

WordNet からの共通概念抽出によるテキスト分類精度の改善

Improving text categorization by extracting common concepts from WordNet

猪野 陽子 松井 藤五郎 大和田 勇人
Yoko Ino Tohgoroh Matsui Hayato Ohwada

東京理科大学 理工学部 経営工学科

Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science.

This paper proposes a new method of extracting feature words in text categorization using WordNet. Thesaurus information in WordNet is useful to improve categorization accuracy. However, it is not practical to use all of those information for categorizing, since the amount of data is too large if we use them. Therefore, in addition to high frequency words, we use common concepts extracted from WordNet. The common concepts are derived from high frequency words and medium frequency words. We consider that a medium frequency word used as paraphrase of a high frequency word is effective for categorizing. We extracted the common concepts of high frequency words and medium frequency words only when they are similar as synonymous words. Experimental results with Support Vector Machine (SVM) and Reuters-21578 are given to show the efficiency of our method.

1. はじめに

電子化された大量の文書が利用可能となったことから、機械学習を利用してテキストを自動分類する研究が進められている。機械学習による分類では、一般的に各テキストをテキスト中に出現する単語を次元とするベクトルを用いて表現する。従って、分類精度はテキストのベクトルの次元としてどの単語を使用するか依存する。このベクトルの次元になる単語のことを特徴語という。

特徴語に関しては、これまでの研究では文書中の単語を使用してきたが、近年の研究ではシソーラス辞書の WordNet などが使用されるようになってきている [福本 02]。WordNet を使用した文書分類は、文書中の単語から得られる情報を手がかりとした辞書情報を利用できる点で、分類精度の向上に有効である。しかし、WordNet 中の情報を全て学習に用いるのは、データ量が増加するため実用的ではない。よって、分類に対して WordNet を使用する場合には、WordNet から特徴語をどのように抽出するかが大きな問題となる。

そこで、本研究では WordNet から分類に有効な情報のみを、特徴語として抽出するための特徴語選択方法を提案する。本手法では、高頻度語ほど出現回数が多いが、ある程度は出現する語を中頻度語として着目する。テキスト中の名詞単語の中から高頻度語と中頻度語のみを使用して WordNet の辞書引きを行い、そこから得られる共通概念と、高頻度語のみを特徴語とする。

本稿では、本手法について詳細に述べた後に、評価実験の結果を示し、提案手法の有効性を述べる。

2. 特徴語選択方法

2.1 手順の概要

本研究における特徴語選択方法は次の手順で行われる。

step1 文書中の名詞単語を DF (Document Frequency) 値によって 3 分割し、大きい順に高頻度語、中頻度語、低頻度語とする。

連絡先: 猪野陽子, 東京理科大学 理工学部 経営工学科 大和田研究室, 千葉県野田市山崎 2641, 04(7124)1501, ino@ia.noda.tus.ac.jp

<高頻度語からの距離2, かつ 中頻度語からの距離1
以下の概念を抽出するとき>

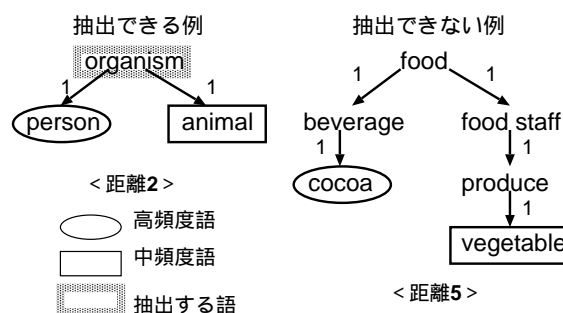


図 1: 共通概念抽出方法

step2 高頻度語と中頻度語の共通概念を WordNet から抽出する。

step3 step1 で抽出された高頻度語と step2 で抽出された共通概念から頻度の高いものを抽出し、特徴語とする。

本研究では、テキスト中の名詞単語のみを特徴語に使用する。

step1 では、a, b を変数とすると、テキスト中の名詞単語から DF 値の上位 a % を高頻度語、その次の b % を中頻度語とする。それ以外の単語は、文書中にほとんど出現しない単語と考え、使用しない。このような単語からは分類精度を向上させる共通概念を抽出できないからである。

step2 において、WordNet 中から共通概念を抽出するために高頻度語と中頻度語を使用する。それは高頻度語の言い替え表現としてテキスト中に存在する中頻度語を、分類に利用するためである。これまでの研究で、特徴語としての高頻度語の有効性は分かっている [Yang 97]。しかし、高頻度ではない語は単語あたりの出現回数が少ないため、分類に有効なもののみを抽出することは難しいと考えられてきた [相澤 03]。本手法では、高頻度語と中頻度語の類似性に着目し、高頻度語と中頻度語が同義語として類似し共通概念を持っている場合にそれを抽出することによって、分類に有効であると考えられる中頻度語のみを利用する。

2.2 共通概念の抽出方法

WordNet 中の単語は、図 1 のように階層構造の形で格納されている。WordNet 中では、同義の単語が synset と呼ばれる

ノードに含まれ、ノード同士はそれらの関係を示すリンクで結ばれている。

共通概念抽出時には、図 1 に矢印で示されている WordNet のリンクの 1 つを、距離 1 として数える。そして、単語間の距離が閾値以下の場合にのみ、それらの共通概念を抽出する。閾値は、高頻度語からの距離における閾値と、中頻度語からの距離における閾値の 2 つを設定でき、全体の閾値はこの 2 つを足したものとなる。

共通概念抽出について、図 1 の例を用いて説明する。高頻度語からの閾値 2、中頻度語からの閾値 1 である距離 3 以下の概念を抽出する場合、左図では高頻度語 person と中頻度語 animal の距離は 2、右図では高頻度語 cocoa と中頻度語 vegetable の距離が 5 である。ここで、左図の共通概念 organism は高頻度語と中頻度語からの距離が双方ともに 1 であり、閾値内の数であるのに対し、右図の共通概念 food は高頻度語からの距離が 2 であり閾値内であるが、中頻度語からの距離が 3 であり閾値 1 より大きい距離になってしまう。よって、この例の場合、左図の共通概念 organism だけが抽出され、右図の共通概念 food は抽出されない。

上記の作業を同一テキスト内の高頻度語と中頻度語のすべての組み合わせに対して行う。この作業によって、テキスト中の中頻度語は、高頻度語と WordNet 中の距離が近い、すなわち同義語としての類似性が大きい場合には共通概念に置き換えられ、WordNet 中の距離が遠い、すなわち同義語としての類似性が小さい場合には削除される。結果として、類似性が大きい高頻度語と中頻度語の組み合わせ数だけ共通概念が得られる。このアルゴリズムを、表 1 に示す。

2.3 頻度による特徴語の決定

本手法では、高頻度語と抽出された共通概念を特徴語としてテキスト分類を行う。しかし、一般に、特徴語は、テキスト中に大量に含まれているか、カテゴリーに独特な単語でないと分類に有効に働かない。高頻度語はテキスト中に大量に含まれていることが分かっているが、抽出した共通概念はどの程度含まれているのかが分かっていない。

そこで、本手法では、高頻度語と共通概念のうち、頻度の高いものだけを特徴語として使用する。これにより、頻度の低い特徴語を排除するとともに、抽出された共通概念よりも頻度の低い高頻度語も排除することができる。

3. 評価実験

提案手法の有効性を確認するために、テキスト分類のベンチマークの 1 つである Reuters-21578 を用いて実験を行った。Reuters-21578 は 12,902 テキストからなり、総計 118 の分野に分類されている。ApteMod を用いて分類のためのテキストを抽出した結果、トレーニングデータ 7,769 文書、テストデータ 3,019 文書となり、分野数は総計 90 分野となった。

実験の前処理では、stop word による不要語処理、Bril's Tagger による品詞付けを行い、名詞単語を抽出した。また、高頻度語と中頻度語の抽出においては、DF 値の大きさが文書中の全単語の上位 10 % を高頻度語、その次の 20 % を中頻度語として抽出した。

テキスト分類には、代表的な Support Vector Machine (SVM) システムである MySVM を使用した。SVM は任意のクラスに属するか否かを判定する二値分類の学習アルゴリズムである [Nello 00]。しかし、テキスト分類は一般的に複数の分野から一分野を決定するマルチクラス分類である。よっ

表 1: 共通概念抽出アルゴリズム。DF(t, B) は文書集合 B 中の単語 t の頻度度を表す。 $d(h, m)$ は高頻度語 h と中頻度語 m の距離を、 $CC(h, m)$ は h と m から得られる共通概念を表す。

```

input: 文書全体の集合  $B$ 
output: 共通概念の集合  $C$ 
local variables:  $x$ : 高頻度語の数
                   $y$ : 中頻度語の数
                   $z$ : 特徴語の数
                   $lim$ : WordNet から共通概念を抽出する際の
                        閾値
1 begin
2  $N \leftarrow B$  の名詞単語のリスト
3 各単語  $t \in N$  の DF( $t, B$ ) を計算
4  $H \leftarrow N$  中の頻度の上位  $x$  個の単語
5  $M \leftarrow N$  中の頻度の次の上位  $y$  個の単語
6 共通概念の集合  $C \leftarrow \emptyset$ 
7 foreach 文書  $D$  in  $B$  do
8   高頻度語の集合  $H_D \leftarrow \emptyset$ 
9   中頻度語の集合  $M_D \leftarrow \emptyset$ 
10  foreach 単語  $t \in D$  do
11   if  $t \in H$  then  $H_D \leftarrow H_D \cup \{t\}$ 
12   else if  $t \in M$  then  $M_D \leftarrow M_D \cup \{t\}$ 
13   endif
14  endfor
15  foreach 高頻度語  $h \in H_D$  do
16   foreach 中頻度語  $m \in M_D$  do
17    if  $d(h, m) \leq lim$  then
18      $C \leftarrow C \cup \{CC(h, m)\}$ 
19   endif
20  endfor
21 endfor
22 endfor
23 end

```

て、本稿では SVM を使用してマルチクラス分類を行うために One-against-the-Rest [Tax 02] を用いた。

SVM への入力データベクトルは、特徴語の出現頻度に基づいて作成した。そして、ベクトルで表現された各テキストに対し、そのテキストが任意の分野に属する (1) か否 (-1) かを示すラベルを付与した。

評価方法には、Recall(精度) と Precision(再現率) の調和平均である F-measure を用いた。F-measure に関しては、分野ごとに F 値を計算してその平均を求めたマクロ平均 F 値、および、分野を区別せず全 5 分野に対して Recall と Precision を計算したものからの F 値であるマイクロ平均 F 値を求めた。

3.1 閾値の違いによる分類精度の比較

提案手法では閾値の設定が、共通概念の抽出量に直接関わっている。よって、共通概念の抽出量が分類精度にどのように影響するかを確認するために、閾値を変化させて実験を行った。今回の実験ではどの程度の閾値が有効であるのかを確認するため、高頻度語と中頻度語から等距離ずつ閾値をとった全体の閾値である 2, 4, 6, 8 を用いた。尚、閾値を変化させても SVM の入力データの次元数は全て等しくした。結果を図 2 に示す。

マイクロ平均、マクロ平均ともに閾値 4 のときが最も精度が良いという結果が得られた。

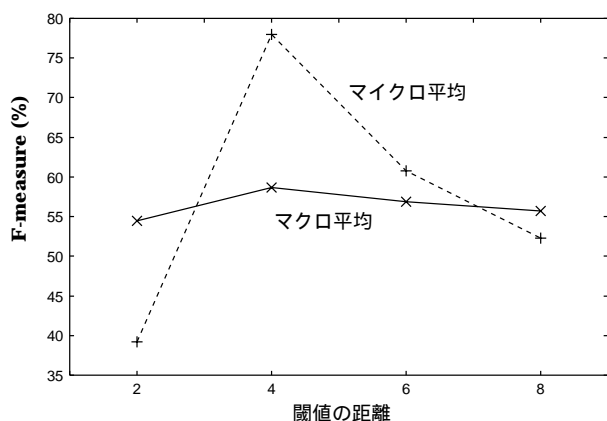


図 2: 閾値による分類精度の違い

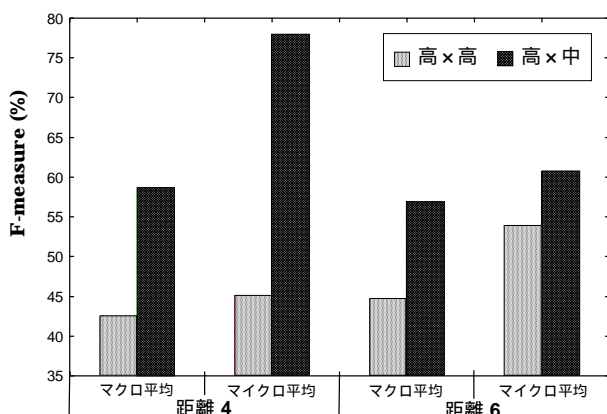


図 3: 中頻度語を使用しない場合との比較

3.2 中頻度語を使用しない場合との比較

本研究では高頻度語の言い替え表現である中頻度語に着目し、それらの共通概念を抽出して分類に使用することにより分類精度の向上の実現を試みている。そこで、高頻度語を中頻度語を用いることの有効性を確認するために、二つの高頻度語の共通概念を抽出した場合との比較を行った。高頻度語のみを用いた実験では、本手法で中頻度語を用いる部分を全て高頻度語に置き換えた。この実験においても SVM の入力データの次元数は全て等しくした。結果を図 3 に示す。

実験は高頻度語と中頻度語から等距離ずつだった全体の閾値 4 と 6 で行ったが、結果は両者ともに高頻度語と中頻度語を使用したときの方が分類精度が高くなった。

4. 考察

4.1 分類精度に与える閾値の影響

閾値を変化させた実験では、図 2 から閾値 4 が最も精度が良く、閾値 4 以上は閾値が大きくなるにつれて分類精度が下がった。これは閾値の設定により抽出される共通概念の量と有効性に関係すると考えられる。

閾値を小さく設定した場合には、高頻度語と中頻度語の距離が近く、同義語として類似性がかなり大きくないと共通概念は抽出できない。よって、この場合抽出される共通概念は分類に有効であることが多いと考えられるが、閾値を小さくするほど抽出量は少量になり、分類精度を変化させる力は減ってしまう。これに対し、閾値を大きく設定した場合は閾値を小さくしたときと逆に、より抽象的な概念が大量に抽出されてしまう。しかし、SVM は特徴語が作成する入力ベクトルの独自性をも

とに分類を行うため、カテゴリに独特でない抽象的な単語を大量に抽出しても分類精度の向上に効果はない。このことから、距離の閾値は、大きすぎず小さすぎない 4 が最適である。

4.2 使用する語の頻度による分類への影響

中頻度語を使用しない場合との比較においては、高頻度語と中頻度語を使用した提案手法の方が精度が高いという結果が得られたが、これは抽出される共通概念の数の違いによるところが大きいと考えられる。

共通概念抽出時には、2 つの単語同士を WordNet にかけて共通概念を抽出したため、抽出可能な単語数は WordNet 中で距離の近い 2 つの単語の組み合わせ数であった。高頻度語はテキスト中に多量に存在するため、WordNet 中の距離が近い高頻度語と高頻度語が多い場合、高頻度語と高頻度語から抽出される共通概念は多量になる。このとき、テキスト中の高頻度語自体よりも多量の概念が抽出されることがある。しかし、本手法では共通概念抽出後に高頻度語と共通概念を合わせたものから頻度度の高い部分のみを特徴語とする。よって、分野に独特で分類に有効である高頻度語の一部でも、頻度度が共通概念よりも小さい場合には除去されてしまう。そして、共通概念は多くのテキストに含まれる一般的な語となってしまう、テキスト分類に有効な語とはならない。このことから、高頻度語と高頻度語からの共通概念の使用が精度の向上に結びつかなかったと考えられる。

一方、高頻度語と中頻度語を使用した場合には、共通概念が適量抽出されたため、有効な高頻度語も除去されることなく使用でき、分類精度の向上につながったものと考えられる。

最後に比較のため、WordNet を使用せずに高頻度語のみを特徴語として実験を行った結果を確認したところ、マクロ平均 54.7 %、マイクロ平均 57.2 %であった。このことから WordNet を使用した本提案手法は有効であるといえる。

5. まとめ

本稿では、WordNet を使用したテキスト分類の特徴語として、テキスト中の高頻度語と、WordNet 中に高頻度語と中頻度語をかけたときの共通概念を使用する手法を提案し、その評価実験を行った。その結果、提案手法における有効な閾値と、高頻度語のみでなく中頻度語を利用した提案手法の有効性が確認された。また、WordNet を使用しない精度との比較から、WordNet を使用した提案手法の有効性も確認された。

今回の実験では、頻度度の上位 10 % を高頻度語として、次の 20 % を中頻度語として使用したが、これらの値を変化させることによって抽出される共通概念も変化する事が予想される。また、共通概念抽出時の閾値の設定法を変更することや、入力データの次元数を変化させることによる分類精度の向上も期待できる。

参考文献

[福本 02] 福本 文代, 鈴木 良弥: WordNet の同義語クラスとその上位関係を利用した文書の自動分類, 情報処理学会論文誌, Vol. 43, No. 6, pp. 1852-1865 (2002).

[Yang 97] Yang, Y., Pedersen, J.O.: A Comparative Study on Feature Selection in Text Categorization, Proc. of the 14th International Conference on Machine Learning (ICML97), pp. 412-420 (1997).

- [相澤 03] 相澤 彰子: 低頻度語の利用によるテキスト分類性能の改善と評価, 情報処理学会論文誌, Vol. 44, No. 7, pp. 1720-1730 (2003).
- [Tax 02] Tax, D.J.M. and Duin, R.P.W.: Using Two-Class Classifiers for Multiclass Classification, Proc. of the 16th International Conference on Pattern Recognition (ICPR16), Vol.2, pp. 124-127 (2002).
- [Nello 00] Nello Cristianini, J. S.-T.: Support Vector Machines, in An Introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, pp. 93-124 (2000).