WordNet からの共通概念抽出によるテキスト分類

Extracting common concepts from WordNet to classify documents

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 大学院理工学研究科 経営工学専攻
Department of Industrial Administration, Graduate School of Science and Technology, Tokyo University of Science

†† 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

ino@ia.noda.tus.ac.jp {matsui, ohwada}@ia.noda.tus.ac.jp

テキスト分類における特徴語選択の際に，英語のシソーラス辞書である WordNet から有効な情報を抽出するための手法を提案する．本研究では，特徴語としての有効性が認められている高頻度語と、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との共通概念に着目する．提案手法では，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，高頻度語から特徴語を抽出する．Support Vector Machine (SVM) と Reuters-21578 を使用した実験の結果，精度の向上が見られ，WordNet を使用した提案手法の有効性と，高頻度語と中頻度語から抽出された共通概念の有効性が確認された．

1 はじめに

World Wide Web やデジタルライブラリー等の出現によって電子化された大量の文書が利用可能となったことから，機械学習を利用した文書の自動分類に関する研究が多く行われている [Lewis 91]．テキスト分類問題においては多くの機械学習手法が利用されているが，近年の研究では， Support Vector Machine (SVM) の有効性が明らかになっている [Joachims 98].

SVM の各事例は，各テキストの特徴を表す特徴空間中において，ベクトルとして表現される．SVM は n 次元 Euclid 空間上に配置されたデータを二分する超平面を探索する．我々が SVM を使用する際には，最善の超平面を得るために，適切な特徴空間を決定しなければならない．テキスト分類問題においては，ベクトルは文書中に出現する単語 (節，文) から生成される．我々は，これらの語を**特徴語**と呼ぶ．分類精度は特徴語にどの単語を使用するかに依存する．

これまでの研究では，特徴語に高頻度語が頻繁に使用されてきた [Yang 97]．しかしながら，相澤は特定のカテゴリーに関連する語である場合，高頻度語以外の単語も有効であることを示した [相澤 03]．また，福本は，単語の関係を使用するために，シソーラス辞書である WordNet を使用した [福本 02]．WordNet の同義語を使用して，文書中の単語をまとめることにより，分類精度を向上させた．このように，近年では，特徴語に関して，テキスト中の情報だけでなく，語の関係が使用されている．

特徴語の選択において，我々は，単語の言い替え表現に着目する．それは，日頃文書を作成している人々は，句や単語の同語反復を用いずに記事を完成する傾向があるという仮定による．故に，新聞記事や通信社の記事には，単語や句の言い替え表現が多く存在する．福本の手法のように，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，この様なケースにも有効である．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，本稿では，WordNet を使用して一般的な概念を抽出し，SVM によるテキスト分類の特徴語として使用する方法を提案する．特徴語の抽出には，出現頻度に基づくスコア (DF, IDF, TFIDF 等) が用いられるが，我々は，スコアが高い語だけでなく，スコアが高い語の言い替え表現を見つけ出し，それらの語の上位概念の語を特徴語に加える．具体的には，スコアが高い語とそれ以外の語を WordNet で辞書引きし，二つの語の共通の上位概念のうち最も下位の概念を**共通概念**として抽出し，特徴語として使用する．

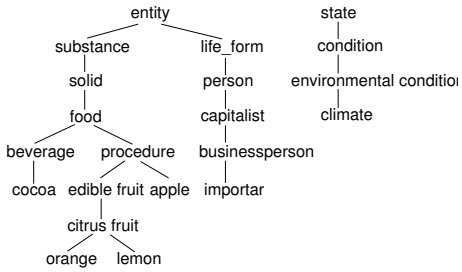

図 1: WordNet の構造

2 WordNet と共通概念

WordNet [Miller 90] は英語のシソーラス辞書であり，図 1 のように単語同士の関係が品詞別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．

WordNet において語 t が語 t' の下位概念であることを $t \prec t'$ と表し，語 t が語 t' の下位概念であるかまたは同じ語であることを $t \preceq t'$ と表す．また，語 t の上位にある語の集合を $\mathcal{G}(t) = \{t' | t \prec t'\}$ と書く．このとき，次の条件を満たす語 t を語 a と語 b の**共通概念**$CC(a, b)$ と定義する．

1. $t \in \mathcal{G}(a) \cap \mathcal{G}(b)$
2. $\forall t' \in \mathcal{G}(a) \cap \mathcal{G}(b)[t \preceq t']$

3 特徴語選択方法

3.1 手順の概要

本研究における特徴語選択は次の手順で行われる．

step1 文書中の名詞単語を DF (Document Frequency) 値の大きいものから順に 3 分割 (高頻度語, 中頻度語, 低頻度語) する．

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出度の高いものを選択し，特徴語とする．

step1 では，文書中の各名詞に対し，DF 値を計算する．そして、分類精度に影響を与える量の共通概念を抽出できる様、DF 値の大きい語のみを使用する．

step2 において，WordNet から高頻度語と中頻度語の共通概念を抽出する．共通概念を抽出する際，我々は中頻度語に焦点を当てる．その理由は，中頻度語のいくつかには，各文書中に存在する高頻度語の言い替え表現となるものが存在するためである．表 1 にこの一例を示す．より良い精度を得るために，我々は，各カテゴリーに対して特別な単語から抽出される共通概念を使用する必要があると考える．各カテゴリーに対して特別な単語を抽出する方法は他にも存在するが，より簡単にそれらを抽出できるよう，中頻度語を使用する．高頻度語と中頻度語が WordNet 中で近い場所に存在する場合，我々はそれらが同じ概念を示すと仮定する．step3 では，抽出された共通概念と高頻度語から，特徴語の選出を行う．

表 1: 共通概念の例.

高頻度語	中頻度語	共通概念
device	system	instrumentality
supertanker	shipping	conveyance
confederation	rally	group_action

3.2 アルゴリズム

表 2 にアルゴリズムを示す．このアルゴリズムを使用して，我々は文書全体の集合 B から特徴語の集合 F を選出する．特徴語の集合 F は，文書全体の集合 B 中の単語と WordNet から抽出された共通概念からなる．前節における step1，step2，step3 はそれぞれ，アルゴリズムの 2-5 行，6-26 行，27-28 行に相当する．

3.2.1 名詞単語分割過程

文書全体の集合 B から名詞単語のみを抽出し，N とする．次に，N 中の各単語 t に対して DF 値を計算する．DF 値は，単語 t を 1 つ以上含む文書の数であり，$\mathrm{DF}(t, B)$ と記述する．

続いて，$\mathrm{DF}(t, B)$ によって N を 3 分割 (高頻度語 H, 中頻度語 M, それ以外) する．この際，高頻度語の数を α，中頻度語の数を β とする．H は N における上位 α 個の単語，M は次の上位 β 個の単語となる．

3.2.2 共通概念抽出過程

共通概念抽出は，全文書 B における各文書 D に対し，独立に行われる．はじめに，各文書 D から H を抽出し，H_D と定義する．次に，同様に各文書 D から M を抽出し．M_D と定義する．

続いて，WordNet から共通概念を抽出する．この

表 2: 共通概念抽出アルゴリズム.

input: 文書全体の集合 B
output: 特徴語の集合 F
local variables:
α: 高頻度語の数.
β: 中頻度語の数.
γ: 特徴語の数.
lim: 閾値.

1. **begin**
2. $N \leftarrow B$ 中の名詞単語の集合
3. 各単語 $t \in N$ に対する $\mathrm{DF}(t, B)$ を計算する
4. $H \leftarrow N$ 中の頻出度上位 α 個の単語
5. $M \leftarrow N$ 中の次の頻出度上位 β 個の単語
6. 共通概念の集合 $C \leftarrow \emptyset$
7. 拡張された全文書 $B' \leftarrow \emptyset$
8. **foreach** B 中の文書 D **do**
9. 高頻度語の集合 $H_D \leftarrow \emptyset$
10. 中頻度語の集合 $M_D \leftarrow \emptyset$
11. **foreach** 単語 $t \in D$ **do**
12. **if** $t \in H$ **then** $H_D \leftarrow H_D \cup \{t\}$
13. **else if** $t \in M$ **then** $M_D \leftarrow M_D \cup \{t\}$
14. **endif**
15. **endfor**
16. 拡張された文書 $D' \leftarrow \emptyset$
17. **foreach** 高頻度語 $h \in H_D$ **do**
18. **foreach** 中頻度語 $m \in M_D$ **do**
19. **if** $d(h, m) \leq lim$ **then**
20. $C \leftarrow C \cup \{\mathrm{CC}(h, m)\}$
21. $D' \leftarrow D' \cup \{\mathrm{CC}(h, m)\}$
22. **endif**
23. **endfor**
24. **endfor**
25. D' を B' に加える
26. **endfor**
27. 各単語 $t \in C$ に対し $\mathrm{DF}(t, B')$ を計算する
28. $F \leftarrow N \cup C$ における頻出度上位 γ 個の単語
29. **end**

際，距離の閾値を lim とする．これは，WordNet 中の二つの単語から離れた場所に存在する共通概念を抽出しないよう，制限するものである．高頻度語と中頻度語の距離が lim 以下であり，意味的に近い場合にのみ，それらの共通概念を抽出する．

$h \in H_D$, $m \in M_D$, $d(h, m) \leq lim$ から共通概念 $\mathrm{CC}(h, m)$ を発見できるとき，$\mathrm{CC}(h, m)$ を共通概念の集合 C と拡張された各文書 D' に追加する．WordNet からの共通概念抽出方法は次節で説明する．

3.2.3 特徴語抽出過程

抽出された全ての共通概念が分類に対して重要であるとは限らないので，共通概念の集合 C と高頻度語の集合 H から特徴語を選出する．

はじめに，拡張された全文書 B' における共通概念 C 中の各単語 t に対し，DF 値を計算する．そして，$H \cup C$ から特徴語として上位 γ 個を選出する．最後に，特徴語の集合を F と定義する．

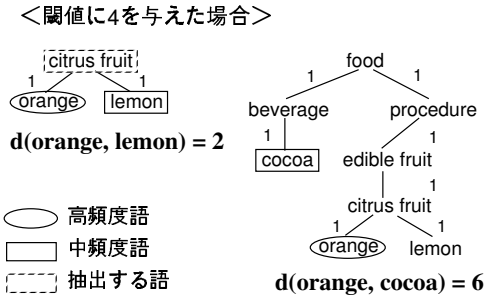

図 2: 共通概念抽出方法

3.3 共通概念抽出方法

本節では，WordNet からの共通概念抽出方法を，詳細に説明する．

本研究では，単語 t_1 と t_2 が一本のリンクでつながれている場合，それらの距離 $d(t_1, t_2)$ を 1 と定義する．故に，2 単語間のリンクの数を数えることにより，それらの距離の計算を行う．ここで得られた計算値に基づき，概念を抽出する．

高頻度語 h と中頻度語 m の距離 $d(h, m)$ が，与えられた閾値 lim 以下である場合 ($d(h, m) \leq lim$ の場合)，我々は最も近くに存在する共通概念を抽出する．

この過程について，図 2 の例を用いて説明する．2 単語間の閾値に 4 を設定する場合 ($lim = 4$)，左図では “orange” と “lemon” の距離 $d(\mathrm{orange}, \mathrm{lemon})$ は 2 であるので，閾値 lim より小さい．よって，最も近くに存在する共通概念 “citrus fruit” をそれらの共通概念として抽出する．他方，右図では，“orange” と “cocoa” の距離 $d(\mathrm{orange}, \mathrm{cocoa})$ は 6 であり閾値を超えているため，それらの共通概念 “food” を抽出しない．

4 評価実験

4.1 実験方法

提案手法の有効性を確認するために，実験データ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．ApteMod という方法を用いて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 [福本 02].

実験の前処理では，stop word による不要語処理，Brill’s Tagger [Brill 94] 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．

分類器には，代表的な SVM システムである MySVM [Ruping 00] 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 [Vapnik 95]．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カテゴリーに属する (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 F1-measure を用いた．F1-measure に関しては，カテゴリーごとに F1 値を計算してその平均を求めたマクロ平均 F1 値，および，分野を区別せず全 90 分野に対して Recall と Precision を計算したものからの F1 値であるマイクロ平均 F1 値を求めた．

また，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い，結果を確認したところ，マクロ平均 F1 値 54.7 %，マイクロ平均 F1 値 57.2 %であった．

4.2 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．閾値には，2，4，6，8 を用いた．尚，閾値を変化させても入力データの次元数は γ によって 2000 に統一した．公正な比較のため，トレーニングデータに 5 フォールドクロスバリデーションを適用した．

結果を図 3 に示す．マイクロ平均，マクロ平均ともに閾値 2 のときに最も精度が良いという結果が得られた．

4.3 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

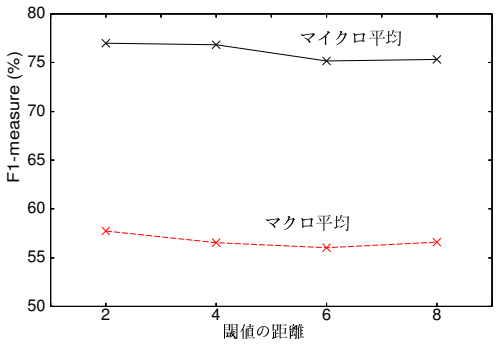

図 3: 閾値の違いによる分類精度の比較

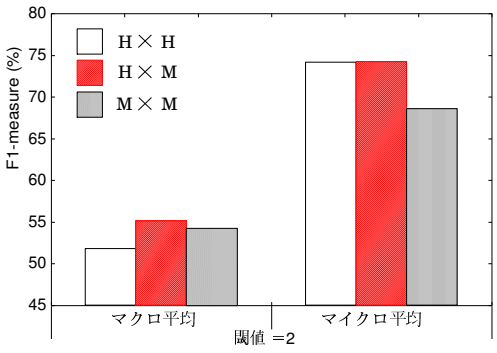

図 4: H×H, H×M, M×M の比較

用している．そこで，高頻度語と中頻度語を用いることの有効性を確認するために，高頻度語同士の共通概念を H × H，高頻度語と中頻度語の共通概念を H × M，中頻度語同士の共通概念を M × M とし，これらを使用して実験を行い，その結果を比較する．我々は，H × H，H × M，M × M のそれぞれの F1 値と処理時間を比較した．処理時間は，計算機が WordNet から共通概念を抽出する時間である．入力データの次元数は 2000 に設定し，閾値には上記の実験結果からその有効性が明らかになった 2 を，データには ApteMod 法によって生成されたトレーニングデータとテストデータを使用した．

図 4 に F1 値の結果を，処理時間の結果を表 3 に示す．これらの結果から，H × M の組合せが最も有効であることが分かった．また，H × M の共通概念抽出時間は，特徴語抽出全過程に費やす時間の約 1/5 であった．

表 3: 処理時間の比較

	マクロ平均 F1(%)	マイクロ平均 F1(%)	Time(s)
H×H	51.77	74.10	15906.98
H×M	55.11	74.20	5237.96
M×M	54.27	68.53	1162.05

5 考察

5.1 閾値の違いによる精度への影響

提案手法における閾値は，抽出する概念の数の制御を行っている．実験では，特徴語数を等しくしたが，閾値の違いによって抽出される共通概念は異なっている．

閾値を大きく設定した場合には，多くの共通概念が抽出される．しかしながら，2 単語がお互いに類似していない場合にも，概念が抽出されてしまうことがある．この場合，より抽象的な概念 (entity, abstract, location 等) も抽出される．故に，特徴語には分類に有効でない単語も含まれてしまうといえる．

閾値を小さく設定した場合には，WordNet 中での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ければ共通概念は抽出できない．それ故，2 単語がより類似する場合に共通概念は抽出されるが，多く抽出することは難しい．共通概念は，より類似性が高い場合に，言い替え表現としての役目を果たす．故に，有効に共通概念を抽出するため，適切な閾値を選出するべきである．

5.2 中頻度語の効果

中頻度語の中には，高頻度語の言い替えになるものが存在するため，共通概念を抽出する際に，中頻度語に焦点を当てた．我々は，高頻度語は多くのカテゴリーに出現する一方，中頻度語は特定のカテゴリーに出現すると考える．

二つの高頻度語 (H × H) の共通概念を抽出するとき，高頻度語が特徴的でないため，獲得された共通概念は一般的すぎてしまう．よって，その共通概念はカテゴリーを特定できない．一方，二つの中頻度語 (M × M) から共通概念を抽出するときは，とても特別な概念のみが抽出されてしまう．

図 4 の実験結果では，高頻度語と中頻度語 (H × M) の組合せが共通概念を抽出するために効果があることを示す．故に，H × M は提案手法に適しているといえる．

6 まとめ

本稿では，WordNet と SVM を使用したテキスト分類の特徴語抽出に対し，共通概念抽出法を提案した．これは，名詞単語を DF 値によって 3 分割 (高頻度語, 中頻度語, 低頻度語) し，WordNet から高頻度語と中頻度語の共通概念を抽出する．そして，DF 値を使用して，高頻度語と抽出された共通概念から，SVM にかける特徴語を選出する．

Reuters-21578 を使用した実験結果から，WordNet から抽出される共通概念は SVM を使用したテキスト分類に有効であることが確認された．また，中頻度語は高頻度語とともに使用されたとき，共通概念を抽出するのに有効であることも確認された．

SVM を使用したテキスト分類の精度は特徴語の数によるため，提案手法に対する最適な高頻度語，中頻度語，特徴語の数を発見することで，更な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).

[Brill 94] Eric Brill: Some Advances in Transformation-Based Part of Speech Tagging, Proc. of the 12th National Conference on Artificial Intelligence, pp. 722–727 (1994).

[Joachims 98] T. Joachims: Text categorization with Support Vector Machines: Learning with many relevant features, Proc. of the 10th European Conference of Machine Learning, pp. 137–142 (1998).

[Lewis 91] David D. Lewis: Evaluating text categorization, Proc. of DARPA Speech and Natural Language Workshop, pp. 312–318 (1991).

[Miller 90] George A. Miller and Richard Beckwith and Christiane Fellbaum and Derek Gross and Katherine J. Miller: Introduction to WordNet: An on-line lexical database, International Journal of Lexicography, 3(4): pp. 235–312 (1990).

[Ruping 00] Stepan Rüping: mySVM-Manual, University of Dortmund, Lehrstuhl Informatik 8 (2000), http://www-ai.cs.uni-dortmund.de/SOFTWARE/MYSVM/.

[Tax 02] D. J. M. Tax and R. P. W. Duin: Using Two-Class Classifiers for Multiclass Classification, Proc. of the 16th International Conference on Pattern Recognition, Vol. 2, pp. 124–127 (2002).

[Vapnik 95] V. Vapnik: The Nature of Statistical Learning Theory, Springer-Verlag, NY, USA (1995).