn't prevent coarse variations of the population of agents. We want to dissociate in time energy discovery variations and population variations by implementing an energy buffer on the pages. If the number of agents is too high the energy of relevant pages is not used and is left behind for future use. Conversely when the number of agents is low, dying agents are "teleported" at no cost on pages with energy remnants. In order not to let behind for too long some pages with energy remnants the stochastic process is modified and includes random walk toward these pages. The modified model is described in algorithm 2.

**Algorithm 2** New model with energy conservation

```
The crawler:
Initialize system with parameters:
- maximal energy cost and gain : $C$ and $G$
- initial number of agents : $N$
- upper and lower boundaries of the number of agents : $UB$ and $LB$
- seeds pages : $S$
- query $Q$
Initialize documents base $DB$ (empty or from precedent execution)
Initialize $N$ agents with energy $E = 1$ and current document $D$ in $S$
loop
    For each agent alive :
    Pick link in $D$ or in $DB$ {only documents with energy remnants can be picked in $DB$}
    Fetch $D$ from the Web or from $DB$
    update_energy()
    reproduce()
end loop

Agent.update_energy():
$r \Leftarrow$ content_relevance($D$,$Q$)
if $D$ not in $DB$ then
    $DB$.set_document($D$,$r$)
    $R \Leftarrow DB$.get_R()
    $DB$.set_doc_energy($D$,energy_function($G$,$C$,$R$,$r$))
end if
if System.number_of_agents() $> UB$ then
    $E \Leftarrow E - \|C\|$
else
    $E \Leftarrow E + DB$.get_document_energy()
    $DB$.set_doc_energy($D$, 0)
end if

Agent.reproduce():
if $E > 1$ then
    Create $\lfloor E \rfloor$ clones with $E \div \lceil E \rceil$ energy
    $E \Leftarrow E \div \lceil E \rceil$
else if $E < 0$ then
    if System.number_of_agents() $< LB$ then
        Force to pick $D$ in $DB$ at next iteration
    else
        die()
    end if
end if
```

## IV. Experiments

### A. Experimental settings

In order to validate the regulating effect of the sigmoidal model and the energy buffer and to study the influence of the parameters, we executed the crawler many times, with varying models, topics and parameters values. The target topics are described as simple text queries and our prototype uses the tf*idf metric to evaluate the relevance of crawled pages and of links contexts. Links contexts are extracted according to the recommendation of Pant & al. [4]. The idf values are estimated before initialization by using the number of results returned by google queries on the keywords of the topic. All crawlers are initialized with two hundred agents and one, two or three manually selected highly relevant pages as seeds.

### B. Impact of the new energy function

Figure 2 shows the variations of population during time of three crawlers based on the original energy model. Those crawlers were executed five times each and only differ on the energy gain rate parameter. The curves clearly illustrate that depending on this parameter the population of agent can either die very quickly, increase brutally or remain usable.

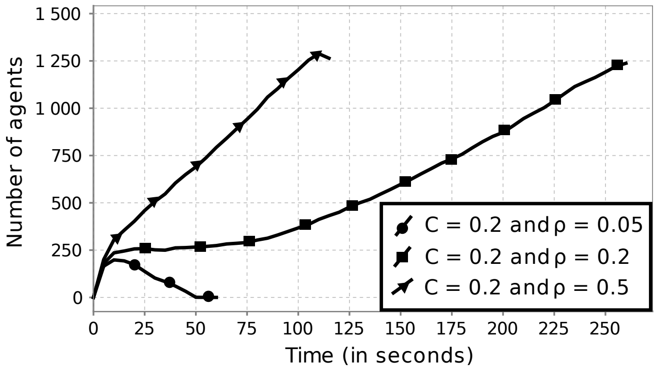

Fig. 2. Variations of population for various parameters of the original model

Figure 3 shows the variations of population during time of three other crawlers based on the new sigmoidal energy model. These crawlers were also executed five times with different values of $C$ and $G$. In the three cases the population of agents increase in time because of the corresponding increase of precision of the system, however, unlike in figure 2 the evolution of the population doesn't variate as drastically depending on user's parameters. Therefore population variations remain controllable and coherent with the activity of the crawler and the energy buffer will be able to erase local variations of the population.

Figure 4 shows the precision results of two crawlers already used in figures 2 and 3. The first one is based on the simple original energy model while the other one uses the new sigmoidal energy model. Values of $C$ are identical but the crawler based on the new model uses a value of $G$ set to 8. Figures 2 and 3 show that the two crawlers have a quite similar evolution of their population in time but the difference of precision seen in figure 4 clearly illustrates the positive impact of the introduction of the parameter $G$.

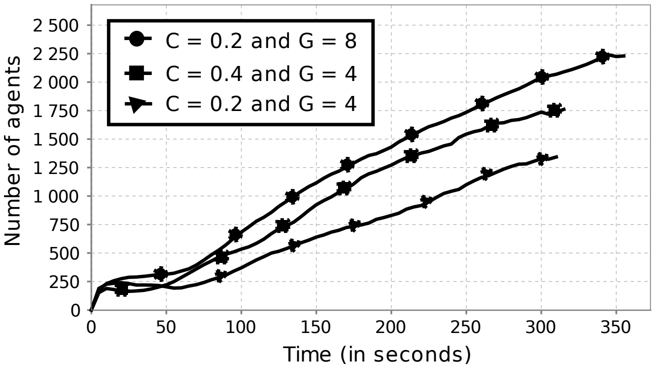

Fig. 3. Variations of population for various parameters of the sigmoidal model

### C. Impact of energy conservation

We have implemented and tested the energy buffer described in section III.C. The variations of the size of the populations of agents for two crawlers with or without the energy conservation buffer are shown in figure 5.

To illustrate the effectiveness of the energy buffer those crawlers were initialized with a high value for $G$ set to 30, implying brutal variations of the population over long crawls of 25000 pages. The energy conservation buffer is effective and the size of the population is clearly bounded by the values given by the user. We can see that the size of the population tends to be equal most of the time to the upper boundary (250 here), that is because the precision of the system increases in time. In the case of longer or more difficult crawlings where the whole accessible topical graph has been visited, the precision decreases and the size of the population tends to be equal to the lower boundary.

Figure 6 shows the precision for the same crawlers as those presented in figure 5. We can see that the precision of the crawler with energy conservation is a little lower than its counterpart. The use of the energy buffer introduces a delay in the scanning of relevant pages: the difference of precision is progressively reduced to zero.

## V. Conclusion

We have introduced some improvements of the endogenous fitness model for focused crawling to gain a better control of the size and activity of the agents-based crawler. Experimental results show that we indeed regulate the crawler and improve its efficiency thanks to the proposed energy management strategy. The previous results are encouraging as a preliminary step toward an efficient and practical crawler for Internet information retrieval.

## References

[1] F. Menczer and R. K. Belew, "Adaptive retrieval agents: Internalizing local context and scaling up to the web," *Machine Learning*, vol. 39, no. 2/3, pp. 203–242, 2000.

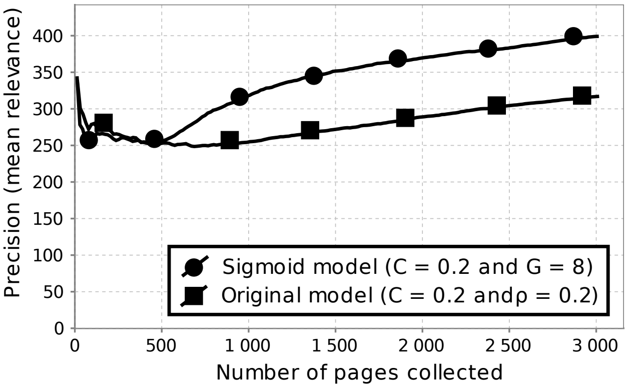

Fig. 4. Comparison of the precision of the two models

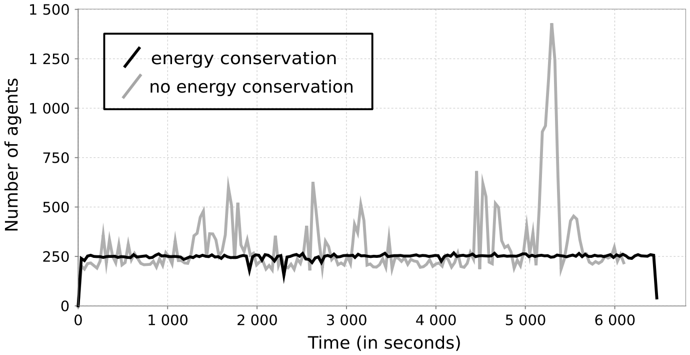

Fig. 5. Effect of energy conservation on the variations of population in the system

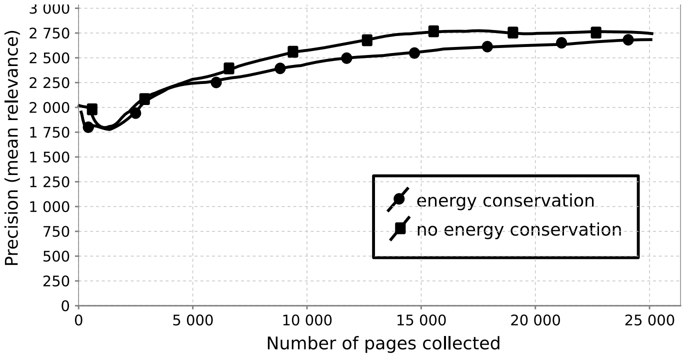

Fig. 6. Effect of energy conservation on the precision of the system

[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)*, vol. 31, no. 11–16, pp. 1623–1640, 1999.

[3] G. Salton, *Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer*. Addison–Wesley, 1989.

[4] G. Pant and P. Srinivasan, "Link contexts in classifier-guided topical crawlers," *IEEE Trans. on Knowl. and Data Eng.*, vol. 18, no. 1, pp. 107–122, 2006.

[5] G. Pant, P. Srinivasan, and F. Menczer, "Exploration versus exploitation in topic driven crawlers," 2002.

[6] ——, "Crawling the web," in *Web Dynamics*, 2004, pp. 153–178.

[7] P. Srinivasan, F. Menczer, and G. Pant, "A general evaluation framework for topical crawlers," *Inf. Retr.*, vol. 8, no. 3, pp. 417–447, 2005.

[8] F. Menczer, R. K. Belew, and W. Willuhn, "Artificial life applied to adaptive information agents," in *AAAI Spring Symposium on Information Gathering*, 1995.

[9] G. Pant and F. Menczer, "Myspiders: Evolve your own intelligent web crawlers," 2002.

[10] A. Revel, "Web-agents inspired by ethology: A population of "ant"-like agents to help finding user-oriented information," in *WI '03: Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence*. Washington, DC, USA: IEEE Computer Society, 2003, p. 482.

[11] F. Gasparetti and A. Micarelli, "Adaptive web search based on a colony of cooperative distributed agents," in *Cooperative Information Agents*, S. O. A. L. H. Klusch, M.; Ossowski, Ed., vol. 2782. Springer-Verlag, 2003, pp. 168–183.

[12] I. Kushchu, "Web-based evolutionary and adaptive information retrieval," *IEEE Trans. Evolutionary Computation*, vol. 9, no. 2, pp. 117–125, 2005.