社会情報学研究, Vol. 9, 99-106, 2003

自己組織化マップを用いた日本語処理の試み

尾崎数也*，薮兼智英*，井上正人**，前原俊信***，岡　隆光****

Japanese Document Processing Using Self-Organizing Maps

Kazuya Ozaki *，Tomohide Yabukane *，Masato Inoue **
Toshinobu Maehara ***，Takamitsu Oka ****

In this paper, we propose a Japanese document processing system by using Self-Organizing Maps (SOM). Japanese documents are represented on a map and similarity relations between documents are visualized. Our system contains three kinds of parameters. The first is the dimension of the vectors, which expresses the Japanese words. The second (third) is the size of the word category map (document map). We obtain optimal values for these parameters by classification of the two different kinds of documents.

Key Words (キーワード)
Japanese Document（日本語の文書），Processing（処理），Self-Organizing Maps（自己組織化マップ），Search（検索），Internet（インターネット）

1. はじめに

情報技術の進歩により，インターネットの利用が益々盛んになってきている．インターネット上にはホームページを始め様々な種類の Web 情報があり，これらを有効に活用することで学習や仕事の能率を上げることが出来るからである．これらの Web 情報を利用する上で，情報検索システムが大きな役割を果たしている．良く使われている情報検索システムとして，例えば，goo[1] や Namazu[2] がある．goo は全文検索型サーチエンジンを使ってインターネットで結ばれている様々なサイトの Web 情報を検索するものであり，Namazu は導入されているコンピュータに含まれている情報を Web からでも検索を可能にした日本語全文検索システムである．これらのシステムでは検索者が検索語(キーワード)を入力して検索を行うため，キーワードに対する知識が十分でない場合は適切な検索結果を得ることが難しい．

情報検索を効率良く行うために，キーワード検索とは異なる方法の開発が求められている．Kohonen 達の研究グループは，検索される情報の特徴を基に分類し，類似情報の収集を容易にした検索システム：WEBSOM（World Wide Web Self-Organizing Map)[3] を開発し，試験的に運用している．WEBSOM は，インターネットのニュースグループ上に登録された 100 万以上の文書を自己組織化マップ(SOM)[4] を用いて文書間の距離を測定し，関連する文書を近くに配置するようにマップをつくるシステムであり，文書のタイトルや内容

* 呉大学大学院社会情報研究科（Graduate School of Social Information Science, Kure University）
** 海上保安大学校（Japan Coast Guard Academy）
*** 広島大学大学院教育学研究科（Graduate School of Education, Hiroshima University）
**** 呉大学社会情報学部（Faculty and Graduate School of Social Information Science, Kure University）

がハイパーリンクによって表示できるようになっている．このシステムは，英語の文書の情報検索システムであり，検索者が入力した英語の文書に対して，類似した内容の文書を検索し，表示するようにできているもので，キーワード検索にかわる検索システムである．

この論文は，SOM を用いた日本語処理システムの構築に関するものである．構築したシステムを有効に機能させるために，分野が異なる 2 種類のメールグループのニュースを用いてこれらの文書を分類し，システムに含まれている変数の最適な値を求めることを試みたものである．

論文の構成は，次の通りである．2.では我々が行った日本語処理の方法について，3.では単語辞書及び単語分類マップの作成，4.では文書分類マップの作成について，5.では結果がそれぞれ示されている．最後の 6.では，今後の課題が述べられている．

2. 日本語処理の方法

日本語を処理するためには，数々の手順を踏む必要がある．それらは，単語辞書及び単語分類マップの作成，文書分類マップの作成の 2 つに大きく分けることができる．これらの過程では，おおよそ次のような処理が行われる．

【単語辞書及び単語分類マップの作成】

事前処理（文書の分かち書きなど）

①分類に用いる文書の収集．

②収集した文書を分かち書きし，単語を抽出．

③出現頻度が極端に多い単語（助詞など）と少ない単語の除去．

単語辞書の作成（単語のコード化）

①単語をベクトルで表すため，ベクトルの次元数（n次元）を決定し，その成分を乱数で生成．

②各単語の直前に現れる単語と直後に表れる単語の成分すべてを各々平均化し，それぞれの単語成分の前と後ろに追加して単語を 3 n 次元のベクトルで表し，単語辞書を作成．

単語分類マップの作成

① SOM の 2 次元平面の縦横のユニットの数（P×P (W)）を決定．

②単語辞書を SOM を用いて 2 次元平面上にマッピングして，単語分類マップを作成．

【文書分類マップの作成】

①事前処理したそれぞれの文書について，単語分類マップの各ユニットに含まれている単語に何回ヒットしたかをカウントし，それぞれの文書ごとにヒストグラムを作成．

②ヒストグラム化された文書をマッピングする SOM の縦横の大きさ（Q×Q (D)）を決定．

③ヒストグラム化された文書を SOM で 2 次元平面上にマッピングして，文書分類マップを作成．

ここで述べたことを図示すると図1のようになる．単語分類マップは 2 次元平面上へのマッピングであり，縦（横）のユニットの数が P（P）のものを P×P (W) と下付の（W）をつけて表し，文書分類マップの場合 Q×Q (D) と区別して表す．

3. 単語辞書及び単語分類マップの作成

ここでは，2.で述べた単語辞書の作成及び単語分類マップの作成について要点を詳しく説明する．

3-1 事前処理（文書の分かち書きなど）

コンピュータを用いて言葉を処理する場合，単語を取り出す必要がある．英語の場合は単語がスペースで区切られているのでこの操作は容易であるが，日本語の場合は区切られていないので容易ではない．日本語の文書を処理し，単語毎に区切る作業を分かち書きと言う．良く用いられている分かち書きソフトには，KAKASI [5] や ChaSen [6] がある．最近になって，MeCab（和布蕪）[7] が注目されている． MeCab は，入力として与えられた文字列を形態素の列に変換する形態素解析（Morphological Analysis）を行いながら文書を分かち書

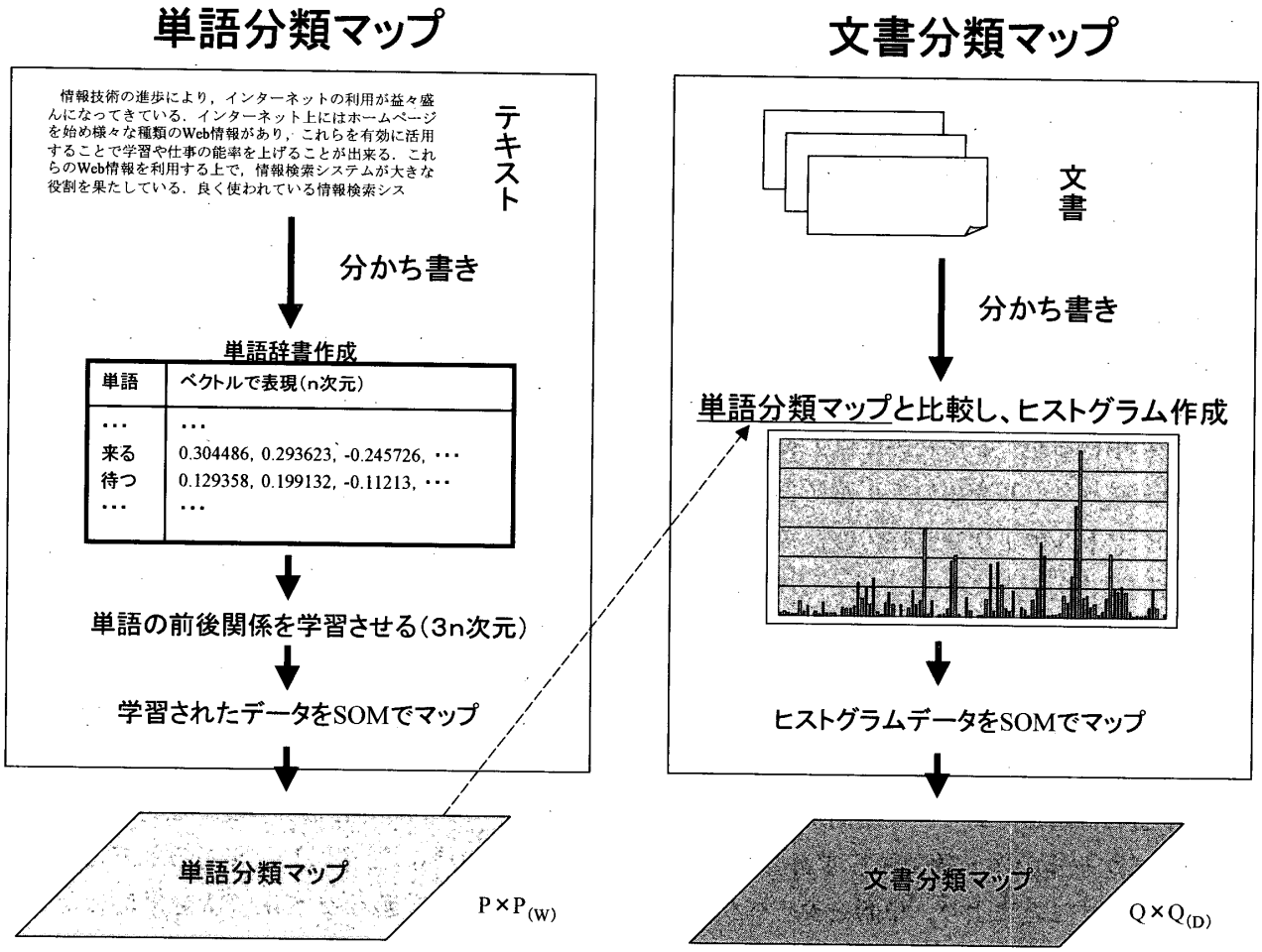

図1　単語分類マップと文書分類マップの作成

きしており，分かち書きの精度が良く，しかも，処理速度が KAKASI や ChaSen よりも数倍早い．また，MeCab は将来，多国語（英語，中国語など）対応予定であることから，将来の拡張性を考えて，我々は MeCab を使用することにした．

本システム作成に使用した文書である，「小泉メールマガジン」と「日経 BP ITPro 記者の目」は共に電子化された文書であり，それぞれの URL から取得した．[8] これらの文書は，文の途中に改行コードが含まれているので，文の終わりにだけ改行コードが付くようにデータを変換した．我々は，テキスト処理言語 Perl を使ってフィルタリングスクリプトを作成し，MeCab で分かち書きしやすいように，全角英数を半角英数へ，半角カナを全角カナへ，全角記号を半角記号へ変換し，特殊記号を取り除いた．

次に，フィルタリングされた文書を MeCab で分かち書きし，文書それぞれに含まれる単語とその単語が現れた重複回数を C 言語で作成したプログラムによって求めた．ここで，出現頻度が極端に大きな単語と極端に小さな単語は取り除く処理を行った．処理をした「小泉メールマガジン」と「日経 BP ITPro 記者の目」の文書の数は，共に 114 文書であった．表 1 には，取り除く処理を行う前と後の 1 つの文書に含まれる平均の単語数が，取り除く基準と共に示されている．取り除いた単語

表 1　文書あたりの平均単語数等

	1 文書あたりの平均単語数		取り除く基準
	取り除く前	取り除いた後	
小泉	2580	374	頻度 1500 回以上と10 回未満
日経	1686	326	

で最も多いのは「改行コード」であり，次に助詞の「の」，「を」，「に」が続いている．また，名詞では，「小泉」，「内閣」及び「日本」の3つである．これらの処理をした後に残った単語は，全部で6632個である．

3-2 単語辞書の作成

a 乱数による単語ベクトルの作成

我々は，多次元のベクトルを用いて単語を表現する方法を用いるが，このためには工夫が必要である．単語をベクトルを用いて正確に表現するためには，ベクトルの1次独立性の性質により，単語の数と同じ次元を持つベクトルが必要になる．例えば，これから扱うところの6632個の単語を表すのには，互いに直交する6632次元の単位ベクトルを6632個用意する必要がある．次元数がこのように大きなベクトルをコンピュータを用いて処理するのは困難であり，実用的ではない．このため，ベクトルの次元数を減らす必要がある．次元数を減らすことにより，単語を表すベクトル間に重なりの部分が生じ，単語同士を完全に分離することが出来なくなるが，ベクトルの成分を乱数で生成することにより，実用上十分に分離されたベクトルを作ることができる．

英文の文献検索システムであるWEBSOM[3]は，英単語に90次元のベクトルを割り当てている．日本語処理においては，何次元のベクトルが適当かは明らかではないので，我々は，60次元，80次元，100次元及び120次元のベクトルを用いて単語ベクトルを作成し，文書を処理し，その結果を比較検討して最適な次元数を求めることにする．

単語ベクトルの成分は，WEBSOMと同様にそれぞれ乱数を生成して求めるが，ベクトルの規格化には注意を払った．我々の計算には，GNU Scientific Libraryに含まれている球面上のランダムベクトル発生方法を用いた．[9] この方法では，ベクトルの大きさを1に規格化する条件をつけて乱数を生成している．その条件は次の通りである．

$$x_1^2 + x_2^2 + \cdots + x_n^2 = 1$$

b 単語辞書の作成

上で求めた単語（単語ベクトル）は，成分が乱数で与えられているため，各単語間には意味のある関係は存在せず，このままでは単なる記号に過ぎない．そこで，単語の前後に現れる単語の情報をその単語に与えるという方法でこれらの単語に意味を持たせることにする．このため，次のような方法で単語ベクトルを拡張し，意味を持った単語（意味を持った単語ベクトル）を作る．

$$X(i) = \{a(i), 0.2x(i), b(i)\}$$

ここで，$x(i)$ は乱数で求めたところの i 番目の単語ベクトルを表し，$a(i)$と $b(i)$は $x(i)$と同じ次元数のベクトルとする．例えば，$x(i)$が100次元のベクトルであれば，$X(i)$ は300次元のベクトルになる．ベクトル $a(i)$と $b(i)$との成分は，それぞれ，i 番目の単語に注目し，文書の中でこの単語の前に現れる単語ベクトルの値を平均したものが，$a(i)$ 後に現れる単語のベクトルを平均したものが $b(i)$ というように求められる．このような方法で言葉の学習をモデル化し，システムに取り入れるのである．$x(i)$に付いている係数0.2は，学習で得た $a(i)$と $b(i)$の値とのバランスを調整するために導入されたものである．ここでは，WEBSOM[3]で使われた値と同じ値を用いることにする．

さて，全ての i について $a(i)$と $b(i)$ を求めることにより，単語辞書が完成する．この論文では，このような方法で6632個（i=1 ～ 6632）の単語について学習し，単語辞書を作成している．

3-3 単語分類マップの作成

前で求めた，単語辞書をSOMを用いて2次元表面上にマッピングし，単語分類マップを作成する．本論文では，SOMに含まれるユニット数が100（10×10(w)），400（20×20(w)）及び900（30×30(w)）の場合の，計3種類の単語分類マップを作成し，ユニット数の違いによって文書の分

類がどう変化するのかを検討していく．単語数が6632個の場合，各ユニットには平均66個(100ユニットの場合)，17個(400ユニット)，7.4個(900ユニット）の単語が含まれており，個数の違いが文書分類に与える影響を調べたのである．

4. 文書分類マップの作成

ここでは，文書分類マップの作成過程を説明する．

4-1 ヒストグラムの作成

分かち書きした文書に含まれている単語が単語分類マップユニットのどのユニットに該当するのかを調べ，重複回数を計算し，ヒストグラムを作成する．作成されたヒストグラムは，単語分類マップのユニットの座標をx軸，重複回数をy軸として表すことにした．我々は，228の文書（小泉メールマガジン 114，日経 BP ITPro 記者の目 114）を扱っているので，228個のヒストグラムを求めることになる．

4-2 文書分類マップ作成

上で作成されたヒストグラムのデータを SOM を用いて，2次元平面上にマッピングして，文書分類マップを求める．ここでは，文書を分類した結果が，SOM のユニット数によってどう変化するのかを調べるために，ユニットの数が 49 ($7\times7_{(D)}$) と 100 ($10\times10_{(D)}$) の場合について文書分類マップを求め，結果を比較していく．

5. 結　　果

今までに述べてきたように，単語辞書及び単語分類マップの作成，文書分類マップの作成には，調節すべき変数が含まれている．これらの変数の最適な値を求めるために，3.でふれた「小泉メールマガジン」と「日経 BP ITPro 記者の目」の2種類のメールマガジンを使用した．まず，単語がどのように分類されているのかを見るために，単語分類マップを調べ，次に2種類のメールマガジンを分類し，分離出来ているかどうかを調べた．2種類のメールマガジンは，取り扱っている内容が異なっている．「小泉メールマガジン」は，小泉総理のメッセージや大臣の本音トーク，小泉内閣の動きなどが書かれており，政治関連のニュースが多い．「日経 BP ITPro 記者の目」は，IT 関係の記事を扱っており，記者や読者のさまざまな意見が述べられている．そして，表現形式はかなり自由であり，討論形式の文書も含まれている．

この論文での SOM の計算には，SOM プログラムパッケージを用いることにする．[10]

5-1 単語分類マップの結果

単語分類マップにおいて，単語がどのように分類されているのかを調べるために，県を表す地名が含まれているユニットに注目した．県を表す地名は，多くの場合，広島県や岡山県など，地名と県（広島と県，岡山と県など）が隣り合って使われている．その結果，単語分類マップの同じユニットに県を表す地名が集まることになる．

表2は県を表す地名が最も多いユニットに含まれる単語数（分母）と県を表す地名数（分子）を求めたものである．表の横方向は，単語ベクトルの次元数（それぞれ60次元，80次元，100次元，120 次元）であり，縦方向は，単語分類マップのユニットの数（100ユニット，400ユニット，900ユニット）である．それぞれの次元数とユニット数について，乱数のシードを5回変えて計算したので，それぞれに5組の数値が記入してある．

この表から，単語分類マップのユニットの数が100 の場合は，県を表す地名の他に別の単語が混じっており，900 の場合は地名だけが集まっていることが分かる．これは，各ユニットに含まれる平均の単語数の違い，すなわち，66 個（100 ユニットの場合）と 7.4 個（900 ユニット）の違いによるものと考えられる．地名の集まり方は，単語を表す次元数にも依存している．ユニットに含まれる県を表す地名の割合が最も大きかったのは，900ユニットと100次元の場合であり，5回の計算の平均値は 0.996 であった．

表2 1つのユニットに含まれる単語数（分母）と県を表す地名の数（分子）

	60次元	80次元	100次元	120次元
100ユニット 10x10(W)	24/123	25/123	19/53	26/109
	24/88	24/118	25/143	26/121
	26/173	26/133	25/149	25/89
	23/125	25/159	25/89	26/57
	25/178	25/99	25/149	25/54
400ユニット 20x20(W)	26/50	24/39	25/29	25/30
	24/30	23/32	24/51	25/58
	25/44	23/31	25/38	25/28
	25/30	23/51	23/34	26/36
	23/29	25/44	24/29	25/39
900ユニット 30x30(W)	22/27	24/30	23/24	22/29
	22/23	24/29	22/23	25/25
	23/23	21/35	22/23	24/25
	21/27	22/24	20/20	23/23
	24/25	22/24	23/24	22/26

このことから，単語分類マップのユニットの数と単語を表すベクトルの次元数は，それぞれ 900 ユニットと 100 次元が適当であると考えられる．

5-2 文書分類マップの結果

次に文書分類マップで文書がどのように分類されるかを調べるため、表 2 を計算した単語分類マップを用いて，各文章毎のヒストグラムを計算した．次にヒストグラムの値を SOM に入力し，文書分類マップを求める方法で，文書を分類した．計算に用いた文書分類マップのユニットの数は 49（7×7(D) と表す）と 100（10×10(D)）である．49 ユニットの場合の結果を表 3 に示す．

表3 2種類の文書の分類

	60次元	80次元	100次元	120次元
100ユニット 10x10(W)	1	2	1	2
	2	2	1	0
	1	1	3	0
	0	1	1	0
	1	0	2	0
400ユニット 20x20(W)	0	0	1	0
	0	0	0	1
	1	0	0	0
	1	1	0	0
	0	0	1	0
900ユニット 30x30(W)	0	1	0	0
	0	0	0	0
	0	0	0	0
	0	0	0	1
	0	0	0	0

この表は，2 種類の文書を文書分類マップで分類したときに，それぞれのグループとは異なるグループに分類された文書の数を表している．この表からも 100 ユニットの場合は，文書の分類が十分出来ていないことが分かる．また，400 ユニットと 900 ユニットではあまり差がないが，僅かに 900 ユニットの方が良いことが分かる．さらに，単語ベクトルの次元数への依存はあまり大きくないことも分かる．同様の計算を，文書分類マップのユニットの数が 100(10×10(D))の場合にも行ったが，結果は似たようなものであった．すなわち，文書分類マップのユニットの数が 49（7×7(D)）と 100（10×10(D)）の場合であまり差がなかったことである．

単語分類マップと文書分類マップの結果から，単語分類マップの SOM のユニットの数は 900（30×30(W)）が適切であり，単語を表すベクトルの次元数は 100 近辺の数で良いことが分かった．英語の文書を扱う WEBSOM は 90 次元のベクトルを用いているが，日本語でも似たような次元数のベクトルを用いて良いことが分かった．図 2 には，100 次元の単語ベクトル，900 ユニット（30×30(W)）の単語分類マップ，49 ユニット（7×7(D)）の文書分類マップで作成した 228 の文書を分類した結果が示されている．ここで，K022（I023）は「小泉メールマガジン」（「日経 BP ITPro 記者の目」）の 22 番目（23 番目）の文書を表している．この図では，類似した文書が近くに集まっており，例えば I010 から I107 までの文書は類似性が強く同じユニットに属している．ユニットに付いた色は，濃い色（薄い色）は隣のユニットとの距離が遠く離れている（あまり離れていない）ことを示している．この図から，「小泉メールマガジン」と「日経 BP ITPro 記者の目」の文書が完全に 2 つのグループに分離されていることが分かる．

分類の有効性を確かめるため，「小泉メールマガジン」と「日経 BP ITPro 記者の目」の最近の文書を新らたにそれぞれ 3 つ用意し，図 2 のどのユニットに属するのかを調べた．結果は，それぞれ

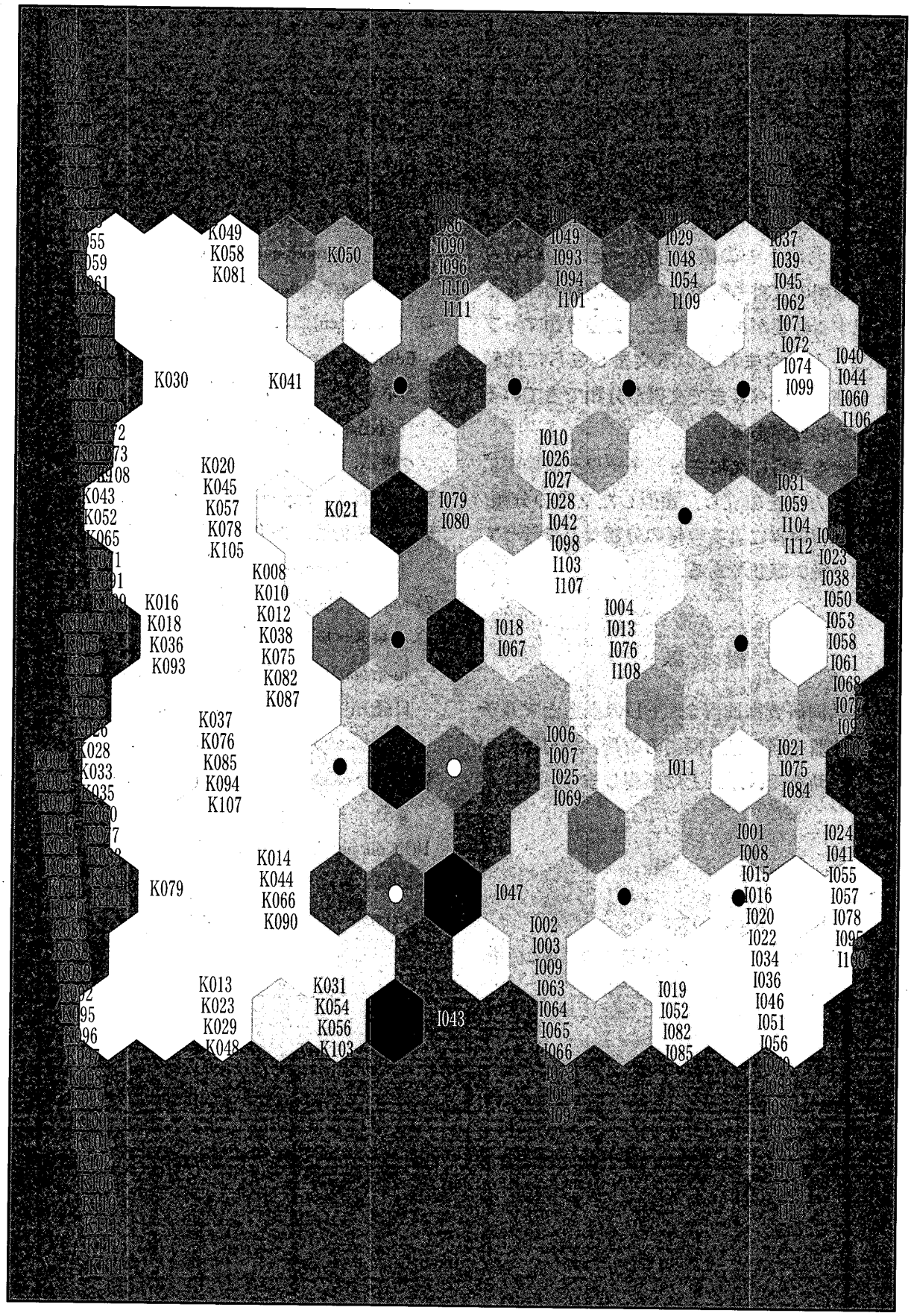

図 2　文書分類マップ（100_30_7）

のグループに属していることが分かった.

6. 今後の課題

この論文では，内容の異なる 2 種類の文書を SOM を用いて分類することを試みた. 日本語を処理するのに適切な単語ベクトルの次元数，単語分類マップのユニット数及び文書分類マップのユニット数のおおよその値が分かった. これからは，別の種類の文書も解析して，その有効性をさらに確認する必要がある. また，文書分類マップの各ユニットに含まれている文書をさらに比較し，より細かなレベルまで文書が分類できているかを調べる必要がある.

この研究をさらに発展させ，キーワード検索に変わる検索方法を確立し，類似した文書の分類，類似した学習内容による教材の分類などを行っていくのが今後の課題である.

謝　辞

呉大学共同研究推進資金（FDの推進とマルチメディア機器を利用した授業の改善、教材開発の研究）の援助を受けたことに感謝します。

文　献

1) http://www.goo.ne.jp/.

2) 馬場　肇，2001，Namazu システムの構築と活用，ソフトバンク.

3) http://websom.hut.fi/websom/.

Kaski S., Honkela T., Lagus K., Kohonen T., 1996, *Creating an Order in Digital Libraries with Self-Organizing Maps, Proceedings* of WCNN'96 World Congress on Neural Networks, Lawrence Erlbaum and INNS Press, 814.

Kaski S. 1998, *Dimensionality Reduction by Random Mapping: Fast Similarity Computation for Clustering*, IEEE International Joint Conference on Neural Networks, Anchorage.

4) T.Kohonen，1997，*Self-Organizing Maps* Second Edition，Springer.

コホネン T.，徳高平蔵・岸田悟・藤村喜久郎訳 1996，「自己組織化マップ」，シュプリンガー・フェアラーク東京.

5) 佐藤雅彦，http://kakasi.namazu.org/.

6) 松本祐治他，2000，形態素解析システム茶筅，http://chasen.aist-nara.ac.jp/.

7) MeCab，http://cl.aist-nara.ac.jp/~taku-ku/.

8) 小泉メールマガジン，http://www.kantei.go.jp/jp/m-magazine/backnumber/index.html.

日経 BP ITPro 記者の目，http://itpro.nikkeibp.co.jp/members/backnumber/200309/bnsearch.jsp/.

9) GNU Scientific Library http://www.gnu.org/software/gsl/.

10) Kohonen T., Hynninen J., Kangas J., Laaksonen J., 1995 *The Self-Organizing Map Program Package Version 3.1* Helsinki University of Technology.