4ZA-06

遺伝的アルゴリズムを用いたWWW空間の探索*

池田 隆文　　伊庭 斉志†

†東京大学工学部電子情報工学科

1 はじめに

近年、WWWの急速な拡がりは検索エンジンのデータベースやその出力結果の巨大化を招き、情報検索作業を再び困難なものにさせつつある。また一方で、IBMのClever Project[1]やgoogleに[2]代表されるような新世代型の検索エンジンが台頭してきた。これらの検索エンジンでは、収集したページ内に現れるキーワード数等のローカルな情報だけでなく、ページ間のリンク構造等のグローバルな情報を考慮することによって検索効率の向上を図っている。

本研究では知的エージェント集団の探索過程を追跡することでWWW空間のグローバルな情報を捉える手法を提案する。そしてこのようにして得られたグローバルな情報も利用してローカルな情報だけでは得られなかったような有用なページを収集することを目指す。

2 InfoSpiders

ユーザフィードバックと知的エージェント集団の進化・学習を通してユーザが興味ある情報を探す出すシステムとして、Filipo MenczerとRik Belewによって提案されたInfoSpiders[3, 4]がある。

InfoSpidersでは、エージェントは遺伝子に記述された探索戦略に則ってWWW空間を探索し、辿り着いた先のページをユーザフィードバックに基づいた評価関数によって評価する。そしてその評価値をエネルギーとして繁殖していく。より優秀な探索戦略を獲得したエージェントは生き残り、より有用なページには多くのエージェントが集中することになる。しかし、このシステムにはいくつかの問題点を指摘できる。

第一に、ページの評価はいくつかの評価関数によって為されるため、このシステムによって推薦されるページはその評価関数値の大きなページに過ぎない。そのためこのシステムの有効性は評価関数の妥当性に依存する。しかし、万能なページ評価関数は未だ提案されていない。そのためこのシステムが既存の検索エンジンより高いパフォーマンスを見せるかどうかは不明である。またそのためにユーザフィードバックを用いているが、エージェントが収集してくるページは膨大な量であるのに較べ、ユーザが評価できるページは限られてしまうので、その効果は小さなものにならざるを得ない。

第二に、このシステムのページ評価基準はページ内の情報のみに依っている。エージェント集団の探索の結果だけでなく、その探索過程へも注意を向ければ、WWW空間の構造を浮き彫りにすることができるかも知れない。そしてそのようなグローバルなWWW空間構造の情報を利用することでより効率の良い探索を行なえる可能性がある。

本研究ではInfoSpidersの以上のような欠点を主にエージェントの探索戦略やエネルギーの獲得方法を改良することで解決することを目指す。

次節では実現したシステムのアルゴリズムについて説明する。

3 アルゴリズム

3.1 探索開始

ユーザからクエリーが与えられると、まず既存の検索エンジンに問い合わせられる。そして検索結果として返されてくるページのリスト中のいくつかの

*WWW space search by means of GA

†Takafumi Ikeda, Hitoshi Iba

†University of Tokyo,7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

e-mail : ikeda@miv.t.u-tokyo.ac.jp

ページにエージェントはランダムに配置される。この時同時に各エージェントには一定のエネルギー初期値と、ページの嗜好性が与えられる。

3.2 探索戦略

本システムではエージェントは特定の探索戦略を持っていない。エージェントは現在のページからリンクされているページの中からランダムに一つのリンクを選び、その先のページに移動する。
一度訪れたページはキャッシュファイルに書き出される。キャッシュファイルにはページ内に現れるキーワード数と、エージェント集団がそのページを訪れた回数も記録される。エージェントの探索はランダムに行なわれるので、訪問回数の多いページはそれだけ他のページから参照されている度合が大きいと考えられる。キーワード出現数はローカルなページの評価値を表し、訪問回数はグローバルなページの評価値を表すと考えられる。

3.3 エネルギー獲得

各エージェントは訪れたページの評価を自分の嗜好性に基づいて決定する。エージェントのページ嗜好性は2つの変数α、βで表される。エージェントが訪れたページのローカルな評価値（本研究ではクエリー出現数）がαに近いほど、またグローバルな評価値（本研究では訪問回数）がβに近いほどより多くのエネルギーをエージェントは得ることができる。

3.4 繁殖、死滅

エージェントはある閾値以上のエネルギーを持っている場合子孫を作ることができる。この繁殖は無性生殖である。繁殖の際、親エージェントのエネルギーは半分になり、その分が子エージェントのエネルギーとなる。また、このときページの嗜好性は突然変異を起こす。子エージェントの嗜好性は親のものとは異なったものになる。

3.5 結果出力

エージェント集団を何世代かに渡って探索させた後、訪問回数の多い順にページを出力する。多様な嗜好性を持ったエージェント集団が十分な探索を行なった結果訪問回数の多いページというのはそれだけ有用なページであると考えられるからである。
出力結果例を以下に示す（図1）。


Search result:

	訪問回数	キーワード数
T. IKEDA Home Page http://localhost/~ikeda/	27	1
HOBBY http://localhost/~ikeda/hobby.html	20	1
T. IKEDA Home Page http://localhost/~ikeda/index.shtml	20	1
OTHERS http://localhost/~ikeda/others.html	18	1
(無題) http://localhost/~ikeda/minipaper1.ps	17	0
CYCLING UP TO SAKATA http://localhost/~ikeda/kisei-top.html	15	0
(無題) http://localhost/~ikeda/minipaper1.dvi	13	0
(無題) http://localhost/~ikeda/french.ps	10	0
STUDY http://localhost/~ikeda/study.html	10	1
CYCLING UP TO SAKATA http://localhost/~ikeda/kisei-data.html	9	0
http://www-cs-staff.stanford.edu/~uno/faq.html	9	0
CYCLING UP TO SAKATA http://localhost/~ikeda/kisei3.html	9	0
CYCLING UP TO SAKATA http://localhost/~ikeda/kisei1.html	7	0
http://www.stevenlevy.com/hackers.html	7	0

図 1: 出力結果例

4 おわりに

本研究では知的エージェント集団の進化を通してWWW空間を探索する手法を提案した。
今後は以下のような拡張を行なう予定である。本研究ではページのローカルな評価値はキーワードの出現数のみとしたが、単語の重み付け計算等も行ないページの意味にまで踏み込んだ評価値を用いるようにしたい。またキャッシュファイルを大規模化及びデータベース化することで、応答速度の向上を図る。これによってより実用に耐えるシステムの構築を目指すつもりである。

参考文献

[1] http://www.almaden.ibm.com/cs/k53/clever.html

[2] http://www.google.com/

[3] F. Menczer, R.K. Belew: Adaptive Information Agents in Distributed Textual Environments. Proc. 2nd Intl. Conf. on Autonomous Agents (Agents '98)

[4] F. Menczer, R.K. Belew: Adaptive Retrieval Agents: Internalizing Local Context and Scaling up to the Web. Technical Report CS98-579 University of California, San Diego, April 1998; shorter version to appear in Machine Learning Journal, 1999