5G−1 自己組織化マップを用いたテキスト自動分類の試み *

津高 新一郎 †
三菱電機株式会社 中央研究所 ‡

1 はじめに

近年社会の情報化に伴い、情報ベースシステムが巨大化の一途を辿っており、それとともにさまざまな問題が顕在化してきた。とりわけ情報の受け取り手側に生じる知識獲得のボトルネックは大きな問題となっており、これを解決するための優れた検索方式の確立が重要な課題となっている。

現在、テキスト検索方式として、カテゴリやキーワードなど予めテキストに付加しておいた情報を利用しつつ、文字列の一致に基づくAND/OR検索を行なうといったことがしばしば行なわれるが、これらカテゴリ分類やキーワード付けは人手による作業に頼っているのが現状である。情報ベースシステムの大規模化に伴い、カテゴリの細分化やキーワードの多様化が進み、今後このような作業の負荷はますます大きくなるものと予想される。

本稿ではニューラルネットワークの一種である自己組織化マップを用いてカテゴリ分類、キーワード付けを自動的に行なう手法について述べる。我々は従来のカテゴリ分類を一般化した手法としてテキストを2次元平面上にマッピングすることを提案し、またテキストを象徴するキーワードを抽出してキーワードマップを作成することを試みる。このような2次元情報をユーザーに提示することにより、よりユーザーインターフェースの優れたテキスト検索システムの構築が可能となることが期待される。

2 自己組織化マップ

自己組織化マップ(Self-Organizing (Feature) Map)はT.Kohonenによって提案された中間層のない2層型の教師なし競合学習モデルである[1]。出力層の各ユニットが層の中で位置を持つ点が他の学習モデルと異なる。自己組織化マップの特徴の一つにトポロジカルマッピングがある。すなわち、距離が接近した2つの入力パターンに対しては、出力層上で近い位置にあるユニットがそれぞれ反応する。

出力ユニットは通常2次元平面（出力平面という）の上に並べられている。出力層の各ユニットは入力パターンと等しい次元のパターンを持っており、学習はこのパターンを入力パターンに選択的に近付けることによって進行する。入力パターンを$x(t)$、ユニットiの保持するパターンを$m_i(t)$、出力平面上でのユニットiの位置をr_iとすると、学習手順は以下のように表現される。

1. 入力パターンに一番近いパターンを持つ出力ユニットcを探す。
2. 出力平面上でcの近傍のユニットの集合$N_c(t)$を求める。
3. ユニット$N_c(t)$の持つパターンを入力に近付ける。

$$m_i(t+1) = \begin{cases} m_i(t) + \alpha(t)[x(t) - m_i(t)] & (i \in N_c(t)) \\ m_i(t) & (i \notin N_c(t)) \end{cases}$$

4. $N_c(t)$とαを次第に小さくしながら1〜3を繰り返す。

αは学習定数を表し、以下のようなガウス関数がよく用いられる。

$$\alpha(t) = \alpha_0(t)\exp(-\|r_i - r_c\|^2/\sigma(t)^2) \quad (i \in N_c(t))$$

$\alpha_0(t)$、$\sigma(t)$としては単調減少の一次関数や指数関数がよく用いられる。

3 テキスト分類手法

前述の自己組織化マップをテキストの自動分類に応用することを試みる。自己組織化マップの持つトポロジカルマッピングの性質により、同じカテゴリに属するテキストのパターンがお互いに近ければ、出力平面上の特定の範囲にカテゴリが形成されることが予想される。またこのとき出力ユニットの持つパターンからキーワードの抽出が可能になる。具体的には以下のような手順を用いた。

1. テキストのパターン化
   分類の対象となるテキストから単語を切り出し、単語の種類を次元とし各要素は単語の出現頻度に比例するようなベクトル表現を用いることによってテキストをパターン化する。
2. 自己組織化マップによるパターン学習
   テキストをパターン化したものを自己組織化マップの入力とし学習させる。
3. テキストのマッピング
   学習後、分類の対象となるテキストのパターン各々に対し最も近いパターンを持つ出力ユニットを探すことにより、出力平面上でのテキストのマップを得る。
4. キーワード抽出
   各出力ユニットの持つパターンについて最大の要素を抽出し、これらを単語に変換することにより、その出力ユニットに分類されたテキストを代表するキーワードを得る。

---

*Clustering of Texts Using Self-Organizing Feature Map
†Shin-ichiro Tsudaka
‡Central Research Laboratory, Mitsubishi Electric Corporation

4 テキスト分類結果

実際に 20 × 20 のマップを用いて計算を行なった。ここで分類の対象のテキストとしては、鉱工業関係のテキストのうち、鉄鋼、電子工業、自動車工業など10カテゴリからランダムに500テキストを抜き出したものを用いた。これらのテキストに計4回以上登場する単語（約3300種）のみを考慮してテキストのパターン化を行なったのち、さらに正規化を行なって自己組織化マップの入力とした。学習回数は2000回、学習式としては前述のものをそのまま用いる。ただし $h(t)$ としては初期値1、終値0の一次関数を、$\sigma(t)$ としては初期値がマップの一辺の長さ／2、終値が0の一次関数を用いる。またユニット c の近傍ユニット N_c としては c を中心とする一辺 4σ の正方形の領域に含まれるユニットと定義する。

4.1 テキストの分布

分類に用いた全テキストの分布を図1、その中で電子工業関係のテキストの分布を図2に示す。各図で正方形の一辺の大きさはそのユニットに分類された記事の数を表している。図1ではかなり平面全体に文書が分布しているが、図2では分布に偏りが見られ、自動分類の試みが成功していることが分かる。

4.2 キーワードマップ

前述の手順に従い各出力ユニットからキーワードを1つ抽出し、出力平面上のキーワードマップを作成した。結果を図3に示す（領域が1×1以下のキーワードは省略）。あきらかに関連が深いと思われるコンピュータ・パソコン・ワープロ・ソフトなどといった語群、あるいはメーカー・米国・半導体・日本といった語群が集中しており、キーワードの抽出に成功しているといえる。

5 おわりに

本稿では自己組織化マップを用いてテキストを自動分類することを試み、2次元平面上でのテキストのマッピングやキーワード抽出の手法について述べた。しかしながら本稿で取り上げたキーワードの抽出方法には、「日本」や「人」など普遍的に存在する（それ故検索を助ける情報としては利用しにくい）単語がキーワードになりやすいという難点がある。これを避けるため、テキスト中での単語の分布状況から算出される単語の「専門度」[2]を考慮した学習モデルを現在考えている。

参考文献

[1] Teuvo Kohonen :
"The Self-Organizing Map," Proceedings of the IEEE, Vol.78, No.9, pp.1464-1480 (1990).

[2] 豊浦潤、有田英一：
「単語の連想関係に基づく意味マップによるテキスト表現の試み」、情報処理学会第45回全国大会、分冊3、pp.247-248 (1991)。

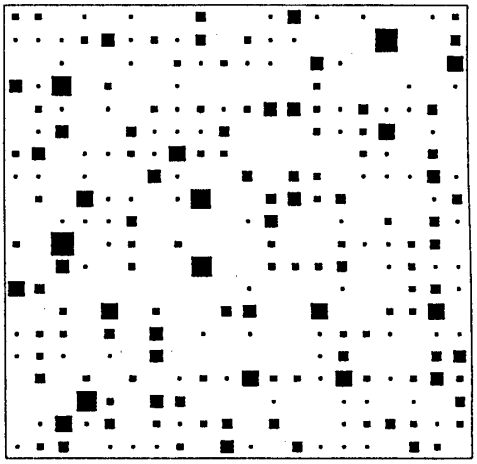

図 1: 全テキストの分布

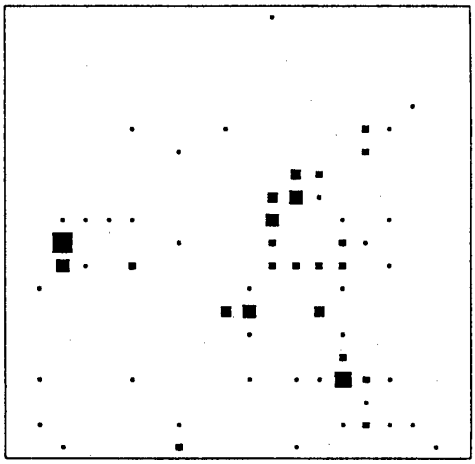

図 2: 電子工業関係のテキストの分布

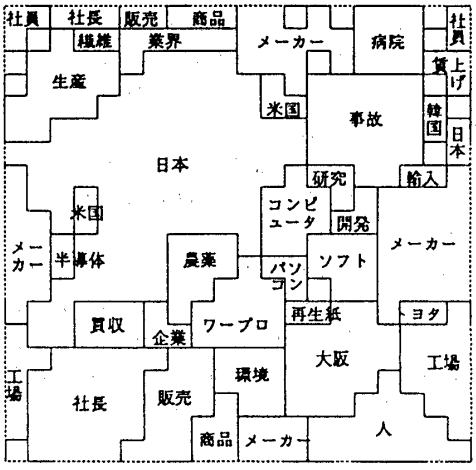

図 3: キーワードマップ