記事マップの自己組織化と情報検索

榎本　康佑*　村田　真樹**　馬　青*
*龍谷大学理工学部
**通信総合研究所けいはんな情報通信融合研究センター
qma@math.ryukoku.ac.jp

1　はじめに

情報検索の研究は 1960 年代に遡ることができる[1]．これまで，ブーリアンモデル，ネットワークモデル，ベクトル空間モデル，そして，確率モデルなど多くの検索モデルが提案されてきた[2]．ネットワークの急速な普及によって情報の流通が進み，ある面で情報過多の状況が生じている今，多くの情報の中から自分に必要な情報を的確に取り出す技術が一層求められるようになった．そのため，従来よりも知的な情報検索を行う必要性が生じ，自然言語処理と結びついた情報検索アプローチが模索されている．

本稿は，Kohonen が提案した自己組織化神経網モデル(Self-Organizing Map, 略して SOM[3])を用いる新しい情報検索モデルを提案する．提案モデルにおいては，与えられた検索要求（クエリ）のセットに対し，まず，記事マップが自己組織化によって自動構築される．記事マップ上には，すべてのクエリと記事が内容的な類似性に従って配置されるため，このマップ自身を可視的かつ連続的な検索結果としてとらえることができる．このようなマップから，与えられたクエリに合致する関連記事を容易に取り出すことができるだけでなく，同一のクエリに合致する記事間の関係や異なるクエリに合致する記事間の関係を見ることができる．そして，マップ上に配置されたクエリと記事の距離を測ることによって，従来の情報検索結果，すなわちランク付けされた関連記事を簡単に得ることができる．

2　提案手法

2．1　データ

記事マップを自己組織化するためには，記事間の類似関係を反映できるような学習データが必要となる．そのために，記事間の類似関係を記事に含まれる名詞を比較することによって結びつけることにした．それは内容的に近い記事どうしは多くの共通する名詞を持ち，内容的に遠い記事どうしはあまり共通する名詞を持たないと考えることができるからである．したがって，記事の中にある名詞を学習データとし，残りを不要語として使用しない．

学習データは日本で開催された情報検索コンテスト **IREX** で用いられた六つの検索要求文及びそれらの適合記事（正解データ）433 個を使用した．クエリは TOPIC-ID で囲まれた数字は設問番号を意味しており，次に DESCRIPTION が検索課題を端的に示すフレーズであり，NARRATIVE が検索要求を厳密に規定する説明文になっている．例を以下に示す．

<TOPIC>
<TOPIC-ID>1001</TOPIC-ID>
<DESCRIPTION>企業合併</DESCRIPTION>
<NARRATIVE>記事には企業合併成立の発表が述べられており，その合併に参加する企業の名前が認定できる事．また，合併企業の分野，目的など具体的内容のいずれかが認定できる事．企業合併は企業併合，企業統合，企業買収も含む．</NARRATIVE>
</TOPIC>

記事については日付，題名，本文で構成されており，日付は<DATE>，題名は<TITLE>，そして全体を<TEXT>で囲まれている．例を以下に示す．

<TEXT><DATE>１９９４年５月１６日</DATE><TITLE>自動車　Ｆ１のベネトン･チーム，リジェを買収</TITLE>　自動車Ｆ１のベネトン・チームは十四日，モンテカルロでマネジング・ディレクターのフラビオ・ブリアトーレ氏

がリジェ・チームの買収を発表．リジェは事実上，ベネトンのジュニアチームとなった．（時事）</TEXT>

クエリと記事から名詞のみを抜き出すために，奈良先端科学技術大学院大学が開発した形態素解析システム『茶筌』[4]を用いた．

２．２　データコーディング

ここで，クエリを記号 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_a_1), (A2_1, ..., A2_a_2), ..., (A6_1, ..., A6_a_6)で表す．但し，a_1, ..., a_6 はそれぞれの適合記事の総数である．適合記事はそれぞれの正解記事のことであり，すべての適合記事をこれからテスト記事と呼ぶ．また区別の必要のない場合，クエリもテスト記事と合わせて記事と呼ぶ．

記事の中の名詞のデータは次のように編集した．

$$D\_i = \{noun_1^{(i)}, f_1^{(i)}, \cdots, noun_{n_i}^{(i)}, f_{n_i}^{(i)}\} \quad (1)$$

ここで，$noun_k^{(i)}(k=1,\cdots,n_i)$ は記事の中に存在している異なった名詞を表しており，$f_k^{(i)}$ は $noun_k^{(i)}(k=1,\cdots,n_i)$ の相対頻度を表す．すなわち，

$$f_1^{(i)} + \cdots + f_{n_i}^{(i)} = 1 \quad (2)$$

仮に記事 D_i と D_j 間の相関あるいは類似性距離 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，

$$V(D\_i) = [d_{i1}, d_{i2}, \cdots, d_{in}]^T \quad (3)$$

これは 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キーポイントとなる．記事間の距離を求める際，クエリ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，与えられたクエリ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クエリとの距離を測ることによって行われるため，クエリどうしがマップ上に遠く離れる位置に配置されなければならないからである．本稿では記事間の類似性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．

$$d_{ij} = \begin{cases} 1 & D\_i, D\_j \text{がクエリの場合} \\ 1-\varphi(C_{ij}) & \text{その他の場合かつ } i \neq j \\ 0 & i = j \end{cases} \quad (4)$$

但し，φ(x)は以下のように定義される．

$$\varphi(x) = \begin{cases} 1 & if \quad x \geq 1 \\ x & if \quad x < 1 \end{cases} \quad (5)$$

明らかに，C_{ij} は記事の類似度を反映するべきものであり，本稿では以下の四つの方法で求めることとした．そのために，まず，式(1)を以下のように書き換えた．

$$D\_i = \{(c_1, f_{c1}^{(i)}, \cdots, c_l, f_{cl}^{(i)}), (n_1^{(i)}, f_1^{(i)}, \cdots, n_{m_i}^{(i)}, f_{m_i}^{(i)})\} \quad (6)$$

$$D\_j = \{(c_1, f_{c1}^{(j)}, \cdots, c_l, f_{cl}^{(j)}), (n_1^{(j)}, f_1^{(j)}, \cdots, n_{m_j}^{(j)}, f_{m_j}^{(j)})\} \quad (7)$$

ここで，c_k $(k=1,\cdots,l)$ は D_i と D_j の共通した名詞であり，$n_k^{(i)}(k=1,...,m_i)$ と $n_k^{(j)}(k=1,...,m_j)$ は D_i と D_j の異なる名詞である．そして，C_{ij} を求める４つの方法は以下の４つの式で表すことができる．

方法１

$$C_{ij} = \sum_{k=1}^{l} \max\left(f_{ck}^{(i)}, f_{ck}^{(j)}\right) \quad (8)$$

方法２

$$C_{ij} = \begin{cases} \sum_{k=1}^{l} \max\left(f_{ck}^{(i)}, f_{ck}^{(j)}\right) & \text{条件１} \\ \sum_{k=1}^{l} \min\left(f_{ck}^{(i)}, f_{ck}^{(j)}\right) & \text{条件２} \end{cases} \quad (9)$$

但し，条件１はクエリとテスト記事との比較であり，条件２はテスト記事どうしの比較である．

方法３

$$C_{ij} = \sum_{k=1}^{l} \frac{f_{ck}^{(i)} + f_{ck}^{(j)}}{2} \quad (10)$$

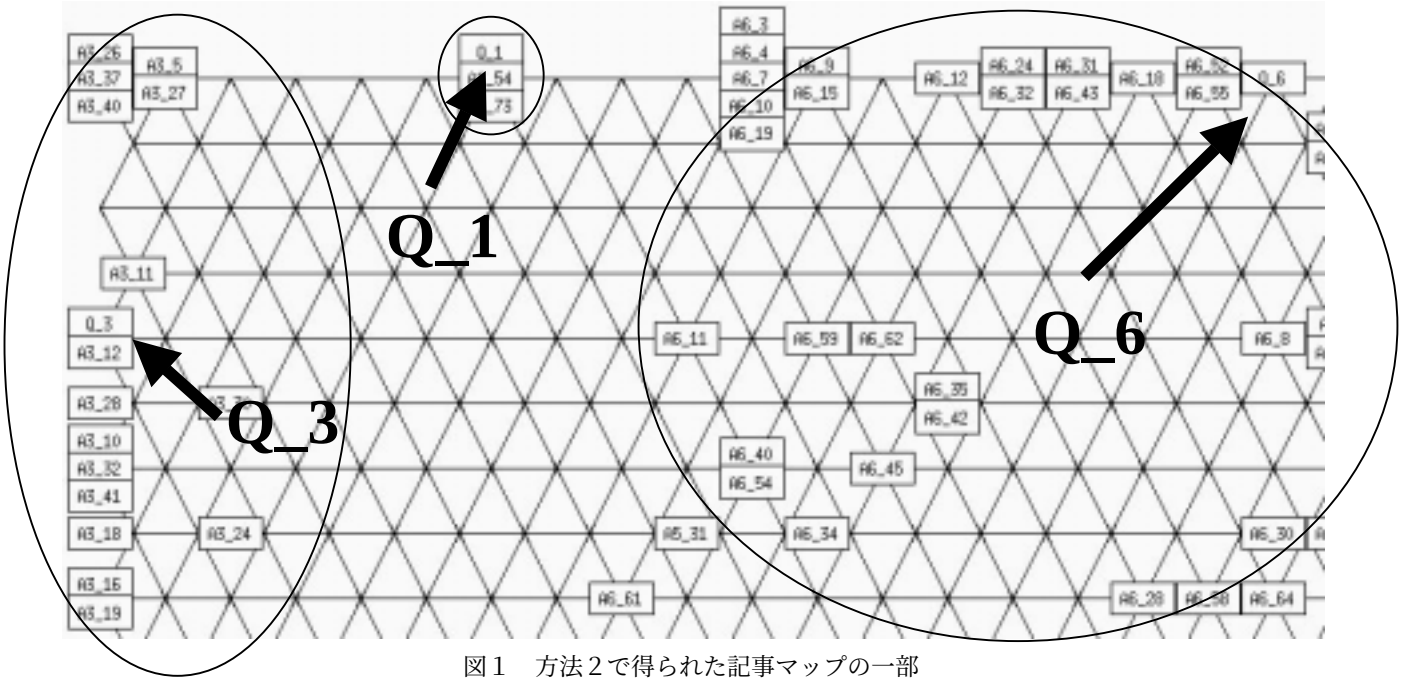

図１　方法２で得られた記事マップの一部

方法４

$$C_{ij}=\begin{cases}\sum_{k=1}^{l}\max\left(f_{ck}^{(i)},f_{ck}^{(j)}\right) & \text{条件1}\\ \sum_{k=1}^{l}\dfrac{f_{ck}^{(i)}+f_{ck}^{(j)}}{2} & \text{条件２}\end{cases} \qquad (11)$$

但し，条件１はクエリとテスト記事との比較であり，条件２はテスト記事どうしの比較である．

３　実験結果

3.1　データ

実験ではクエリは 6 個とそれらの適合記事 433 個使用した．適合記事の分布は以下の表で示す．

表１　実験に用いた適合記事の分布

a_1	a_2	a_3	a_4	a_5	a_6
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には方法２を用いて得られた記事マップの一部を示す。このマップを見ればわかるようにクエリ Q_3 とその適合記事 A3_*，Q_6 とその適合記事A6_*が互いに近い位置に配置された. しかし，クエリ Q_1 についてはわずかな適合記事しか，その近隣にマップされていない．このマップから個々のクエリの記事との距離を測り，最も近い 10 個の記事を取り出した結果は表２に示す．そして，表３は４つの方法で求め，それぞれ表２のように 10 個取り出した時の精度を示す.

表２　方法 2 を用いて得られた情報検索

Q_1	A1_73 A1_54 A6_19 A6_10 A6_7 A6_4 A6_3 A6_15 A6_11 A6_9
Q_2	A2_6 A2_27 A2_62 A2_66 A2_5 A2_3 A2_50 A2_85 A2_58 A1_66
Q_3	A3_12 A3_28 A3_11 A3_41 A3_32 A3_10 A3_30

	A3_18 A3_24 A3_40
Q_4	A4_72 A4_67 A4_13 A4_91 A4_71 A4_44 A4_73 A4_77 A4_75 A4_12
Q_5	A6_56 A5_42 A6_21 A6_20 A6_55 A6_52 A6_49 A6_5 A6_18 A6_8
Q_6	A6_55 A6_52 A6_21 A6_20 A6_18 A6_43 A6_31 A6_56 A6_32 A6_24

表 3　情報検索の精度

方法 1	方法 2	方法 3	方法 4
0.55	0.70	0.17	0.67

3.4　考察

まず明らかに方法３では良い結果が得られていない．クエリどうしの距離を最大に設定しているにもかかわらず，作成されたマップ上では全てほぼ同じ位置になっていた．これは計算方法３ではそれぞれの記事の類似度に差がなくなることに原因があると考えられる．

方法１の結果はクエリ１と２が全くの不正解であり，全体のマップを見る限りではすべてのクエリがそれほど離れていない．この方法はクエリとテスト記事の名詞の数がクエリのほうが少ないので試してみた計算方法であるが，クエリと記事の名詞の数に差がありすぎると問題が起こる場合がある．例えばクエリの名詞の数が５個でテスト記事の名詞の数が 10 個ほどならそれほど問題はないと考えられるが，テスト記事の名詞の数が 50 個など差が出ると，適合記事でなくてもクエリの名詞と一致する名詞を含む記事が出る確率が高くなる．名詞の出現頻度は個数を総数で割ったもので表しているので，もちろんクエリの値のほうが高い．すなわち，テスト記事の名詞の数が多くなるにつれて適合記事かどうかの区別がつかなくなる．また，クエリの名詞の数が記事のそれに比べ少ないという問題は単純にクエリの名詞をその中の同じ名詞を用いて増殖させても，テスト記事の名詞の数も様々なので解決にはならない．

方法２と４の計算方法はそれほど変わらないが方法２のほうが正解率は高かった．テスト記事どうしの比較において頻度の低い値を取るか，平均値を取るかの違いで精度が変わることがわかり，今後様々な比較においての検証が必要である．

4　おわりに

本稿は神経回路網に基づく情報検索手法を提案した．提案手法を用いれば，従来の検索結果（ランク付けの適合記事リスト）が得られるだけでなく，２次元上の可視的連続的な情報検索結果も同時に得られる．小規模の実験においては提案手法が高い検索精度を有すことが確認された．

本研究を通じ，提案手法を用いて高精度な情報検索を実現するためにはまず，如何に異なるクエリをマップ上の離れる位置にマップできるかが重要であることが分かった．また，情報検索においてすべてのクエリを同時に与えることは本来あまり現実的ではない．以上のことを一括に解決するために，今後は，記事マップを先に自動構築し，それにクエリでラベル付ける方法を考案していく予定である．

参考文献

[1] Menzel，H: Information needs and uses in science and technology, *Annual Review of Information Science and technology*，1，pp.41-69, 1966.

[2] 徳永健伸：情報検索と言語処理，東京大学出版会，1999.

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

[4] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正幸：形態素解析システム『茶筌』version2.3.3 使用説明書，2003 年 8 月