自然言語処理 126－11
(1998. 7. 24)

構文情報を利用した電子ニュース記事のクラスタリングシステムの作成と評価

清田 陽司　黒橋 禎夫　中村 順一　長尾 真
京都大学大学院情報学研究科 知能情報学専攻
〒606-8501 京都市左京区吉田本町
e-mail: kiyota@pine.kuee.kyoto-u.ac.jp

あらまし
本研究では、電子化されたニュース記事をKohonenの自己組織化マップを利用して自動整理するシステムを作成した。ユーザーは、欲しい記事がありそうなユニット、及びその近傍のユニットを参照することにより、目的に近い記事を得ることができる。このシステムでは、ニュース記事文を構文解析し、構文情報、特に文中の助詞を利用して単語を重みづけすることによって、ユーザの目的により適した自動整理をすることができる。例えば、助詞「は」「も」が附属する単語により高い重みを与えることにより、ニュース記事が企業名を中心に整理され、助詞「を」「の」が附属する単語により高い重みを与えることにより、製品の種類を中心に整理された。
キーワード　自己組織化マップ, クラスタリング, 構文解析, 単語の重みづけ, 電子ニュース記事, 自然言語処理

A Clustering System for Electronic News Articles using Sentense Structure

Yoji Kiyota　Sadao Kurohashi　Jun-ichi Nakamura　Makoto Nagao
Department of Intelligence Science and Technology
Graduate School of Informatics, KYOTO University
Yoshidahonmachi, Sakyo, KYOTO 606-8501 JAPAN
e-mail: kiyota@pine.kuee.kyoto-u.ac.jp

Abstract
We developed a clustering system for electronic news articles using Kohonen's Self-Organization Map. The advantages of this system are: (1) user can get intended articles by referring a unit which seems to contain them, (2) by parsing sentenses and giving larger weights to some terms using postpositional particles of Japanese, this system can arrange articles to fit to user's purpose. For example, when words followed by postpositional particles "wa" "mo" are given larger weights, news articles are arranged by company names.

key words　self organization map, clustering, parsing, term weighting, electronic news article, natural language processing

1 はじめに

現在、電子メールなどを利用したニュース記事の自動配信サービスが普及しつつある。このサービスは、迅速な情報の提供が可能であり、安いコストで大多数の相手に情報を配信することができるなど、数々の特長を持っている。しかし、利用者にとって配信される記事がすべて必要となるわけではない場合が多い。そこで、配信される大量のニュース記事の整理や取捨選択を自動化したいという要求が生じている。

一つの手法として、村井らはT.Kohonenが提案した自己組織化マップを用いたWorld Wide Web(WWW)上の情報の自動整理システムを作成している[中尾 96][村井 97][村井 98]。このシステムをニュース記事の自動整理に応用すれば、新しく届いたニュース記事のうちユーザの興味のあるものだけを抽出したり、ある記事について過去に配信された関連記事を参照するなど、数々の利用方法が考えられる。

本研究では、村井らの作成したシステムを改良し、ニュース記事の自動整理に応用した。それに加えて、構文情報、特にニュース記事の文中に現れる助詞に着目し、それを利用してキーワードを重みづけすることによってシステムの性能の改善を試みた。

2 自己組織化マップによるニュース記事の自動整理

2.1 自己組織化マップを用いた情報整理システム

村井らが作成したシステムは図1に示すように、ユーザからの指示によりWWW空間からHTMLファイルを選択・取得する“Page Selector”、HTMLファイルを使用されている単語を元に多次元ベクトルに数値化する“Word Vector Extractor”、自己組織化マップを利用して多次元ベクトル化したHTMLファイルを二次元マップに整理する“Map Maker”、結果の二次元マップをHTML形式に変換し出力する“Map HTML Generator”から構成されている。

2.1.1 キーワードの抽出とベクトルパターンの作成

Word Vector Extractorは、HTMLファイル中からタグを除去して本文のみを抽出し、日本語形態素解析ツールJUMAN[黒橋 97a]により品詞分類し、普通名詞、固有名詞のみをキーワードとして抜き出す。さらに各HTMLファイルをキーワードを元にtf×idf法によりベクトルパターン化する[有田 95]。

各HTMLのもつベクトルの各要素はそれぞれ1つのキーワードに対応し、そのキーワードの重要度を値としてもつ。HTMLファイルの数をM、利用する異なりキーワードの数をnとする。HTMLファイル$H_i(i=1,\ldots,M)$に対するキーワード$k_j(j=1,\ldots,n)$の重要度$W_j(i)$は、キーワードk_jのH_i内における出現頻度(回数)をF_{ij}、キーワードk_jの全HTMLファイル数Mに対する局所性をM/M_j(M_jはキーワードk_jが含まれているHTMLファイル数)、キーワードk_jの文字長をL_jとして、以下のように計算する。

$$\text{重要度} w_j(i) = F_{ij} \times \log\left(\frac{M}{M_j}\right) \times \log(L_j)$$

HTMLファイルH_iのもつパターンベクトルは、

$$\boldsymbol{x}_i = \{w_1(i),\ldots,w_j(i),\ldots,w_n(i)\}$$

となる。

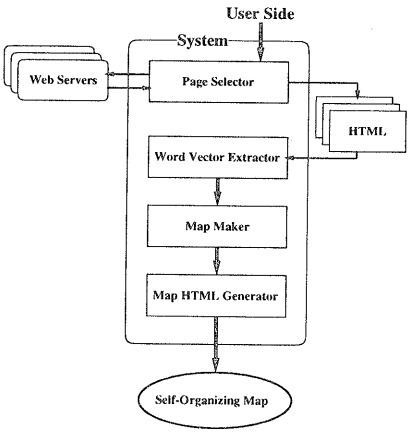


図 1: WWW情報自動整理システムの構成図

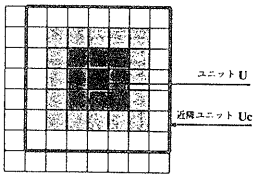


図 2: Kohonenアルゴリズムの概要

2.1.2 自己組織化マップによる情報整理

Map Makerは、Kohonenの提案した**自己組織化マップ学習アルゴリズム** [Kohonen 90](以下**Kohonenアルゴリズム**)によって、多次元ベクトル化されたHTMLファイルを二次元マップ上に整理する。

Kohonenアルゴリズムの概要を図2に示す。二次元マップは図のような正方形であり、$N \times N$個(Nは正の整数)の正方形のユニットをもつ。それぞれのユニット$U_k(k=1,\ldots,N\times N)$は、各HTMLファイル$H_i(i=1,\ldots,M)$を表すパターンベクトル$\boldsymbol{x}_i$(これを入力パターンベクトルと呼ぶ)と同次元のパターンベクトル$\boldsymbol{m}_j(t)$をもつ。

学習は、以下のような過程で行われる。あらかじめ決めておいた学習回数をTとする。

$\boldsymbol{m}$ を初期化する (具体的には、全てを零ベクトルとする)。$t = 0, \ldots, T-1$ に対して以下の処理を反復する。

$i = 1, \ldots, M$ に対して以下の処理を反復する。

1. 入力パターンベクトル $\boldsymbol{x}_i$ について、最も近いパターンベクトルをもつユニット U を決定する (コサイン値の計算により決定)。
2. U と、その近隣ユニット U_c を $\boldsymbol{x}_i$ に近づける。ユニット U とユニット U_j との間のマップ上での距離を r とすれば、

$$\boldsymbol{m}_j(t+1) = \begin{cases} \boldsymbol{m}_j(t) + \alpha(t)\{\boldsymbol{x}_i - \boldsymbol{m}_j(t)\} & (U_j \in U_c) \\ \boldsymbol{m}_j(t) & (U_j \notin U_c) \end{cases}$$

$$\alpha(t) = \alpha_0(t) \cdot \exp(\frac{-r^2}{\sigma(t)^2})$$

$$\alpha_0(t) = 0.5 \times (\frac{1-t}{T})$$

$$\sigma(t) = (\frac{1-t}{T}) \times (\text{マップサイズ}) \times 0.5$$

$\alpha(t)$ は学習率であり、学習が進むにつれて単調減少する。

なお本研究では、ユニット U の近隣ユニット U_c を、U を中心とする 7×7 の正方形内に含まれるユニットとした。

2.1.3 K-mean 法の利用

Kohonen アルゴリズムには、学習に時間がかかり過ぎるという欠点がある。そこで村井らは、**K-mean(K-平均) クラスタリングアルゴリズム** [鳥脇 93](以下 **K-mean 法**) を利用することにより、この欠点を解消した [村井 98]。

K-mean 法は、初期入力によって結果が大きく異なる反面、一般的に少ない反復回数で収束状態になるといった長所をもっている。村井らは、Kohonen アルゴリズムの各ユニットの初期値として K-mean 法のクラスタリング結果を与えることにより、K-mean 法を用いずに 100 回学習させた場合と同程度の性能を、10 回程度の少ない学習回数で得られることを示した。

なお、K-mean 法におけるクラスタの数は、自己組織化マップのユニット数と同一にしている。

2.2 ニュース記事の自動整理への応用

2.2.1 対象とするニュース記事

自動整理の対象とするニュース記事は、株式会社インプレスの電子ニュースサービス「PC Watch」[Impress 97] である。PC Watch は、パソコンに関する最新情報を WWW や電子メールで提供するサービスである。本研究では、PC Watch のニュース記事の取得に「PC Watch メールサービス」を利用した。

個々の記事はあらかじめ何種類かのカテゴリーに分類されている。本研究では、このうち [NEWS] というカテゴリーに属する記事のみを対象とすることとした。その理由は、[NEWS] カテゴリーの記事は、全記事数の約半数を占めており、内容も新製品情報、技術動向、業界動向、企業間紛争など様々であり、情報の自動整理をする効果が大きいと考えられるからである。

本研究では、1997 年 7 月 23 日から 12 月 16 日までに配信された記事のうち、[NEWS] カテゴリーに分類されている 599 記事を対象とした。

2.2.2 抽出するキーワード

本研究では、下記のようなルールでキーワードの抽出を行う。

- JUMAN によって「普通名詞」「固有名詞」「サ変名詞」「地名」「人名」「組織名」「未定義語・カタカナ」と品詞分類された単語は、無条件にキーワードとする。
- JUMAN によって「未定義語・その他」と品詞分類された単語は、次の条件を全て満たす場合に限ってキーワードとする[1]。
  - A ～ Z または a ～ z の文字のみからなる。
    (商品の型番や URL などを除くため)
  - 頭文字が大文字である。
    (固有名詞以外の英単語などを除くため)
  - 3 文字以上である。
    (アルファベットからなる固有名詞は略号も含めて 3 文字以上である場合がほとんど)

2.2.3 システムの構成

PC Watch のニュース記事に対して自己組織化マップによる自動整理を適用するために、以下の 2 つのモジュールを新たに作成した。

- 電子メールの受信箱より PC Watch のニュース記事を選択し、記事単位で抽出する "Article Extractor"
- JUMAN によってニュース記事文を形態素解析し、前項で述べたルールによって記事ごとにキーワードを抽出してキーワードリストを作成する "Keyword Extractor"

第 2.1項で述べた Map Maker、Map HTML Generator は、村井らの作成したものをそのまま利用した。また Word Vector Extractor は、村井らの作成したものからキーワード抽出の機能を取り除いたものを利用した。

2.2.4 評価方法

自己組織化マップによるニュース記事の自動整理の結果を評価するために、企業別、製品の種類別などで分類できる記事グループを任意に決め、それぞれの記事グループに属する

[1] JUMAN Ver. 3.4 では、アルファベットを含む単語は「未定義語・その他」と品詞分類される。

と判断される記事を手作業により全記事中より選び出しておく。全記事に対して自動整理を実行した後、それぞれの記事グループに属する記事が自己組織化マップ上でどのように分布するかを調べる。同じ記事グループに属する記事がまとまって配置されているほど、望ましい結果であるといえる。

本論文では、以下の記事グループを用いる。

- 企業別の記事グループ

  Intel、松下電器産業(「松下」と略記)、IBM、シャープの4企業について、それぞれの企業の新製品やサービスに関する記事を選び出した。
- 製品の種類別の記事グループ

  MPU、液晶ディスプレイ(「液晶」と略記)、モデムの3製品に関する記事をそれぞれ選び出した。
- 特定の事件についての記事グループ

  「Microsoft 対 米司法省」に関する記事を選び出した。この記事グループを「MS vs DOJ」と略記する。
- 一般的な記事グループ

  「企業業績」に関する記事を選び出した。

なお、以上の記事グループのうち、「Intel」と「MPU」は一部が重複している。

まとまりの良さを評価する方法としては、村井らが提案したモーメント値による方法[村井 98]を用いる。

モーメント値は、ある記事グループについて以下のように計算する。まず、自己組織化マップ上でその記事グループの重心点を求める。記事グループに属する記事数をN、各記事$i(i = 1, \ldots, N)$が配置されたユニットの座標[2]を(x_i, y_i)とすると、重心の座標$\mathrm{J}(j_x, j_y)$は式

$$\mathrm{J}(j_x, j_y) = \left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N} y_i}{N} \right)$$

によって求まる。重心点Jより、記事グループ内の各記事iの配置されたユニットの座標(x_i, y_i)と重心点Jの間の距離をd_iとすると、正規化されたモーメント値Mは式

$$M = \frac{1}{N} \sum_{i=1}^{N} d_i$$

によって求まる。モーメント値Mが小さな値を示す記事グループほどまとまっていると判断する。

また、村井らが提案した8連結手法[村井 98]を用いて、自己組織化マップ上に配置された(ある記事グループに属する)記事をさらにグループ分けし(分けられたこれらの記事グループをサブグループと呼ぶ)、最大のサブグループに含まれる記事数を調べる方法も併用する。最大のサブグループに含まれる記事数が多いほど、記事グループがまとまっていると判断する。

[2] マップサイズを$S \times S$とするとき、一番左上のユニットを(1,1)、右下のユニットを(S,S)と座標付けした

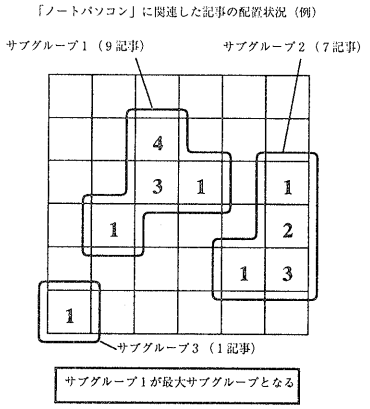


図 3: 8連結手法の適用例

表 1: モーメント値と最大サブグループの記事数・割合

記事グループ		全記事数	モーメント値	最大サブグループ記事数(割合)
企業別	Intel	19	0.459	19(100%)
	松下	20	1.061	18(90%)
	IBM	15	2.167	7(46%)
	シャープ	20	1.679	17(85%)
製品別	MPU	30	1.446	25(83%)
	液晶	13	2.078	10(76%)
	モデム	14	2.279	6(42%)
MS vs DOJ		11	0.826	10(90%)
企業業績		14	0.297	13(92%)

8連結手法とは、ある記事グループに属する記事Aが配置されたユニットから見て上下左右斜めにある8つのユニットに、同じ記事グループに属する記事Bがあれば、記事Aと記事Bは同じサブグループに含まれるとする手法である。8連結手法の適用例を図3に示す。

2.3 実験

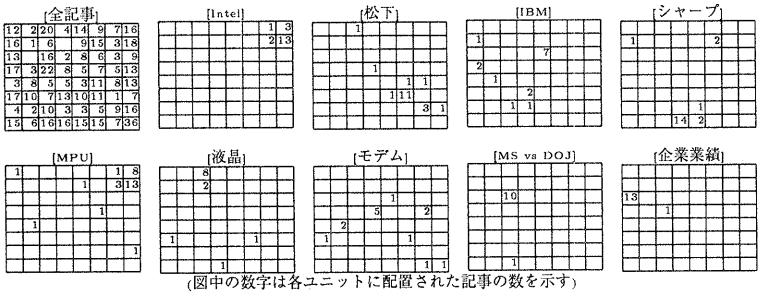
2.3.1 実験方法

前節で述べたシステムにより、実験を行った。

アルゴリズムは、村井らの実験結果を踏まえてK-mean法を適用したのち、そのクラスタリング結果をKohonenアルゴリズムのユニット初期値として与える方法を用いた。Kohonenアルゴリズムの学習回数は20回とした。また、自己組織化マップのサイズは8×8とした。キーワードとして利用する異なり単語数は、5990単語とした。

2.3.2 実験結果と考察

第2.2.4項で述べた記事グループごとの記事の自己組織化マップ上での配置状況を図4に示す。また、記事グループごとのモーメント値と最大サブグループに含まれる記事数・割



(図中の数字は各ユニットに配置された記事の数を示す)

図 4: 記事グループごとの配置状況 (重みづけなしの場合)

合を表1に示す。

なおアルゴリズムの実行時間は、SUN SPARCstation 5(Main RAM 64MBytes) 上で形態素解析にかかる時間を除き30分程度であった。

図4、表1より、各記事グループについて自己組織化マップ上でのまとまり具合を検討した。

- 企業別の記事グループについては、「Intel」、「松下」、「シャープ」についてはほぼまとまって配置されたが、「IBM」についてはやや分散して配置された。
- 製品の種類別の記事グループについては、「MPU」、「液晶」についてはほぼまとまって配置されたが、「モデム」についてはやや分散して配置された。
- その他の記事グループについては、「MS vs DOJ」、「企業業績」の両者ともまとまって配置された。

このように、9つの記事グループのうち7つについてはうまくまとまったが、2つ(「IBM」、「モデム」)についてはあまりうまくまとまらなかった。そこで、この2つの記事グループについて、各記事の (tf×idf 法による) 重要度の高いキーワードがどのようになっているかを調べた。

その結果、記事グループに直接関連しないキーワードの重要度が最も高い記事は、最大サブグループから外れたユニットに配置されている場合が多いことがわかった。例えば、IBM の音声認識ソフトに関する記事は、「音声」「認識」というキーワードの重要度が高いため、音声認識ソフトに関する記事が集まっているユニットに配置されていた。

よって、記事グループごとにうまくまとまらない原因は、記事グループに直接関連しないキーワードの重要度が高い記事があるためだと推測される。

3 構文情報の利用

3.1 構文情報を利用したキーワードの重みづけ

3.1.1 キーワードの重みづけの必要性

第2.3節で行った実験では、村井らの研究と同様にニュース記事の文中に出現するキーワードを全て同列に扱った。しかし全てのキーワードを同列に扱うことには問題点がある。例えば、PC Watch のニュース記事の自動整理を利用するユーザには次のような要求があると思われる。

1. 企業名ごとに分類したい。
   例: ソニーの新製品に関する情報が知りたい。
2. 製品の種類ごとに分類したい。
   例: ノートパソコンの新製品に関する情報が知りたい。

しかし、全てのキーワードを同列に扱うだけでは、両方の要求を満たすことは以下に述べる理由から困難である。

多くの企業は複数の種類の製品を手がけており、一つの種類の製品は多数の企業が発売している。一方、PC Watch のニュース記事の文中に現れるキーワードのうち、企業名を表すキーワードと、製品 (あるいはそれに関連した語) を表すキーワードはそれぞれかなりの部分を占める。このため、1の要求については製品を表すキーワードが存在するために自己組織化マップ上に企業名ごとにうまくまとまらないことが考えられる。逆に、2の要求については企業名を表すキーワードが存在するために製品の種類ごとにうまくまとまらないことが考えられる。

この問題を解決するために、**キーワードの重みづけ**を導入する。これは、ニュース記事文中に出現するキーワードのうちの一部に、他のキーワードより高い重みを与えるというものである。例えば、企業名を表すキーワードに他のキーワードよりも高い重みを与えることができれば、ニュース記事は自己組織化マップ上で企業名ごとに分類されることが期待できる。

3.1.2 構文情報の利用

前項で述べたキーワードの重みづけを実現するには、どのようなキーワードに重みを与えるかを決定する必要がある。ここでは、キーワードに附属する助詞を利用して、キーワードに重みを与える手法を提案する。

助詞を利用したキーワードの重みづけ PC Watchのニュース記事で、新製品関連のニュースは次のような書き出しで始まることが多い。

> 米Apple社は22日、Mac OSの新バージョン、Mac OS 8を発表した。

この文中、企業名を表すキーワードは「米Apple社」、製品を表すキーワードは「Mac OS 8」である。

ここで、それぞれのキーワードに附属している助詞に注目する。「米Apple社」には「は」、「Mac OS 8」には「を」という助詞が附属している。そこで、例えば助詞「は」が附属するキーワードに他のキーワードより高い重みを与えてから自己組織化マップによる自動整理を行うと、企業名を中心に分類されるのではないか予想できる。また、助詞「を」が附属するキーワードに高い重みを与えれば、商品名を中心に分類されるのではないかと予想できる。

構文解析の必要性 助詞によってキーワードの重みづけをするには、形態素解析のみでは不十分な場合がある。以下の例文について考える。

> 株式会社メルコは、Macintosh用UltraSCSI対応HDD、PC-9821用内蔵型HDD、NECEXPRESS5800サーバ用増設RAMボードなどを発表、9月上旬から順次出荷を開始する。

下線で示した部分が商品名を表す。ここで、助詞「を」が附属するキーワードに重みをつける場合は、文中で並列関係にある全ての下線部にキーワードに同じ重みを与えなければならない。しかし、形態素解析のみではこのような処理は難しく、構文解析によって並列関係を解析する必要がある。

そこで本研究では、構文解析システムKNP[黒橋 97b]を利用した。KNPは文中の並列関係などを解析することができる。

3.1.3 助詞による重みづけ

本研究では、以下に挙げる助詞が附属するキーワードについて重みづけをして、それが自己組織化マップによる自動整理にどのような影響を与えるかを調べた。

1. 格助詞「が」
2. 副助詞「は・も」
3. 格助詞「を」
4. 格助詞「に」
5. 格助詞「で」
6. 接続助詞「の」

このうち、1～5の助詞によるキーワードの重みづけの対象となるキーワードをそれぞれキーワード「が」、キーワード「は・も」、キーワード「を」、キーワード「に」、キーワード「で」と略記する。また、「AのB」の形で文中に出現している場合、Aをキーワード「の」(A)、Bをキーワード「の」(B)と略記する。

表2: 各助詞による重みづけの対象となる単語の数・割合

	「の」(B)	「の」(B)以外	合計
「が」	731 (1.2%)	2203 (3.5%)	2934 (4.7%)
「は」	917 (1.5%)	4787 (7.6%)	5704 (9.1%)
「も」	312 (0.5%)	658 (1.1%)	970 (1.5%)
「を」	1799 (2.9%)	4095 (6.5%)	5894 (9.4%)
「に」	920 (1.5%)	2768 (4.4%)	3688 (5.9%)
「で」	389 (0.6%)	1649 (2.6%)	2038 (3.2%)
「の」(A)	799 (1.3%)	7019 (11.2%)	7818 (12.4%)
その他	1840 (2.9%)	32056 (50.9%)	33896 (53.9%)
合計	7707 (12.2%)	55235 (87.8%)	62942 (100.0%)

なお、キーワード「の」(B)が、「の」以外の助詞による重みづけの対象でもある場合は、より高い重みづけを優先して適用することとした。

本研究で利用したPC Watchの599記事に使われている延べ62942単語のうち、助詞による重みづけの対象となるものの数と割合を表2に示す。

3.2 個々の助詞についての重みづけ

3.2.1 実験方法

個々の助詞についてキーワードの重みづけを変化させる実験を行った。

キーワード「が」、「は・も」、「を」、「に」、「で」、「の」(A)、「の」(B)について、それぞれ他のキーワードのX倍の重みを与え、Xの値を変化させて実験した。用いたXの値は2、3、5、10である。他の条件は、第2.3節の実験と同様とした。

3.2.2 実験結果と考察

モーメント値の変化のグラフを、キーワード「は・も」に重みを与えた場合の企業別の記事グループについて図5に、キーワード「を」「に」「の」(A)「の」(B)に重みを与えた場合の製品の種類別の記事グループについて図6～9に示す。

それぞれの記事グループに属する記事が、キーワードの重みづけを変化させることによってどのようにマップ上に配置されたかを検討した。

企業別の記事グループ いずれの企業に関する記事も、キーワード「は・も」に5倍以下の重みを与えた場合は、重みづけなしの場合よりもまとまって配置される傾向がみられた(図5)。

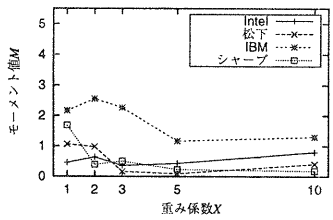


図 5: モーメント値の変化 (企業別) キーワード「は・も」

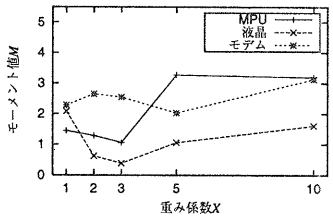


図 6: モーメント値の変化 (製品別) キーワード「を」

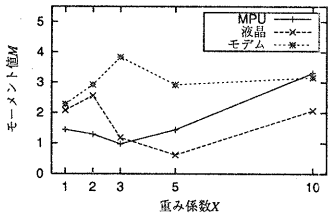


図 7: モーメント値の変化 (製品別) キーワード「に」

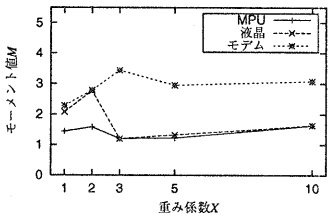


図 8: モーメント値の変化 (製品別) キーワード「の」 (A)

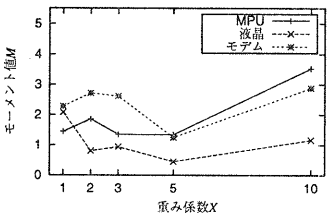


図 9: モーメント値の変化 (製品別) キーワード「の」 (B)

製品の種類別の記事グループ

MPU に関する記事 キーワード「を」「に」に 3 倍以下の重みを与えた場合と、「の」(B) に 5 倍以下の重みを与えた場合は、重みづけなしの場合よりもまとまって配置される傾向がみられた。

液晶ディスプレイに関する記事 キーワード「を」に 3 倍以下の重みを与えた場合と、「に」「の」(A)「の」(B) に 5 倍以下の重みを与えた場合は、重みづけなしの場合よりもまとまって配置される傾向がみられた。

モデムに関する記事 どの場合でも、「MPU」、「液晶」に関する記事ほどまとまらなかった。ただ、キーワード「を」、「の」(B) に 5 倍以下重みを与えた場合は、重みづけなしの場合よりもまとまって配置される傾向がみられた。

その他の記事グループ 「Microsoft 対 米司法省」「企業業績」に関する記事は、大抵の場合重みづけなしの場合と同様にまとまって配置された。キーワード「が」「は・も」「に」「の」(A)「の」(B) に重みを与えた場合は、かえって分散して配置される傾向がみられた。

以上の実験結果から、次のことがいえる。

1. 記事を企業別にまとめたいときは、キーワード「は」「も」に重みを与えると明確な効果があった。
2. 記事を商品の種類別にまとめたいときは、キーワード「を」「に」「の」(A)「の」(B) に重みを与えるとある程度の効果がみられたが、1 ほど明確ではない。
3. 重みづけなしの場合でもうまくまとまっていた「MPU」、「MS vs DOJ」、「企業業績」に関する記事は、重みを上げ過ぎるとかえって分散した。

3.3 複数の助詞についての重みづけ

3.3.1 複数の助詞についての重みづけ

前節では、個々の助詞についてキーワードを重みづけし、それが自己組織化マップによる情報の自動整理に与える効果を調べた。

企業別の記事グループについては、キーワード「は・も」に重みを与えるだけでかなり結果が改善された。しかし、製品の種類別の記事グループについてはそれほど効果は明確ではない。そこで、複数の助詞についてキーワードを重みづけすることによって改善を試みた。キーワード「を」「に」「の」(A)「の」(B) のうち、2 つ以上を組み合わせて実験した。

3.3.2 実験方法

複数の助詞についてキーワードの重みづけを変化させる実験を行った。

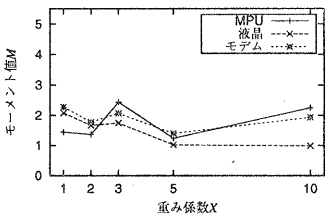


図 10: モーメント値の変化 (製品別) ケース (f) (キーワード「を」「の」(A)「の」(B))

以下の組み合わせのキーワードについて、他のキーワードの X 倍の重みを与え、X の値を変化させて実験した。用いた X の値は 2、3、5、10 である。

ケース (a) キーワード「を」「に」
ケース (b) キーワード「を」「の」(A)
ケース (c) キーワード「を」「の」(B)
ケース (d) キーワード「を」「に」「の」(A)
ケース (e) キーワード「を」「に」「の」(B)
ケース (f) キーワード「を」「の」(A)「の」(B)
ケース (g) キーワード「を」「に」「の」(A)「の」(B)

他の条件は、第 2.3節の実験と同様とした。

3.3.3 実験結果と考察

モーメント値の変化のグラフを、ケース (f) の製品の種類別の記事グループについて図 10に示す。

製品の種類別の記事グループに含まれる記事が、キーワードの重みづけを変化させることによってどのようにマップ上に配置されたかを検討した。

MPU に関する記事 ケース (b)、(d)、(g) と、ケース (f) で $X = 5$ 以下の場合では、重みを上げるにつれてまとまる傾向がみられた。

液晶ディスプレイに関する記事 ケース (f) と、ケース (a) で $X = 5$ 以下、ケース (b)、(e) で $X = 3$ 以下、ケース (c)、(d) で $X = 2$ 以下の場合では、重みを上げるにつれてまとまる傾向がみられた。

モデムに関する記事 ケース (b)、(f)、(g) で $X = 5$ 以下、ケース (d)、(e) で $X = 3$ 以下の場合では、重みを上げるにつれてまとまる傾向がみられた。

個々の助詞について重みづけした場合と比較すると、ケース (b)、(d)、(e)、(f)、(g) については全般的に結果が改善された。

最適な重み係数についてはケースによって若干異なるが、ケース (b)、(d)、(e) では $X = 3$、ケース (f)、(g) では $X = 5$ の場合で全般的に最も結果が改善された。特に、ケース (f) で $X = 5$ の場合には、個々の助詞について重みづけした場合にあまりまとまらなかった「モデム」も含めた 3 つの記事グループが全てうまくまとまった。

以上の結果から、製品の種類別に記事を分類したいときは、キーワード「を」「の」(A)「の」(B) を組み合わせて 5 倍程度の重みを与えると効果があるといえる。

4 結論

本研究では、村井らの作成した WWW 情報自動整理システムを PC Watch のニュース記事の自動整理に応用し、その有効性を確かめた。さらに、構文情報、特に文中の助詞を用いてキーワードを重みづけすることによって、ユーザの利用目的に応じた自動整理を行うことができることを示した。

今回の実験の範囲では、以下のようにすれば適切な整理ができることが明らかになった。

1. 企業に注目して分類する場合は、キーワード「は・も」に 5 倍の重みを与える。
2. 製品に注目して分類する場合は、キーワード「を」「の」(A)「の」(B) に 5 倍の重みを与える。

謝辞

本研究を進めるにあたり適切な御指示をいただきました村井幸一氏に心から感謝いたします。

参考文献

[中尾 96] 中村順一, 中尾学: 自己組織化マップを利用した Web 情報整理システムの作成と評価, 言語処理学会第 2 回年次大会 (1996)

[村井 97] 中村順一, 甲斐郷子, 村井幸一: 自己組織化マップによる WWW 日本語情報検索システムの作成と評価, 言語処理学会第 3 回年次大会 (1997)

[村井 98] 中村順一, 村井幸一, 馬場博巳, 甲斐郷子: クラスタリングアルゴリズムを利用した WWW 情報整理システムの作成と評価, 言語処理学会第 4 回年次大会 (1998)

[Kohonen 90] T.Kohonen: The Self-Organizing Map, Proceedings of the IEEE, Vol. 78, No. 9 (1990)

[Impress 97] 株式会社インプレス: PC Watch http://www.watch.impress.co.jp/pc/ (1997)

[黒橋 97a] 黒橋禎夫, 長尾真: 日本語形態素解析システム JUMAN 使用説明書 ver. 3.4, 京都大学大学院工学研究科 (1997)

[黒橋 97b] 黒橋禎夫: 日本語構文解析システム KNP 使用説明書 ver. 2.0b3, 京都大学工学部 (1997)

[有田 95] 有田英一, 安井照晶, 津高新一郎: 単語集合の自動構造化機能を持つ「情報散策」方式, 電子情報通信学会「言語理解とコミュニケーション」(1995)

[鳥脇 93] 鳥脇純一郎: パターン認識とその応用, テレビジョン学会教科書シリーズ (1993)