社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2004－ＩＣＳ－135 (32)
2004／3／17

類義語辞書とドキュメントの特性を用いた類似度の獲得

小川　智也† 　犬塚　信博†† 　加藤　昇平† 　世木　博久†

†††名古屋工業大学　〒466-8555 名古屋市昭和区御器所町
E-mail: †{mei,shohey,seki}@ics.nitech.ac.jp, ††inuzuka@nitech.ac.jp

あらまし　ドキュメントを検索するためにドキュメント間類似度を与える方法を考える。特に専門分野ドキュメント集合に対しても有効に働く類似度を求めたい。従来、ドキュメントを語句の出現数などで特徴づけ、それにより類似度を定義する手法等がある。統計的に次元を圧縮する、類義語辞書のカテゴリで語句をまとめる、またそれらを併用する手法がある。本研究では専門分野のドキュメントに応じて類義語辞書のカテゴリを動的に併合する手法を与え、ドキュメントに応じた次元圧縮をし、類似度を求める手法を提案する。

キーワード　テキストマイニング、類似度、類義語辞書

Similarity of Documents Using Thesaurus and Statistical Characteristics

Tomoya OGAWA†, Nobuhiro INUZUKA††, Shohei KATO†, and Hirohisa SEKI†

†††Nagoya Institute of Technology　Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
E-mail: †{mei,shohey,seki}@ics.nitech.ac.jp, ††inuzuka@nitech.ac.jp

Abstract　We study a method to give similarity between documents, the similarity which can be used for search. This paper aims to give a method for documents in specific fields. Some conventional methods treat a document as a vectors of frequency of words in the document and a similarity is given in the vector space. In this case the large dimension is a problem. Some solution includes counting words in a semantic category in a dimension, a statistical method, and combining these. This paper proposes a method that merges semantic categories depending on the given collection of documents. This is expected to give an appropriate similarity for documents in specific fields with reasonable dimension.

Key words　text mining, similarity, thesaurus

1. まえがき

近年、電子化されたドキュメントは、PCやインターネットの普及と共に急増している。それぞれのドキュメントの内容についても、新聞社のウェブサイトのような一般的なもの、電子化された論文のように専門的なものや、個人の趣味を記述しているようなホームページまで多岐に渡っている。

幅広い分野にわたる大量なドキュメント中から、自分が欲しい情報を引き出すことは非常に煩雑な作業を要する。検索サイトにアクセスし、語句検索を行ない、その語句を含むウェブページを表示させる。そして、ユーザは自分の欲しい情報を含むドキュメントがどれかを見比べる必要がある。語句検索だけでは欲しい情報を探すには不十分であり、どのような条件下でも使えるわけではないと考えられる。

この解決法はドキュメントからドキュメントを探すことである。単純に検索語句をいくつか並べるよりも、ドキュメントの特性を用いたドキュメント間の関連を利用した方が有用なことがあるだろう。ドキュメントを手がかりにすることで、数個の検索語句を検索エンジンに与えるよりも有用な情報を用いることができる。

ドキュメント間類似度の求める有用なものの1つに類義語辞書を用いる手法が有る。意味が類似している語句をカテゴリに分類し、ドキュメントをカテゴリ空間に写像することで、カテゴリ空間内で類似度を測る。広く知られている類義語辞書として、日本語では日本語語彙大系[2]、英語では WordNet[3] が有る。

本稿では専門分野のドキュメントに対する類似度を考える。語句や統計情報を用いた手法を適用することはできるが、類義語辞書が対応できない問題を挙げ、専門分野ドキュメントに対応できる類義語辞書の用い方を提案する。そのため、類義語辞書を用いる手法、統計的手法、またそれらを組み合わせた手法の長所・短所を検討し、専門分野のドキュメントに対応できない原因を調べ、その解決法を提案する。まず節2.で、これら従来手法の説明をし、節3.で性質を述べる。節4.で従来手法の

欠点を解消する手法を提案する。節 5. では提案手法の有用性を実験によって示す。

2. ドキュメント間類似度

この章では、ドキュメント間類似度を求める既存の手法について述べる。n 個のドキュメントを含む $D=\{d_1,d_2,\cdots,d_n\}$ と語句集合 $T=\{t_1,t_2,\cdots,t_m\}$ を関係づける。ドキュメント d_i が、語句 t_j に関連する度合を v_{ij} とする。すると、ドキュメント d_i は $d_i=(v_{i1},v_{i2},\cdots,v_{ij},\cdots,v_{im})$ のように表される。このときドキュメントの集合 D は行列

$$D=\begin{pmatrix} d_1 \\ \vdots \\ d_i \\ \vdots \\ d_n \end{pmatrix}=\begin{pmatrix} v_{11} & \cdots & v_{1j} & \cdots & v_{1m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{i1} & \cdots & v_{ij} & \cdots & v_{im} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ v_{n1} & \cdots & v_{nj} & \cdots & v_{nm} \end{pmatrix} \quad (1)$$

で表すことができる。値 v_{ij} の与え方については、多くの手法が存在する [1] [5]。本稿での v_{ij} はドキュメント d_i 中で語句 t_j の 出現回数とする。

各ドキュメント d_i を $\mathbf{R}^m$ の要素と見なし、その要素間の類似度を考える。余弦による類似度は一つの候補であるが、語句数 m が大きくなるにつれて問題となる。m が大きい場合は、適切な $T_{sub}\subseteqq T$ を生成することや、T を分割してその代表元で空間を構成するなどし、語句数に比べてできるだけ小さい次元の空間内で距離を求めることが望ましい。

次元を下げる方法には (1) 重要語句を選択 (2) 語句のカテゴリ化 (3) 特異値分解による次元の削減等が挙げられる。以下にこれらの手法について説明する。

2.1 重要語句を選択する手法

一般的に広く知られている、語句の部分集合 T_{sub} を求める手法には TF - IDF [8] 重み付けを用いる手法がある。tf はドキュメント $d\in D$ における語句 t の生起頻度 $tf(d,t)=(d$ での語句 t の出現回数$)/(d$ での各語句出現数の総和$)$ である。idf (Inverse Document Frequency) は全ドキュメントの数 $n=|D|$ と、語句 t が 1 回以上生起するドキュメントの数 $df(t)$ から $idf(t)=\log\frac{n}{df(t)}$ のように定義される。idf 値が高い語句は、多くの文書に生起せず、文書を特定する性質を持つ可能性が高い。$tf\cdot idf$ は $tf(d,t),idf(t)$ を掛け合わせた $w(t,d)=tf(d,t)\cdot idf(t)$ であり、ドキュメントでの語句の重みを表す。

2.2 類義語辞書を用いたカテゴリ化による手法

カテゴリ $C=\{c_1,\cdots,c_k\}$ は T の部分集合の族である。各 $c_i(1\leqq i\leqq k)$ は $c_i\subseteqq T$ であり、$\bigcup_i c_i=T$ という性質を持つ。カテゴリ化によって $k<<m$ となる空間へ写像することで、適切な次元を得られる可能性がある。次の K_T を用いて、ドキュメントと語句の関係を表す行列をドキュメントとカテゴリとの関係を示す行列に変換する。K_T の j 版目の行 $\mathbf{k}_j$ は以下のようになる。

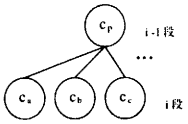

図 1 カテゴリの上下段関係

$$\mathbf{k}_j=\left(\frac{c(j,1)}{C(j)},\cdots,\frac{c(j,i)}{C(j)},\cdots,\frac{c(j,k)}{C(j)}\right)$$

$$c(j,i)=\begin{cases} 1 & if \quad t_j\in c_i \\ 0 & if \quad t_j\notin c_i \end{cases}$$

$$C(j)=\sum_{i=1}^{k}c(j,i)$$

複数のカテゴリに含まれる多義語は、含まれるカテゴリ数によって割っている。カテゴリを用いる手法は、この $K_T=(\mathbf{k}_1,\mathbf{k}_2,\cdots,\mathbf{k}_m)^T$ を用いて、語句空間行列として表されているドキュメント集合 D からカテゴリ空間行列とした D_C を

$$D_C=K_TD \quad (2)$$

と、求めることができる。

カテゴリ化を行なうには類義語辞書を用いる手法が挙げられる。類義語辞書はカテゴリに木構造を持たせたものである。カテゴリがノードであり、カテゴリ同士のリンクは枝となっている。上位のルートにより近いノードは、より一般的な意味を示すカテゴリであり、逆に葉に近いノードはより狭い意味を持つカテゴリである。

類義語辞書を用いて語句をカテゴリ化しても、D によってはカテゴリ空間 $\mathbf{R}^k$ ですら相関の取りづらい疎な空間になることがある。この時、全てのカテゴリを用いず、上段から一定以内の深さまでのカテゴリを用いる。

2.3 特異値分解による空間直交化手法

式 (1) で示した行列は目的変数の無い多変量データと見なすことができる。相関に基づき、空間の各基底ベクトルが直交化するように線形変換する多変量解析には、様々な手法が存在する [10] 。例えば、参考文献 [4] で挙げられている特異値分解 [9] を用いて、類似度を求めるのに適した空間を生成する手法がある。

任意の行列 G について下記のような特異値分解が常に可能であると知られている。

$$G=U\Sigma V^T \quad (3)$$

G の階数を r とした場合、U は n 行 r 列、V^T は r 行 m 列の直交化行列となる。また、Σ は r 行 r 列の対角行列で、その対角成分とは G^TG の固有値の絶対値の平方根を値とした特異値 σ ($\sigma_1\geqq\sigma_2\geqq\cdots\geqq\sigma_r$) である。U 中の上位 k 列を用いた行列 U_k、対角行列 Σ 中で k 番目の特異値まで含む k 行 k 列の行列、V^T 中の上位 k 行を用いた行列 ${V_k}^T$ の積で表される行列 G_k は、階数が k 以下の任意の行列中で、最も良い G の近似であると知られている。

Σ_k は、各属性に対する重み付けを表す対角行列であるから、

U_k 自体は重み付けがされていない基底ベクトルを要素とする列ベクトルであると見なすことができ、かつ直交は保たれている (${U_k}^T R_K = I$)。そこで、この U_k を行列 G を数学的に直交化させた行列 G_M とする。

$$U_k = D_k V_k {\Sigma_k}^{-1} \quad (4)$$

$$= G_M \quad (5)$$

式 (1) で示す行列 D に特異値分解を行なう場合を考える。本節で示した特異値分解による手法で用いた行列 G に D を代入することで、n 行 k 列の直交化された行列 D_M を得ることができ、空間 $\mathbf{R}^m$ は直交化された空間 $\mathbf{R}^k$ となる。

2.4 類義語辞書と特異値分解の併用法

類義語辞書を用いる手法と特異値分解による直交化手法、それぞれに欠点がある。類義語をカテゴリ化することで異なる語句の相関を取れるようになるものの、カテゴリ間は数学的に直交でない。また、特異値分解では空間は直交化されているものの、類義語の考慮はされていない。

そこで、類義語辞書を用いる手法と数学的直交化手法を併用した手法 [4] がある。これは、類義語辞書を用いて語句をカテゴリ化した後、基底となるカテゴリ間の関連性を特異値分解により考慮し、その後数学的に直交化された空間を生成する手法である。即ち、次のように K_T によりカテゴリ化して D_C を得て、さらに特異値分解により k' 次元の D_{CM} を得る。

$$D_C = DK_T$$

$$D_{CM} = D_{Ck'} V_{k'} \Sigma_{k'}^{-1}$$

3. 既存手法の性質と問題点

ここでは、前節で述べた既存手法の性質と問題点について述べる。

3.1 既存手法の性質

(1) 類義語辞書を用いる手法の利点は以下の 2 つである。1 つ目は、語句をカテゴリに分類して考えるため、人が考える意味での類似を取り入れて、ドキュメント間の相関を取ることができる。2 つ目に、各軸に対して 1 つのカテゴリを対応させられるため、1 つの次元が 1 つの概念に対応していると考えることもできる。欠点として、カテゴリ間が数学的に直交していない。つまり、意味の面で直交しているからといって、語句をカテゴリ化しただけのカテゴリ空間において、余弦等の尺度を用いてドキュメント間の類似度を求めることや、距離を測ることが適していない可能性がある。

(2) 特異値分解を用いた手法の利点は、ドキュメントを表す語句空間を数学的な視点から直交化しているため、この空間内で類似度や距離を求めることは適切と言える。欠点として、対象とするドキュメント集合の規模や、各ドキュメントの内容量によっては相関が得にくい場合がある。また、語句間の関連を用いるために起こる同義語の異なる表記の問題は数学的直交化のみでは解決できない。

3.1.1 類義語辞書と特異値分解併用法

類義語辞書と特異値分解の併用法では、カテゴリ化によって類義語を考慮することで、特異値分解によりカテゴリ空間を直交化することで、それぞれの欠点を、それぞれの利点で補うことができている。類義語辞書を用いて語句をカテゴリに分類した後に特異値分解を用いることで、類義語の考慮もしながら、数学的に直交化できている。しかし、専門分野ドキュメントに適用する上では次節のような問題点が現れる。

3.2 専門分野ドキュメントに適用する場合の問題点

ここでは、専門分野ドキュメントに類似度を適用する場合に従来法で解決されていない、解消すべき問題点を示す。

まず、類義語辞書そのものの問題点を示す。類義語辞書が持っているカテゴリは語句の一般的な意味の視点で構成されており、一般的なドキュメント集合に対して適用する範囲でのみ適切である。しかし、内容が専門的なドキュメント集合に対しては、問題がないとは言えない。一般的に類義語とされる語句でも、専門的な視点からは、全く異なる意味を示す語句であることは多い。専門的なドキュメント集合に適用する上では、単純に類義語辞書を用いる手法を使う限りは、この問題が解消されることはない。

次に、類義語辞書を用いる手法で行なっている処理の問題点を示す。そのカテゴリは意味の広さに応じた木構造となっている。類義語辞書を用いる手法では、用いる段を指定する必要がある。対象とするドキュメント集合に含まれるドキュメントや語句が少ない場合、カテゴリ化するのみで相関を得られるほど行列が密にならない。それゆえに、上位から任意の段数までで下位のカテゴリを代表させることで、カテゴリ間の関連を見つけ易くすることができる。しかし、分野による語句の意味を反映していないため、ドキュメント集合の特性を損なう可能性が出てくる。

4. 提案手法

4.1 ドキュメント集合に応じたカテゴリ併合

ここでは、カテゴリを用いる新しい手法として、対象とするドキュメント集合に応じて動的にカテゴリ併合を行なう、動的カテゴリ併合法について述べる。この手法は、3.2 節で述べた問題点を解消できる。

多くの分野に均等に語句が散在しているような国語辞典や新聞記事等を対象とするなら、ある一定の段まで引き上げる場合に全てこのように処理すれば良いが、狭い分野のドキュメント集合に対してこの処理を行なうことは適切でない。そこでカテゴリを併合する・しないを、カテゴリ単位で決める手法を提案する。カテゴリ間の相関関数 $corr$ を以下のように定め、カテゴリ c_i に対してそのカテゴリの語句を 0 回以上含むドキュメント部分集合 D_{c_i} について、

$$D_{c_i}(c_j) = D_{c_i} \text{でカテゴリ } c_j \text{の語句が現れる回数}$$

とし、カテゴリ間の相関を、このドキュメント部分集合間集合間の相関として、以下のように定める。

$$corr(c_i, c_j) = \frac{|\{c_x | D_{c_i}(c_x) \geqq 1 \wedge D_{c_j}(c_x) \geqq 1\}|}{|\{c_x | D_{c_i}(c_x) \geqq 1 \vee D_{c_j}(c_x) \geqq 1\}|}$$

$$(x = 1, \cdots, k)$$

$corr$ が低いカテゴリ同士は、共起することが少ないので例え親カテゴリが同一であっても、その兄弟カテゴリに含まれるの語句と使われ方が異なるから併合しない。逆に $corr$ が高いカテゴリ同士は使われ方が同じなので、兄弟カテゴリと併合することが好ましいと考えられる。

この相関関数の値に応じて併合を行なうことで不必要な次元の収縮を行なうことができる。類義語辞書における一般的な視点からは同一のものと考えられて併合されるような場合でも、専門分野に特化されたドキュメント集合中では全く違う意味で用いられている場合があるとすると、共起する語句の相関は低いであろうと予測できる。相関関数 $corr$ は、ドキュメント集合における語句の特性を効果的に用いられるものと考えられる。

この相関関数を用いたカテゴリ併合法を述べる。まず、類義語辞書で、同一段に存在するカテゴリに対しての処理は、その上位カテゴリが同一である場合、カテゴリ c_i, c_j が予め定める閾値 $bound(0 \leqq bound \leqq 1)$ を超えた時に、カテゴリ c_i に c_j を併合する。

階層が上下段に別れたカテゴリの併合は同一の段において併合されるカテゴリが無くなった場合に行なう。あるカテゴリ c_i の親カテゴリが c_p である時、$corr(c_i, c_p)$ が閾値を超えた時、親カテゴリに 子カテゴリを併合する。

このアルゴリズムは表 4.1 に示す通りとなる。例えば、図 1 のような関係を持つカテゴリに対し、提案法を用いてカテゴリ併合を用いた場合、閾値に応じて図 2 のうちのいづれかのように併合される可能性がある。

この手法は、カテゴリ間相関の閾値を小さくすると、殆んどのカテゴリが併合されてしまうことが容易に想像できる。閾値を 0 とすると、全てのカテゴリが無条件で併合されてしまい、1 カテゴリしか残らない。逆に、閾値を大きくする程、併合されるカテゴリは少なくなり、閾値を 1 としたときは全く併合されない。このカテゴリの動的併合法を用いた類似度計算手法は、表 2 のようになる。

5. 実　　験

5.1 実験方法

従来法においては、類義語辞書を用いた手法の後、得られたカテゴリ行列に直接、特異値分解を適用している。ここでは、提案手法をカテゴリ行列に適用した後に特異値分解を行った場合との比較を行なう。

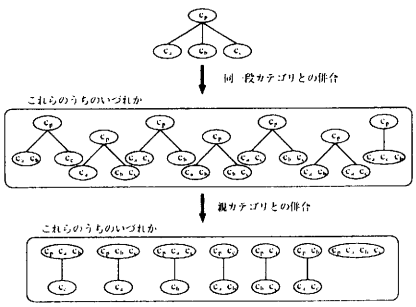

図 2　提案法により起き得るカテゴリ併合

動的カテゴリ併合法

入力：ドキュメントとカテゴリの関係行列 D, ドキュメント集合中のカテゴリ集合 C, 閾値 $bound$ 類義語辞書の最深層 l_{max}, カテゴリ数 n, ドキュメント数 m, カテゴリ間相関の閾値 $bound$

出力：更新された D

for $l := l_{max}$ **downto** 1 **do**
  repeat
    for all $c_i, c_j \in C(1 \leqq i, j \leqq |C|, i \neq j)$ **do**
      if c_i と c_j は兄弟 $\wedge$ c_i と c_j は l 段目 $\wedge$ c_i と c_j は併合条件を満たす
      cen c_i, c_j を併合手法を用いて併合
    end for
  until 全ての c_i, c_j が併合条件を満たさなくなるまで
  repeat
    for all $c_i, c_j \in C(1 \leqq i, j \leqq |C|, i \neq j)$ **do**
      if c_i と c_j は親子 $\wedge$ c_j は l 段目 $\wedge$ c_i と c_j は併合条件を満たす
      cen c_i, c_j を併合手法 を用いて併合
    end for
  until 全ての c_i, c_j が併合条件を満たさなくなるまで
end for
D を出力

併合条件

c_i, c_j の相関 $corr(c_i, c_j)$（ 式（ 6 ）参照 ） が閾値 $bound$ 以上

併合手法

for all $d_x \in D$ **do**
  ドキュメントのカテゴリに対する値 $d_x(c_i)$ に $d_x(c_j)$ を併合
  c_j を取り除く
end for

表 1　動的カテゴリ併合法

同一階層カテゴリ併合法と、異階層カテゴリ併合法を組み合わせて用いる。併合法は、最下段のカテゴリ全てに最初に適用し、適用する階層を 1 段づつ上げていく。適用する階層より上位の階層が無くなった時点で終了とする。

表 2　提案法による類似度計算

動的カテゴリ併合法＋従来法
（1） D に含まれる全ての語句の集合 T に対し類義語辞書を用いて語句をカテゴリに変換
（2） カテゴリ間の関係と D の特性から、カテゴリの動的併合法 を適用し、$D_{C'}$ に変換
（3） ドキュメントとカテゴリの関係行列 $D_{C'}$ に特異値分解適用
（4） $D_{C'M}$ 中の上位 k 個の空間 $\mathbf{R}^k$ 中 でドキュメント間類似度を求める

対象としたドキュメント集合は、情報処理学会論文誌 [6] に採録された論文の集合のうち、2000 年 1 月 から 2003 年 12 月までのものとし、これを 1 年単位で区切ることで、4 回実験を行なった。但し、論文の概要のみを対象とし、本文は用いなかった。また、タイトルや概要が英語であったり、概要が空欄

となっているドキュメントは対象としない。各ドキュメント集合の論文数、語句数、分野数は、表 3 のようになった。

表 3　ドキュメント集合の情報

ドキュメント集合名	ドキュメント数	語句数	論文の分野数
Vol.41	292	2739	103
Vol.42	281	2658	98
Vol.43	350	2914	132
Vol.44	279	2606	93

語句数は類義語辞書に含まれている語句の数を示す

形態素解析ソフトの茶筅 [7] とそれに対応する辞書 ipadic [7] をこの論文集合に対して用いる。ドキュメントを語句集合に分解し、ドキュメントに於ける出現数を重みとして、ドキュメントと語句の関係を示す行列 D を生成する。値 v_{ij} はドキュメント中の語句の出現数とした。類義語辞書は日本語語彙大系 CD-ROM 版 [2] を用いた。これはおよそ 30 万語を 2715 のカテゴリに分類したものである。上下には意味の広さに応じて 12 段に別れている。

特異値分解手法は、MaTX [11] に実装されている関数 *svd* を用いた。この *svd* 用いると対象とする行列のランクに関わらず、行と同じ次元数の正方行列の U ($D = U\Sigma V^T$) を得ることができる。提案法、比較対象とする従来法においては、これらを用いることで、語句をカテゴリに分類したカテゴリ空間、カテゴリ空間の直交化を行なった。特異値分解によって得られた行列の列は、Σ 中の特異値の大きい順に並んでおり、上位から任意数で区切ることによって空間を狭めることができる。今回はこの数をパラメータとして用いた場合の類似度の変化も調べる。

従来法と提案法それぞれの場合の実験の流れは表 4, 2 のように行なう。

表 4　従来法による類似度計算

類義語辞書と特異値分解の併用法
（1） D に含まれる全ての語句の集合 T に対し、類義語辞書を用いて語句をカテゴリ化
（2） K_T を用いてドキュメント空間 D を D_C に変換
（3） 上位から定めた段数までのカテゴリを用いて、下位のカテゴリを代表させる
（4） ドキュメントとカテゴリの関係行列 D_C に特異値分解適用し、D_{CM} を得る
（5） D_{CM} 中の上位 k 個の空間 $\mathbf{R}^k$ 中 でドキュメント間類似度を求める

5.2 評価方法

人が考えるようにドキュメント間類似度を得ることが目的となっているが、専門的なドキュメント集合の 1 つ 1 つの内容を人手を用いて吟味することは、評価が主観的になってしまいがちであることや、評価することに多大な労力がかかってしまうので避けた。今回は、ドキュメント間の類似度を得ることに使わなかった論文の分野情報を、精度を測る基準として用いることで 2 つの尺度を定めた。

(尺度 1) 全ての論文から、他の全ての論文に対する類似度を求め、類似度の平均を分野単位で求める。ある論文に対して最も類似する分野が、その論文が含まれていた分野のドキュメント集合である割合を最初の尺度として用いる。

(尺度 2) ある論文に類似する論文を類似度の大きい順に並べ替えた時、同一分野の論文が最初に現れる順位を 2 番目の尺度とした。論文に類似する論文を探す場合、類似度順に並べて、同一分野の論文が現れることが早いということは、自分が持っているドキュメント (論文) に対して、適切な類似度を与えられていることを示す。

5.3 実験結果

ここでは、提案法を用いた場合と用いなかった場合での結果の差異を示す。尺度 1 についての実験結果は 図 3 のようになった。対象とするドキュメント集合は 4 種あるので、それぞれの値は 4 回実験を行なった平均となっている。横軸に特異値分解適用後の次元数を 30 - 240 と取り、縦軸は同一分野が現れる順位を示す。既存手法で用いる階層は 6, 8, 10, 12 の 4 通り提案法で用いた相関のパラメータは 0.3, 0.4, 0.5, 0.6, 0.7 とした 5 通りである。この結果、従来法と提案法共に、パラメータの値による差は見られなかった。

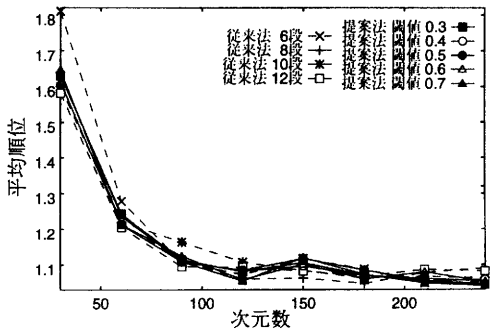

図 3　尺度 1 での比較結果

次に、尺度 2 で比較した場合は図 4 のようになる。これを見ると、提案手法は殆んどの部分で従来法を改善していることがわかる。また、提案法は用いた閾値 : 0.3 - 0.7 では、大きく精度が変わることが無く、常に従来法の精度を改善していることがわかる。両方法で類似しているとされた論文の一部を表 5, 6

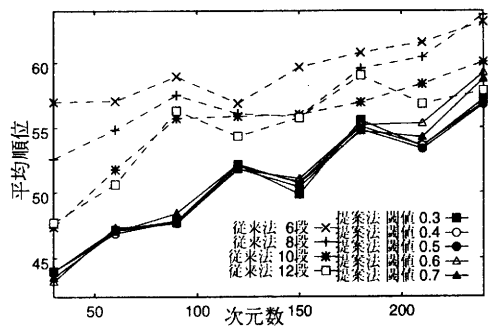

図 4　尺度 2 での比較結果

表 5　従来法 12 段 30 次元

類似度	対象：線形変数変換に不変な自乗値ペナルティ項の構成法
0.571	離散確率分布を持つ リアルタイムシステムの詳細化検証手法
0.472	振舞い近似手法を用いた ステートチャートに対する不変性の検証
0.446	3 次元形状・運動復元のための 高速非線形最適化計算法
0.444	ホモトピー法のパラメータに 一次分数変換を 適用した近接根問題の解法について
0.441	64bit 計算機環境に適した多倍長数値計算環境の 構築と非適切問題の数値計算
0.427	実時間ビジョンシステムのための信頼度駆動メモリ
0.418	マルチ OS 環境を利用した アクセス制御システムの実装と性能評価
	⋮

対象論文と従来法で類似するとされた上位 7 論文を示す

表 6　提案法 閾値 0.7 30 次元

類似度	対象： 線形変数変換に不変な自乗値ペナルティ項の構成法
0.551	EM アルゴリズムの最適ループ回数の予測を用いた 語義判別規則の教師なし学習
0.515	完全 一致法を用いた手書き住所文字列の認識
0.464	3 次元形状・運動復元のための 高速非線形最適化計算法
0.408	ニューラルネットワークを用いた最適化問題 における重み付けの対称性の破れとその効果
0.406	IP ネットワーク上での映像配信サービスを 対象とした利用者指向 QoS 制御手法の提案
0.4045	(同 一分野) 方向線素特徴とノイズ重畳を用いた ニューラルネットワークによる手書き文字認識
0.4044	(同 一分野) 重ね文字を認識する複写学習モデル
	⋮

対象論文と提案法で類似するとされた上位 7 論文を示す

に表す。

6. まとめと考察

ある分野に偏った専門的なドキュメント集合に於いては、専門的語句や、その通常と異なった意味で語句が用いられることが多々あると考えられ、また逆に一般的に用いられる語句が殆んど現れないこともある。類義語辞書の構造に従って、常に均一にカテゴリにまとめては特性を損なってしまう。ドキュメント集合での語句の出現数や共起する語句を考慮することで、より適切な類似度を導くことが可能となる。一般的に広く知られ、多く用いられる語句であっても、対象とするドキュメント集合において現れることが殆んどないのならば、ノイズとなることも有り得る。よって、このような語句は他の語句やカテゴリと併合することで、ノイズとなる可能性を減らすこととなる。実験により、これらの語句を他のカテゴリと併合することや、ドキュメントの特性を示す語句を併合しないことによって精度の向上を見ることでこれらの問題点、解消法を示した。

また、実験により一般的な辞書を専門ドキュメントに利用する事の限界を感じた。ドキュメント集合から語句を得るために用いたソフトウェア茶筅 や、語句間の意味の類似を得ることができる日本語語彙大系は、日本語を扱う形態素解析や類義語辞書としては著名であり、その語句解析能力が高性能であることや、含まれる語数が非常に多いことが広く知られている。しかし、論文のような専門分野のドキュメント集合を扱うことは難しく、未知語と処理された語句が多くあった。

本稿で示す提案手法において、カテゴリ間の相関を決める関数は基本的なものである。それにも関わらず従来法と比べて精度の向上を見ることが出来ている。カテゴリ間の相関求める関数を異なる手法にすることで更に精度の向上が期待できる。

実験の評価手法でのみ、同一分野に分類されるかを精度として用いた。分類問題として捉えれば、機械学習 [12] の前処理としても使える。本稿で提案した手法は学習精度を向上させる特徴選択 [13] や特徴生成手法 [14] [15] と考えることもできる。機械学習分野への応用は非常に興味深い。

文　献

[1] J. Han, M. Kamber. Data Mining – Concepts and Techniques. Morgan Kaufmann Publishers, 2001.

[2] 池原 , 宮崎, 白井 , 横尾, 中岩, 小倉, 大山, 林. 日本語語彙大系 CD-ROM 版. 岩波書店 1999.

[3] C. Fellbaum. WordNet An Electronic Lexical Database. The MIT Press 1998.

[4] 笠原, 稲子, 加藤. 単語の属性空間の表現方法. 人工知能学会論文誌 17 巻 5 号 B, pp. 539–547, 2002.

[5] 笠原, 松澤, 石川. 国語辞書を利用した日常語の類似性判別. 情報処理学会論文誌 Vol. 38 No. 7, pp. 1272–1283, July 1997.

[6] 社団法人 情報処理学会 http://www.ipsj.or.jp/

[7] 形態素解析器 : 茶筅, ipadic. http://www.chasen.org/

[8] G. Salton, C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing and Management Vol.24 No.5, pp. 513–523, 1988.

[9] M.Berry, M. Browne. Understanding Search Engines : Mathematical Modeling and Text Retrieval. SIAM 1999.

[10] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning. Springer, 2001.

[11] 古賀 MaTX. http://www.matx.org/. 2002

[12] T. Mitchel. Machile Learning. McGraw Hill, 1997.

[13] H. Liu & H. Motoda. Feature Selection. Kluwer Academic Publishers, 1998.

[14] S. Markovitch & D. Rosenstein. Feature Generation Using General Constructor Functions. Machine Learning, 49, pp. 59 – 98, Kluwer Academic Publishers, 2002

[15] 大西, 大原, 馬場口. 帰納学習のためのメタ属性を用いた属性生成手法. 第 51 回 人工知能基礎研究会資料, 人工知能学会, pp. 79 – 84, 2003.