

自己組織化マップ (SOM) による文書の分類とキーワードの抽出

藪兼智英*, 井上正人**, 金 明旭*, 岡 隆光***, 前原俊信****

Classification and Extraction of Keywords for Japanese Documents Using Self-Organizing Maps

Tomohide Yabukane*, Masato Inoue**, Jin Mingxu*
Takamitsu Oka***, Toshinobu Maehara****

By using Self-Organizing Maps (SOM), we propose a Japanese document processing system. Japanese documents are represented on a map and similarity relations between documents are visualized. For each group of the documents, extraction of keywords has performed. Comparison between our method and a simple keywords search system has been done.

Key Words (キーワード)

Japanese Document (日本語の文書), Processing (処理), Keyword (キーワード), Self-Organizing Maps (自己組織化マップ), Search (検索)

1. はじめに

我々の研究グループは、自己組織化マップ (SOM)¹⁾を用いた日本語処理について研究を続けてきている。前回の論文では、分野の異なる2種類のメールグループのニュースを用いて単語分類マップと文書分類マップを作成し、これらの文書を分類することによりシステムに含まれる変数の最適な値を求めた²⁾。その結果、単語分類マップではユニット数が900ユニット、単語ベクトルの次元数は100次元が適当であり、文書分類マップではユニット数が49ユニットの場合と100ユニットの場合ではそれほど差が無く、また、単語ベクトルの次元数への依存はあまり大きくなく100次元で良いことが分かった。そして、これらの数値を用いることにより、2種類のメールグループのニュースは完全に2つのグループに分離され、さ

らに文書分類マップ上では類似した文書が近くに集まっていることを示した。

今回は、これらの研究を発展させ、類似した文書のグループを代表するキーワードを自動的に求める方法について述べるものである。

論文は次のように構成されている。2.では「キーワードの抽出法」について、3.では我々が行った「日本語処理の方法」について、4.では「結果」がそれぞれ示されている。最後の5.では、「今後の課題」が述べられている。

2. キーワードの抽出法

キーワードは、文書の内容を代表するものであり、キーワード検索など文献検索に用いられている。従来は、文書の内容を人間が理解し、その文書に対してキーワードを付与する方法をとってい

* 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)

*** 呉大学社会情報学部 (Faculty and Graduate School of Social Information Science, Kure University)

**** 広島大学教育学研究科 (Graduate School of Education, Hiroshima University)

た。しかしこの場合、付与されるキーワードはそれを付与した人間の感性にゆだねられることになり、検索する人が想定しているキーワードと異なる場合が生じるなどの問題点が含まれている。また大量の情報が流通する高度情報社会を迎えた現在、人の手でキーワードを付与するには限界が生じている。

このような中で、文書からキーワードを自動抽出する方法が考えられ、様々な研究が進められている。これらには形態素解析 (Morphological Analysis) を用いたものや N-gram を用いたものなどがある³⁾。これらは、多くの場合、文書中でてくる単語の出現頻度を調べ、出現頻度の高い単語をキーワードとして抽出する方法を採っている。この方法は文書からキーワードを抽出しやすいという利点はあるが、助詞などキーワードになり得ない単語が多数出現するという問題がある。また、共通のキーワードを持った文書が、類似性の高い文書であるとは限らないことも問題である。

先に述べたように、SOM は類似した文書を近くに配置することが出来るので、我々は SOM を用いて文書を分類し、グループ化した。そして、グループ化された文書中に含まれている単語のうち出現頻度の高い単語を求め、それらをキーワードとして抽出することにした。ここで求めたキーワードはグループ全体を代表するキーワードということが出来る。この方法で問題となるのは、形態素解析などの場合と同じように助詞などキーワードになり得ない単語をいかに上手に取り除くかである。次に、これらを含めた具体的な日本語処理の方法について述べて行く。

3. 日本語処理の方法

SOM で日本語を処理する場合、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。単語分類マップや文書分類マップの作成方法については、前回の論文²⁾に詳しく述べてあるのでそちらを参照することと

し、ここでは、今回特に注意を払った日本語文書の事前処理、分かち書きした文書に含まれる助詞などをどの段階で取り除くのが良いのか、について述べる。

前回の論文では、文書を分かち書きし、全文書(228文書)に含まれる単語のうち、出現回数10回以下と1,500回以上の単語を除去する条件(以降条件①と表記)をつけて、単語分類マップを作成し、文書分類マップを作成した。これと同じ方法でメールグループのニュースである「IT Pro 記者の目」⁴⁾(今回は、234文書をURLから取得)を用いて、同じ種類の文書が内容によってどう分離されるかを調べ、キーワードの抽出を試みた。同じグループに属する文書を比較するために、文書分類マップ作成の過程で作成した2次元ヒストグラムを比較し、出現頻度の高い単語の現れ方を調べた。出現頻度の高い単語(言葉)に例えば「であるが」などが含まれ、これらをキーワードとするには問題があることが分かった。

次に、これらの単語(言葉)を除いて単語分類マップを作成することを試みた。このため、条件①に加えて、以下の条件を(以降条件②と表記する)を追加し単語の前後関係を学習させた。結果は、良くなかった。

- 1, 何桁でも数字だけの単語は消す。
- 2, 記号のみの単語は消す。
- 3, アルファベットなどの1バイト文字やカタカナ1文字の単語は消す。
- 4, 何文字でも平仮名だけの単語は消す。

原因は、条件②を追加することにより、文書の中で除かれる単語数が増加し、これを除いて単語の前後関係を学習することになり、学習がうまくいっていないものと考えられる。

そこで我々は、前回のように単語分類マップ作成の段階で入力する単語をすべて決めてしまうのではなく、まず単語分類マップの学習のため、ある程度の前後関係を保った単語数を確保し、文書分類マップ作成時において単語分類マップ上から条件②に該当する単語を除去し文書分類マップを

作成する方法を採用した。

また、2 次元ヒストグラムを文書の比較検証に用いることにした。図 1 に今回行った日本語処理のフローチャートを示す。

なお、単語分類マップと文書分類マップのユニット数は前回の結果を基にそれぞれ 900 ユニットと 100 ユニットで行い、単語ベクトルの次元数は 100 次元(前後関係の学習を行った後には 300 次元)で計算を行い、SOM の計算には Viscosity SOMine 4.0 Plus⁵⁾ を用いることにする。

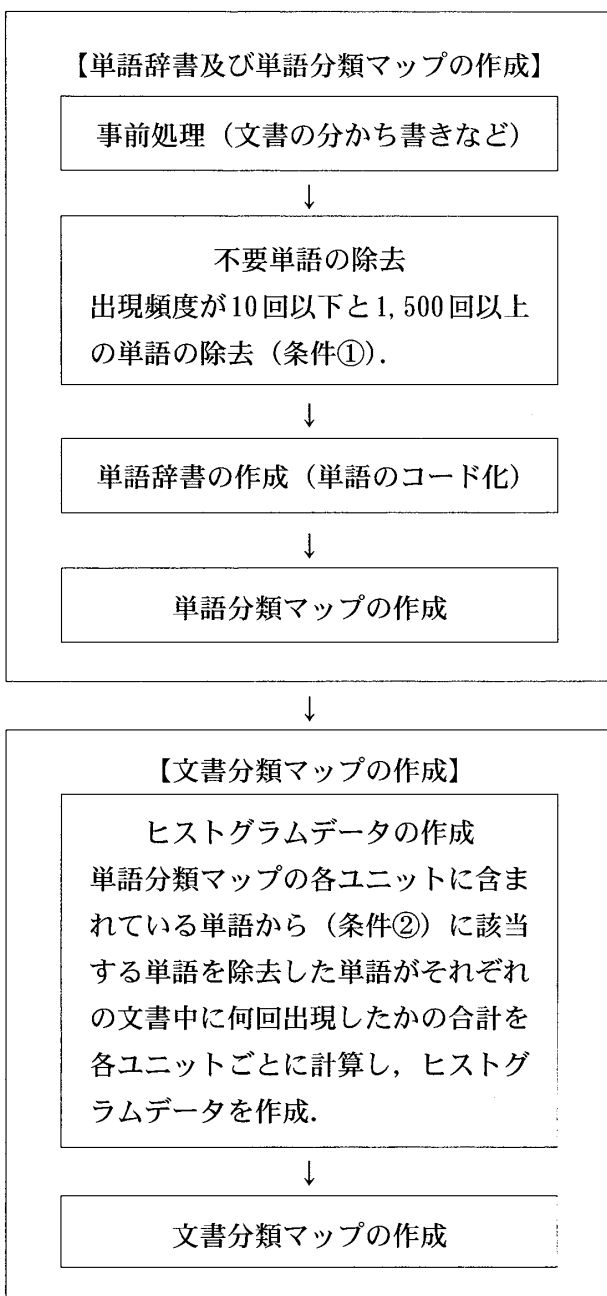


図 1 日本語処理フローチャート

4. 結 果

今回は、合計 234 の文書を扱うが、これらの文書を MeCab（和布蕪）⁶⁾ で分かち書きした結果、合計 15,904 個の異なる単語が存在した。そして、条件①を適用して除去された単語数は 12,674 個であり、文書分類マップ用ヒストグラムを条件②を適用して作成した結果、2,726 個の単語が残った。我々は、この 2,726 個の単語を用いてキーワード抽出を行った。

さて、上の数は、前回の論文²⁾での単語数 6,632 個と比較すると約半分と少ない。この理由は、今回は一種類のメールマガジンの文書しか扱っていないので、文書の数も 234 と増やしても、文書の中に現れる単語は同じ単語が多く、単語の種類が少ないことによると思われる。文書分類マップ用ヒストグラムに現れる単語数を増やすため、条件①を緩和して 5 回以下を除去するようにした結果、4,473 個の単語が残った。この、4,473 個の単語を用いて文書の分類を行ったが、同じグループの中に関連性が薄い文書が属しており、結果は良くなかった。

2,726 個の単語を用いて作成した文書分類マップを図 2 に示す。なお、図中の数字は文書の番号を示す。番号は、1～114、200～320 と振っている。1～114 は前回用いた文書と同じ文書であり、200～320 は今回新たに増やした文書である。なお、262 は欠番である。

文書分類マップ上に類似した文書が配置されているかどうかを調べるため、我々は同じユニットに属する文章を比較した。次に、文書分類マップ上のユニット 38 番目にグループ化されている文書番号 023, 061, 092, 291, 299 について述べる。これらの文書は、1 文書あたり 1063～1259 単語ある。よって、これらの文書をすべて掲載することは無理なので、表 1 にタイトルと冒頭数行を掲載する。表 1 から分かるように、ここに現れる文書は、全て個人情報についての文書であり、グループ化が上手くいっていることを示している。さらに、それぞれの文書の 2 次元ヒストグラムを

図 3 に表す. 2 次元ヒストグラムの各データは, 単語分類マップ上の単語から条件②を除いた単語と, 文書中に含まれる単語が一致する単語分類マップ上のユニットに対し, 該当する単語の文書中の出現頻度を各ユニットごとに合計し, 全体を

1に規格化した値を示している. これらの 2 次元ヒストグラムから, それぞれ類似性が高いことが分かる. ここで, グループ合計とあるのは, ユニット 38 に属する 5 つの文書の 2 次元ヒストグラムデータを合計して求めた 2 次元ヒストグラムの

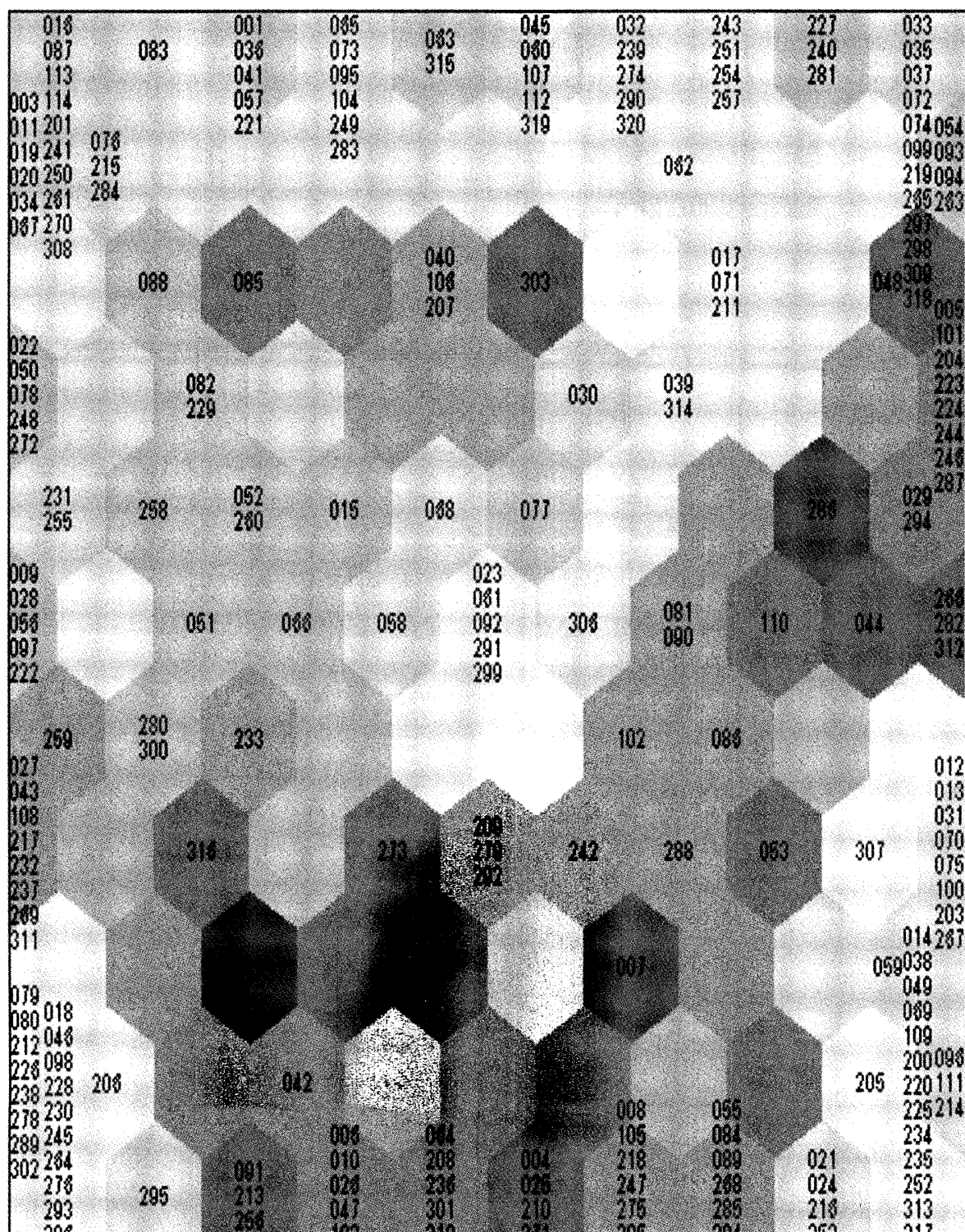


図 2 文書分類マップ (文書番号)

ことである。この合計した2次元ヒストグラムは全体の傾向を表しており、このヒストグラムからキーワードを推定することは可能であるが、正しく求めることはできない。

その理由として、単語分類マップの一つのユニットには複数の単語が含まれていることがあ

り、その場合は、該当する文書に含まれる単語と含まれない単語が混在することがある。よって、該当するユニットに含まれる単語全てをその文書のキーワードとすることはできない。

キーワードを正しく求めるためには、ユニット38に属する5つの文書に現れる単語の出現頻度を

023	どうやって個人情報を守りますか？
<p>「ITでは個人情報は保護できない。重要なのは社員教育だ」——日経システム構築9月号の特集「個人情報を守る」の取材活動を通じて、何度となく聞かされた言葉である。教育の重要性は否定しない。しかし、冒頭のような言葉を発する企業の教育内容を聞いてみると、配属すら決まっていな入社時点にビデオ教材を見て終わりだったり、昇進時の研修の中で外部講師の話を聞いて終わりだったりする。とても教育だけで個人情報を守ることができる、とは言えないと感じた。</p>	
061	ずさんな個人情報の管理が企業経営を揺さぶる
<p>6月26日、大手コンビニエンス・ストア「ローソン」の顧客データが外部に流出する事件が明るみになった。この数年前から、企業や自治体などから情報の漏洩が頻繁に発生している。特に個人情報の漏洩は、企業のイメージ・ダウンだけにとどまらない。訴訟問題に発展すれば大きな経営リスクになる。この5月23日に成立した個人情報保護法は早ければ来年にも施行されるだろうが、そうなれば個人情報のより厳密な取り扱いが求められることになる。企業や自治体の情報セキュリティ対策はどうなっているのだろうか。</p>	
092	個人情報保護法が成立、あなたの会社は大丈夫？
<p>いろいろと物議を醸した「個人情報の保護に関する法律（個人情報保護法）」が5月23日、参院本会議で可決した。5月中にも公布され、基本的な部分については直ちに施行され、残りは公布後2年以内に施行されることになる。</p> <p>この法律は、個人データを保有する企業や団体に対して、個人情報を取り扱う際の義務を定めたもの。条文では、どのような企業・団体が対象になるのか、また保護すべき個人情報とは何を指すのかなど、あいまいな点が多かったが、国会の審議過程で少しずつ全体像が見え始めた。</p>	
291	【結果発表】退会後の個人情報も「一定期間は保持すべき」が大半「個人の要望に応じて削除」に期待
<p>「個人情報は収集した企業ではなく、顧客個人のもの。目的を達した個人情報は廃棄すべき」「個人的には、退会した後も自分の情報が管理されていると思うと気持ちが悪い。即座に廃棄してほしい。しかし、システムを運用する立場としては保管せざるを得ない」</p>	
299	業務上使わなくなった個人情報は捨てるべき？
<p>「あなたの会社から個人情報が漏れたという噂が流れています。さて、あなたはどうしますか」「あなたの所属部門に、架空請求を受けたという顧客から情報漏洩を指摘する連絡がありました。どうしますか」</p>	

表1 ユニット38に属する文書のタイトルと冒頭数行

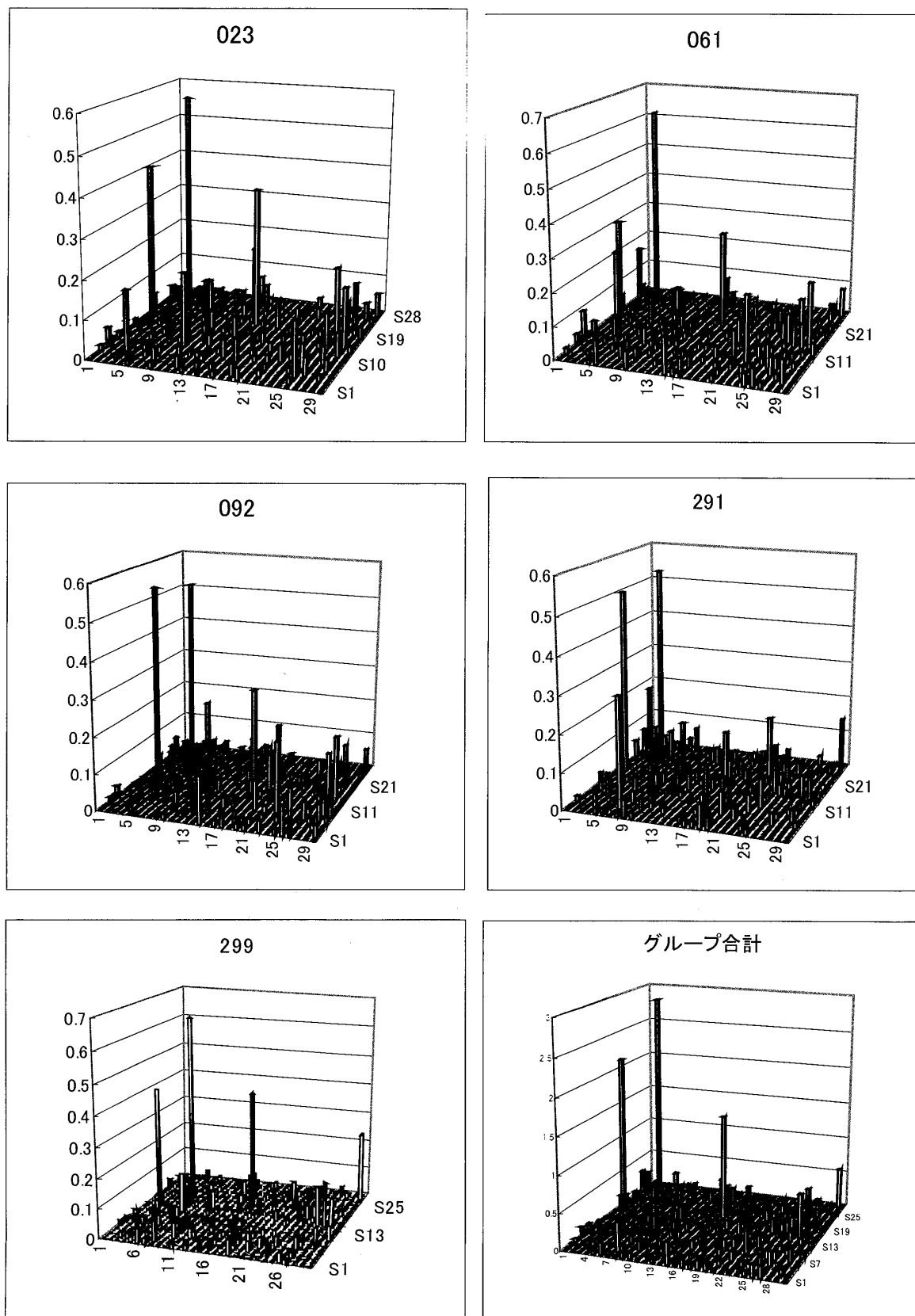


図3 ユニット38に属する文書の2次元ヒストグラム

求める必要がある。つまり、グループ化された文書全体で最も出現頻度の高い単語から順に単語を取り出し、それらをそのユニットに含まれる文書全体のキーワードとするのである。

図4は、文書分類マップの各々のユニットに属する文書全体のキーワードを5単語取り出して図示したものである。先に述べたユニット38にグループ化された文書全体には「情報」、「個人」、

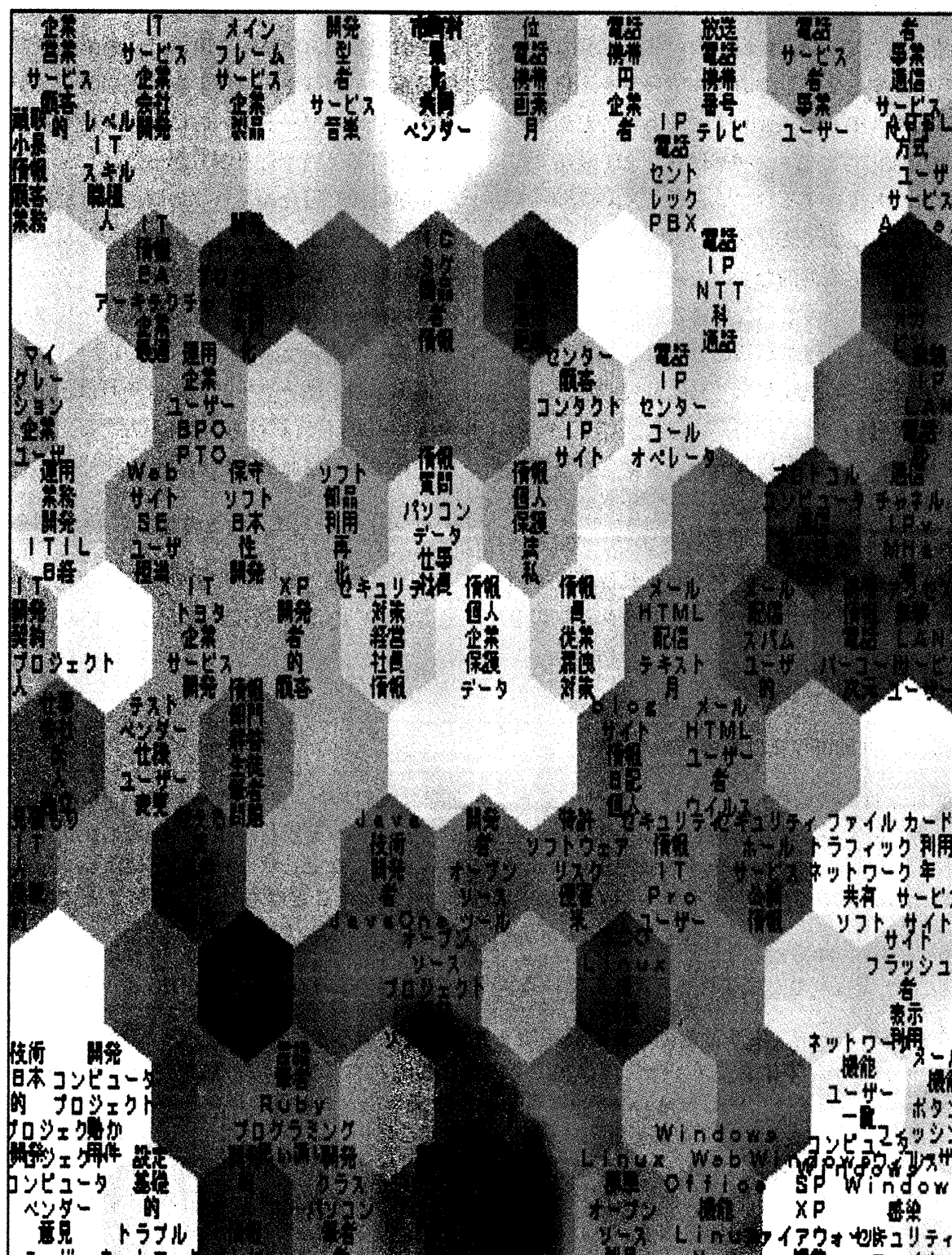


図4 文書分類マップ^o（キーワード）

「企業」、「保護」、「データ」がキーワードとして抽出された。図4から、全体的に見てほぼ関連するキーワードが同じグループに抽出されていることが分かる。

次に、我々が行ったキーワード抽出の妥当性を調べるため他の方法と比較することを試みる。直接比較することは困難なので、我々が求めたユニットに属する文書全体のキーワードを、日本語全文検索システム Namazu⁷⁾ に入力し、どのような文書が抽出されるのかを調べた。Namazu は、文書を分かち書きするソフトとして KAKASI⁸⁾ をもちいている。ただし、KAKASI による分かち書きはあくまで簡易的なものであり、MeCab⁶⁾ や ChaSen (茶筌)⁹⁾ のような形態素解析を行いながら分かち書きを行っているのではない。したがって分かち書きされた結果は、MeCab と異なる。

「情報」、「個人」、「企業」、「保護」、「データ」の5つの単語を Namazu に入力し、検索された文書は、023,092,291,039,025 であった。SOM の場合は、023,061,092,291,299 であったので、023,092,291 が共通で、他の2つの文書が入れ替わっている。新たに現れた、039,025 の文書のタイトルと

冒頭数行を表2に示す。文書分類マップでグループ化された文書はいずれも企業の個人情報保護についてかかれていた。一方 Namazu で検索を行った場合、全く内容の異なる文書 039,025 も検索されてしまい、文書番号 061,299 は検索されなかった。

これにより、SOM を用いて、単語間及び文書間の距離を測りキーワードを抽出したシステムの方がよりキーワードの内容に近い文書を抽出することが可能ということが分かった。

しかし、文書分類マップの周辺部にグループ化されている文書については、類似した文書の集まりが悪い。これらはグループ化できる文書が少ないか内容が多岐にわたっている文書が集まっていると思われる。

抽出されるキーワード自体はそれほど意味が離れているわけではないので、入力する文書の量を増やせば解消されると思われる。

5. 今後の課題

本論文では、前回の文書のグループ化を発展させ、SOM によりグループ化した文書自身から

025	知的財産権は誰のためにあるのか
ありふれた技術を応用してアプリケーションを開発、販売した。その後、突然、ほかの会社から「御社の製品は弊社が所有する特許権を侵害している」という警告状が届いた。誰もが使っている技術なのに……。——「休眠特許」や「サブマリン特許」などと呼ばれる特許によって、ありふれた技術がある日突然自由に使えなくなる。近年、そうしたことはもはや珍しいことではなくなった。	
039	深夜の電話オペレータはどこにいる？ 企業での IP 電話の新しい活用法
深夜のテレビで、実演付きの通信販売番組を見たことがある人は少なからずいるだろう。お腹を引っ込める器具とか、ジュースとか、車の傷を目立たなくする塗料とかいろいろなものを扱っている。中にはテレビを見て、さっそく電話で注文した人もいるに違いない。買う買わないは別として、この深夜の電話を受けるオペレータはどこにいるのか考えたことがあるだろうか？ ヘッドセットを付けたオペレータが大部屋に居並んで、パソコンに向かっていて——こうしたイメージを持った人は正しい。ところが、ブロードバンド回線の普及と IP 電話技術によって、こうしたイメージは変わり始めている、というのが今回のお話である。	

表2 日本語全文検索システム Namazu で検索された文書（抜粋）

キーワードを抽出することを試みた。単語間および文書間の距離を測り、キーワードを抽出することにより、グループ化した文書に対する有用なキーワードを抽出することが出来た。

今後の課題として、抽出したキーワードから文書分類マップに登録された文書を検索するシステムを構築し、また、より精度の高いキーワード抽出のためシステムの調節を行っていくことが今後の課題である。

文 献

- 1) 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.
- 2) 尾崎数也, 薮兼智英, 井上正人, 前原俊信, 岡隆光, 2003, 「自己組織化マップを用いた日本語処理の試み」, 呉大学社会情報学部紀要社会情報学研究 p99-p106.
- 3) 徳永健伸, 1999, 情報検索と言語処理, 東京大学出版会.
- 4) 日経 BP ITPro 記者の目,
<http://itpro.nikkeibp.co.jp/members/backnumber/200309/bnsearch.jsp/>.
- 5) Viscovery SOMine 4.0 Plus,
<http://www.mindware-jp.com/some/>.
- 6) 工藤 拓, MeCab (和布蕪),
<http://chasen.org/~taku/software/mecab/>.
- 7) 馬場 肇, 2004, 改訂 Namazu システムの構築と活用, ソフトバンク.
- 8) 佐藤雅彦, <http://kakasi.namazu.org/>.
- 9) 松本祐治 他, 形態素解析システム茶筌,
<http://chasen.naist.jp/hiki/ChaSen/>.