自己組織型神経回路網を用いた可視化情報検索

榎本康佑[†]　村田真樹[‡]　馬　青[†‡]
[†]龍谷大学
[‡]情報通信研究機構
qma@math.ryukoku.ac.jp

1 はじめに

情報検索は、それ自身に対する需要もさることながら、情報抽出や質問応答などおおよそすべての情報アクセスの必要不可欠な基礎技術であり、その研究開発は他の情報アクセス技術よりも早い時期から、そしてもっとも多方面からなされてきた。しかし、いまだに情報検索の主要手法とされている TFIDF 重み付けに基づくベクトル空間法（以降略して TFIDF 法と呼ぶ）を超える画期的なものが開発されておらず満足のできる検索精度が得られていない。また、大量の検索結果をいかに分かりやすくユーザ側に提示するかという可視化の視点からの研究もあまり見られていない。

われわれは情報検索の精度向上を図りながら可視化も実現させるために、自己組織型神経回路網に基づく可視化情報検索手法を提案した[1]。提案手法により、従来の順位付き検索結果が得られるだけでなく２次元上の可視的かつ連続的な情報検索結果も同時に得られた。しかし、検索精度はわれわれの意図に反し従来の TFIDF 法に比べ著しく低下した。そこで、精度向上を図るため、個々の記事に対応する神経回路網への入力ベクトルの構成に単語の頻度情報の代わりに単語の tfidf 値を用いることにした。さらに、分類語彙表[2]より得られた上位概念の情報の利用も試みた。計算機実験の結果、tfidf 値を利用することにより、前回の結果に比べ提案手法の精度が大幅に向上し、TFIDF 法よりも高い精度が得られた。さらに、上位件数を絞って検索する場合、提案手法の方が TFIDF 法より精度がはるかに高かった。したがって、提案手法が検索結果を可視化できるだけでなく（特に再現率よりも検索の精度要求が重視された場合の）検索精度の向上にも大きく寄与できると考える。

2 提案手法

2.1　基本的な考え方

われわれが目指す可視化情報検索は二段階の処理から構成される。第一段階はいわゆる従来通りの情報検索である。すなわち、1 つの検索要求に対し、TFIDF 法などを用い新聞やウェブサイトなどからその関連記事群を収集してくる。第二段階においては検索結果の精密化と可視化を行う。ここでいう可視化とは、第一段階の処理で得られた記事群を分類しその結果を 2 次元マップ上に表示することであり、検索結果の精密化とは可視化によって得られた分類結果を利用し検索結果を少数のよい結果に絞ることによって精度向上を図るものである。同時に、検索要求文を構成する重要語もマップ上に表示する。以降、このようなマップ

を記事マップと呼ぶ。したがって、このようなマップが構築されれば記事と記事の関係だけでなく記事と重要語の関係も一目瞭然になる。

本稿の提案手法は上記第二段階の処理に当たるものである。提案手法の性能を評価しやすくするために、１件の検索要求に関連する記事群とその検索要求の重要語に関する記事マップを構築する代わりに、情報検索コンテスト IREX で用いられた６件の検索要求とそれらの適合記事（正解記事）のマップを構築することにした。記事マップの自動構築に自己組織型神経回路網モデル（Self-Organizing Map, 略して SOM[3]）という多次元データを２次元に圧縮しながらクラスタリングできる教師なし機械学習手法を用いた。

2.2 データコーディング

ＳＯＭは実数ベクトルしか扱えないため、個々の記事または検索要求を多次元ベクトルに変換（データコーディング）する必要がある。よい記事マップが形成できるか否かはデータコーディングに大きく依存するので、最適なデータコーディング法を用いることがもっとも重要となる。

ここで、検索要求を記号 *Q_1, Q_2, Q_3, Q_4, Q_5, Q_6* で表し、それぞれの適合記事を記号 (*A1_1,...,A1_a1*), (*A2_1,..., A2_a2*), ..., (*A6_1, ..., A6_a6*)で表す。但し、*a1, ..., a6* はそれぞれの適合記事の総数である。以降、適合記事であるか否かが分からない任意の記事をテスト記事と呼ぶ。また、区別する必要がない場合、検索要求もテスト記事と合わせて記事と呼ぶ。このようにあわせて呼んでいる記事を *D_i(i=1,...,d)*で表す。但し、*d* は記事の総数である。ここで、個々の記事を以下のように名詞の集合で定義する。

$$D\_i = \left\{ noun_1^{(i)}, w_1^{(i)}, \cdots noun_{n_{ii}}^{(i)}, w_{n_i}^{(i)} \right\} \quad (1)$$

但し、$noun_k^{(i)}(k=1,\cdots n_i)$ は記事の中に存在する名詞の異なりであり $w_k^{(i)}$ は $noun_k^{(i)}(k=1,\cdots,n_i)$ の重要さを表す重みである。その重みはそれぞれ、出現頻度 tf または tfidf 値で求める。出現頻度で求める場合、それらを足し合わせて１であるように正規化している。また、分類語彙表を利用した場合、検索要求のみに対し、それに含まれる名詞と同じ上位概念を持っているすべての名詞を検索要求の式（１）に追加した。追加された名詞の重みには検索要求にあった名詞と同一のものを与えることにした。

仮に記事 *D_i* と *D_j* 間の相関あるいは類似度距離 d_{ij} を要素とする相関行列を求めることができれば、その行列の各行を用いることによって各記事を符号化することができる。すなわち、

$$V(D\_i) = [d_{i1}, d_{i2}, \cdots, d_{in}]^T \quad (2)$$

で表すことができる。このベクトルは SOM への入力となる。従って、記事の符号化において、記事間の類似性距離を求めることが鍵となる。

記事間の距離を求める際、検索要求どうしの類似性距離を最大にするという条件を考慮に入れた。それは、与えられた検索要求の関連記事の取り出しは記事マップからテスト記事とその検索要求との距離を測ることによって行われるため、検索要求どうしがマップ上に遠く離れる位置に配置され

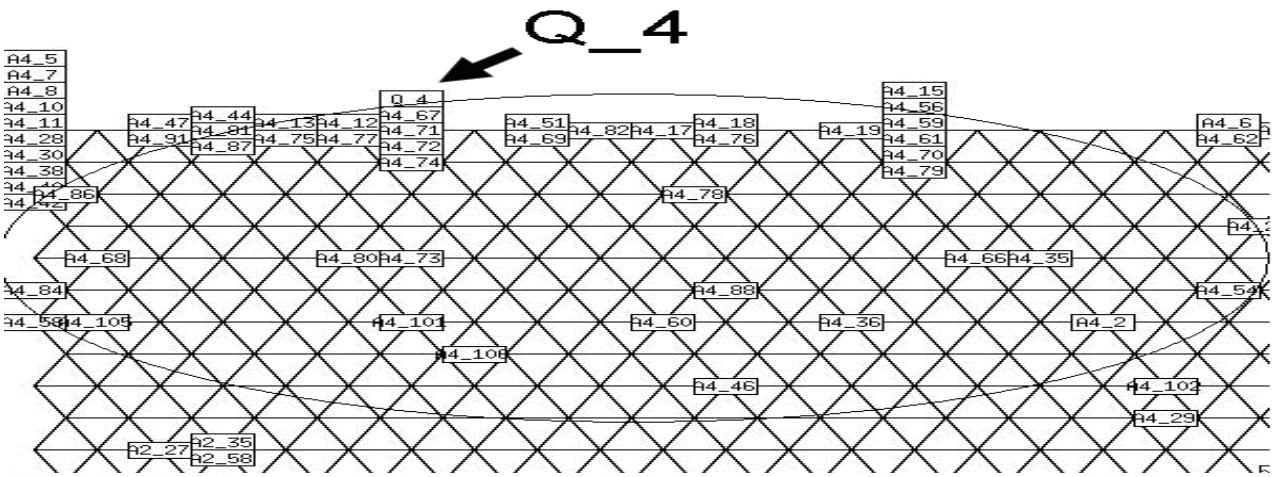

図１　実験で得られた記事マップの一部

なければならないからである。本稿では記事間の類似度距離の計算にこの条件を満足させるように以下の計算式を用いた。

$$d_{ij}=\begin{cases}1 & D\_i, D\_j \text{が検索要求の場合}\\ \phi(x) & \text{その他の場合かつ} \quad i\neq j\\ 0 & i=j\end{cases} \tag{3}$$

ただし、*Φ(x)*は以下のように定義される。

$$\phi(x)=1-C_{ij}/(D_i+D_j-C_{ij}) \tag{4}$$

ここでの D_i と D_j はそれぞれの記事の名詞の重みを足し合わせた値である。C_{ij} は記事の類似度を反映するものである。これを求めるためにまず、式(1)を以下のように書き換えた。

$$D\_i=\{(C_1^{(i)}, w_{c1}^{(i)}, \cdots C_{c_l}^{(i)}, w_{c_l}^{(i)}), (n_1^{(i)}, w_1^{(i)}, \cdots n_{m_i}^{(i)}, w_{m_i}^{(i)})\} \tag{5}$$

$$D\_j=\{(C_1^{(j)}, w_{c1}^{(j)}, \cdots C_{c_l}^{(j)}, w_{c_l}^{(j)}), (n_1^{(j)}, w_1^{(j)}, \cdots n_{m_j}^{(j)}, w_{m_j}^{(j)})\} \tag{6}$$

ここで、$C_k(k=1,\cdots,l)$ は Di と Dj の共通した名詞であり、$n_k^{(i)}(k=1,\cdots,m_i)$ と $n_k^{(j)}(k=1,\cdots,m_j)$ は Di と Dj の異なる名詞である。そして、C_{ij} を先行研究[1]でもっともよいとされた以下の方法で求める。

$$C_{ij}=\begin{cases}\sum_{k=1}^{l}\max\left(w_{ck}^{(i)}, w_{ck}^{(j)}\right) & \text{検索要求と記事の場合}\\ \sum_{k=1}^{l}\min\left(w_{ck}^{(i)}, w_{ck}^{(j)}\right) & \text{記事と記事の場合}\end{cases} \tag{7}$$

3　実験結果

3.1　データ

実験では６件の検索要求とそれらの 433 件の適合記事（正解記事）を使用した。適合記事の分布は表１に示す。

表1　実験に用いた適合記事の分布

A1	A2	A3	A4	A5	A6
80	89	42	108	49	65

3.2　SOM

SOM は 40×40 の 2 次元配列のノードで構成し、近傍の形状は六角形にした。整列のフェーズにおいては、学習回数 T を 10000 に、学習率の初期値α(0)を 0.1 に、そして近傍の初期半径σ(0)を 30 に設定した。整列のフェーズにおいて学習回数 T を 15000 に学習率の初期値α(0)を 0.01 に、そして近傍の初期半径σ(0)を 5 に設定した。

3.3　結果

図１は tfidf 値を重みとし、分類語彙表を利用しない場合に得られた記事マップの一部を示す。この図より、検

索要求 Q_4 とその適合記事 A4_*が互いに近い位置に配置され、適合記事どうしも比較的近い位置に配置されていることが分かる。

表2は、提案手法の四つの場合で得られた記事マップから個々の検索要求と記事との距離を測り、もっとも近い、適合記事数分の記事を取り出した時の精度（すなわち、正解した記事の数／適合記事数）と TFIDF 法の精度を示す。ただし、分類語彙表を BGH と略して記している。

表2　各手法の精度

SOM(w=tf)	0.45
SOM（w=tfidf）	**0.78**
TFIDF 法	0.67
SOM(w=tfidf, BGH)	0.73
TFIDF 法（BGH）	0.75

表3　上位 N 個の記事を取り出した場合の各手法の精度

手法 / N	TFIDF	TFIDF (BGH)	SOM (w=tf)	**SOM (w=tfidf)**	SOM (w=tfidf, BGH)
10	0.83	0.88	0.75	**1.0**	0.97
20	0.79	0.86	0.68	**0.98**	0.97
30	0.73	0.84	0.62	**0.97**	0.91
40	0.71	0.82	0.58	**0.97**	0.87

また、SOM(w=tf)は前回の研究に用いた手法と同一のものである。この表より、tfidf 値を重みとし、分類語彙表を利用しない場合の提案手法の精度が TFIDF 手法を含めたすべての場合においてもっとも高かったことが分かる。一方、分類語彙表の利用は TFIDF 法においては精度向上が見られたが、提案手法においては逆に精度を低下させてしまった。これは SOM の入力ベクトルを求める過程にある式（4）の分母を求める際、分類語彙表から得られた名詞の重みの計算がまだ最適になっていないことに起因するものではないかと考えられる。

表3は上位 N 個（N=10, 20, 320, 40）の記事を取り出した場合の各手法の精度を示す。この表からも今回の提案手法を用いることにより前回より精度が大幅に向上していることがわかる。そして、上位件数を絞って検索する場合、TFIDF 手法よりも提案手法の精度の方がはるかに高いことが分かる。これは提案手法が情報検索の精密化に大きく寄与できる可能性を示唆していると考える。

4　おわりに

本稿では自己組織型神経回路網に基づく可視化情報検索手法は検索結果の可視化を可能にしたことに加え、TFIDF 法以上の精度を有することと特に検索件数を絞った場合に TFIDF 法よりはるかに高い精度を有すること、すなわち、検索結果の精密化に大きく寄与できる可能性を示した。以上のことは 2.1 節に述べた基本的な考え方の裏づけにもなったと考える。

今後は提案手法への分類語彙表の導入効果の再確認を行うとともに、提案手法のさらなる精度向上を図っていきたい。そして、大規模新聞データを対象に、従来法による情報検索と提案手法による検索結果の精密化・可視化という二段階の検索手順を用い、高性能で実用的な可視化情報検索システムの開発を目指す。

参考文献

[1] 榎本康佑、村田真樹、馬青： 記事マップの自己組織化と情報検索、言語処理学会第１０回年次大会、2004 年 3 月

[2] 国立国語研究所: 分類語彙表、**大日本図書**、1964 年 3 月

[3] Kohonen, T.: Self-organizing maps, Springer, 2nd Edition, 1997