# 蟻の行動パタンを用いたテキスト分類の試み*

4L−4

安井 照昌†
三菱電機株式会社 中央研究所‡

## 1 はじめに／テキスト情報ベースの構築

最近の情報の量的な氾濫はすさまじく、利用可能かも知れないすべての情報をユーザが直接見て選別・処理することは不可能になりつつある。ユーザの必要とする情報だけを迅速に得るには、カテゴリをあまり限定しない大量の文書を何らかの方法で自動分類し、ユーザに必要と思われる文書を集める機能が必要となる。こうした機能を持つシステムは本質的に大規模であり、かつ情報は分散しておかれていると考えられる。

我々は以前からニューラルネットワークを用いた文書分類システムの実現について研究してきた[2]が、専用のハードウェアを用いない場合、これらの方法が並列/分散処理にあまり適しているとはいいがたい。

更新の頻発する巨大な分散系では、一括してすべての情報を処理したり、整合性を保ったりすることは困難であり、各部分で処理が行なわれる結果として徐々に全体の整合性を高めるような方式が必要である。

Artificial lifeの研究分野では昆虫などの挙動を模倣し、集合的に分散問題解決を行なう研究がある[1]。このような手法は処理を行なうエージェント間に複雑な通信を発生せず、また、計算がローカルに進むという特徴を持つ。

分散環境において大量のテキストをこのような手法を用いて分類する方法について報告する。

## 2 単語の出現パタンによる分類法

領域を限定しない大量の文書を一様に扱う方法として、単語の共起関係を用いる方法がある。同じ単語を多く用いている文書間には内容的にも関係があるという仮定に基づき、単語の出現パタン（ある単語がどの文書に何個含まれているか）によって、文書を類似のもの同士のクラスタに分けるという方針をとる。ただし、どの文書にも一様にたくさん含まれる単語はあまり文書間の違いを表さない上に、分類において重要になる単語の影響をマスクするため、これの重みを下げる必要がある。実際にはこのthresholdをどのくらいに設定するかによって大分類をしたり細かい分類をしたりすることになる。

[2]などで用いた方法を扱いやすくするために単純化し、文書$a$,$b$間の一致度$F(a,b)$を以下のように定義する。

$n$個の文書が$m$個の単語を含む時、文書$i$に含まれる$j$番の単語の個数を$w(i,j)$とし、全体で$h$回以上現れる単語を除くとすると、

$$W(j)=\sum_{i=0}^{n} w(i,j)$$

$$c(i)=\begin{cases} 1, & W(i)<h \\ 0, & otherwise \end{cases}$$

$$f(a,b)=\sum_{i=0}^{m} w(a,i)\cdot w(b,i)\cdot c(i)^2$$

を用いて、

$$F(a,b)=\frac{f(a,b)}{\sqrt{f(a,a)\cdot f(b,b)}}$$

これによってサンプル文書($n=55, m=1952$)を分類した結果を図1に示す。数字は文書に付けられた番号で、同一の並びに属する文書は同一のクラスタに分類されたことを示す。

```
A: 0 2 4 6 7 8 10 11 12 13 14
B: 1
C: 3
D: 5
E: 9
F: 15 16 17 18 19 20 21 22 23 24
G: 25
H: 26 27 28 34
I: 29 30 32 33
J: 31
K: 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54
```

図1: サンプルの分類結果

## 3 蟻を用いた実現

[1]によると、蟻が巣の中で卵や繭・餌などのコロニーをつくる挙動を非常に単純な方法で模倣することができる。

蟻の挙動は以下のように記述される。

- 蟻はランダムに動く。
- てぶらの蟻がものを見つけた場合、そのもののにおいと、その場所に漂うにおいが似ていなければ、高い確率でそれを持ち去る。
- ものを持ったありはそのにおいと、その場所に漂うにおいが似ていれば、高い確率でそれを置いていく。

*A document clustering method with ant-like agents.

†Terumasa YASUI, e-mail: yasui@sys.crl.melco.co.jp
‡Mitsubishi Electric Corporation Central Research Laboratory

これを長時間続けると、同じあるいは似たもののコロニーが形成されていく。

この方法を拡張して、それぞれの文書中に含まれる単語の数をその文書が発するその単語のにおいの強さとすることにより、文書を類似文書のクラスタに分けると言う問題を集合的・分散的に解決することができる。

ただようにおいの強さは、物体から発生したものの拡散をシミュレートする方法もあるが、[1] にあるように、蟻が出会った物体のにおいを覚えておくことでも実現できる。広さ $A$ の field に $n$ 個の物体があり、$i$ 番の物体のにおいの強さが $o(i)$ であるとき、$A$ が十分狭く、忘却率 $p << 1$ であれば、$j$ ステップ前に出会った物体のにおいを $o'(j)$ とするとき、

$$\sum_{j=0}^{\infty} o'(j)\cdot(1-p)^j \longrightarrow \frac{1}{p\cdot A}\cdot\sum_{i=0}^{n} o(i)$$

となる。

広さ $A$ の field に $n$ 個の物体があり、$m$ 種類のにおいを出しているとする。$i$ 番の物体が発する $j$ 番のにおいの強さを $o(i,j)$ とすると、その場のにおい $j$ の強さは、

$$O(j) = p\cdot A\cdot\sum_{j=0}^{\infty} o'(j)\cdot(1-p)^j$$

で計算できる。$A$ がある程度広く、$p$ がある程度大きい時、$O(j)$ はその領域におけるローカルなにおい $j$ の強さを表し、逆に、$A$ がある程度狭く、$p$ がある程度小さい時、$O(j)$ は field 全体におけるにおい $j$ の強さを表す。

領域中に物体を一つだけ置いて、領域の広さと忘却率を変化させた時の、物体からの距離と計測されたにおいの強さの関係を図 2に示す。

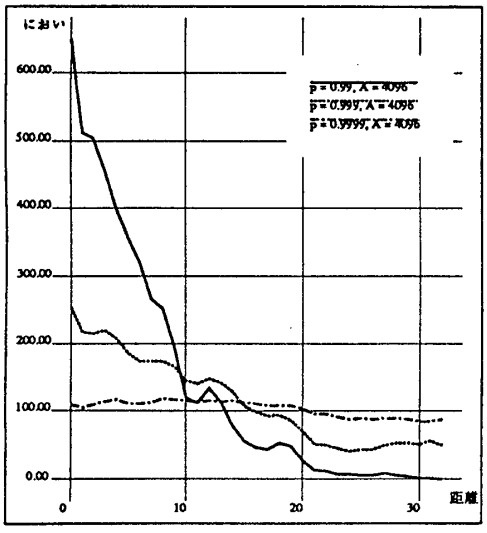


図 2: 距離とにおいの強さ

そこで、2 種類の忘却率 $pl$ と $ps(pl > ps)$ を用意して、$pl$ を用いて field 全体のにおいの強さ $ol(j)$、$ps$ を用いてその領域のにおいの強さ $os(j)$ を測る。前者は全文書中でのその単語の使用頻度を表し、後者はその時点での蟻の近辺に散在する文書におけるその単語の使用頻度を表す。

$$c(j) = \begin{cases} 1, & ol(j) < h \\ 0, & otherwise \end{cases}$$

としたとき、物体 $i$ のその地点のにおいとの一致度を

$$u(i) = \frac{\sum_{j=0}^{m} o(i,j)\cdot c(j)\cdot os(j)}{\sqrt{\sum_{j=0}^{m}(c(j)^2\cdot o(i,j)^2)\cdot\sum_{j=0}^{m}(c(j)^2\cdot os(j)^2)}}$$

として計算する。$u(i)$ が大きいほど高い確率で物体を置き、$u(i)$ が小さいほど高い確率で物体を持ち去る。これを続けることにより、類似の文書どうしは互いに近いところに集まる。

結果を図 3に示す。

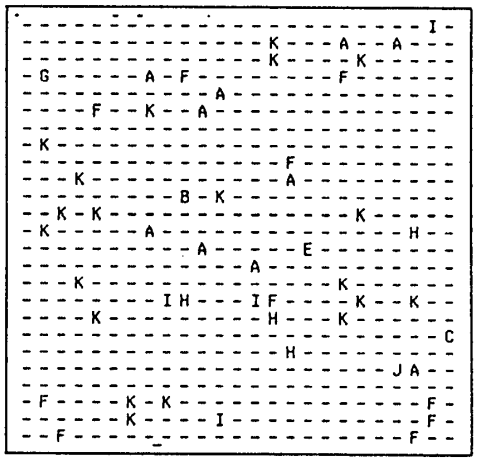

図 3: 文書のコロニー

## 4 おわりに

大規模な情報ベースは、本質的に並列 / 分散システムとなり、これらの上で行なう処理も分散処理となるため、そのための一方式として、集合的な問題解決の方式を試みた。

ここで用いた単純なテキストのパタン化の方式自体にも改良が必要であることはもちろんだが、そうした場合、単純に分類するのではなく、もっと複雑な構造を作ったり、もっと複雑なエージェントを用いる必要が生じると思われる。今後の研究課題である。

### 参考文献

[1] Denenbourg, J. L., et al. The dynamics of collective sorting robot-like ants and ant-like robots, From animals to animats (eds.Meyer, J.-A. and Wilson, S. W.), The MIT Press (1991).

[2] 豊浦潤, 有田英一 テキストの内容を表すワードマップ作成の試み, FI 28-3, 情報処理学会研究報告 (Nov. 1992).