単語の共起関係を利用した概念的特徴ベクトルの生成

福元　伸也†　　渕田　孝康†

† 鹿児島大学大学院理工学研究科　〒 890–0065 鹿児島市郡元 1-21-40
E-mail: †{fukumoto,fuchida}@ibe.kagoshima-u.ac.jp

あらまし　Web 上の膨大なデータの中から有益な情報を見出そうとする試みが広く行われており，テキストデータ解析に関する多くの研究が行われている．本研究では，情報検索において必要な特徴ベクトルの生成において，語を意味によって分類・整理した分類語彙表を利用して概念的な特徴ベクトルを生成し，機械学習フレームワークの 1 つである Jubatus を用いてテキスト分類する．出現単語のみによる共起行列では，特徴ベクトルの次元数増大やベクトルのスパース性の問題がある．分類語彙表を用いた特徴ベクトルの生成では，次元数の抑制や，より的確なベクトル表現が期待できる．実験では，ニュース記事におけるテキスト分類を通じ，Jubatus 上で複数の学習アルゴリズムを用いて分類精度を検証した．

キーワード　文書分類，特徴ベクトル，シソーラス，共起行列，機械学習

1. はじめに

近年，インターネットの普及やクラウド環境の充実により，膨大な量のデータを扱う機会が増大しており，ビッグデータに関する研究が脚光を浴びている．情報検索においては，テキストデータを，ある特徴に従いグループ分けするために，テキスト分類に関する研究が行われてきた．大量の文書データを，効率よく分類する手法も数多く提案されている [1], [2]．テキスト分類では，文書を特徴ベクトルで表現し，文書ベクトル間の類似度を定義し，分類を行っている．そのため，テキストは文字データとして扱われ，語の持つ意味までは考慮されていない．

本研究では，特徴ベクトルの生成において，文書中に現れた出現単語とシソーラスの単語の意味属性を用いて，共起頻度による共起行列を生成する．シソーラスには，分類語彙表を用いる [3]．分類語彙表は，長い年月にわたり語を意味によって分類・整理した類義語集であるため，語の持つ意味的な概念を特徴ベクトルの生成にうまく反映させることが期待できる．

通常，文書内には，似たような意味を持つ複数の単語が存在する．ある単語に隣接して別の単語が現れることを共起と言い，単語同士の共起頻度を利用した共起行列を用いて文書分類を行うと，本来似ている意味の単語が，距離の離れた特徴ベクトルとして表現されるため，分類精度の低下が生じてしまうという問題がある [4]．

本研究では，語を意味により分類したシソーラスである分類語彙表を用いて，単語の意味を考慮した共起行列を作成し，その共起行列を学習データとして，分類のための学習器に与える．学習器には，大規模データのさまざまなデータ分析に優れた性能を示している機械学習フレームワークである Jubatus を利用する [5]．実験では，ニュース記事の分類を通じて，Jubatus 上でクラスタリングのための複数の学習アルゴリズムを用いて，分類精度について検証する．

2. 関連研究

ネット上の膨大なテキストデータより，新たな知見や有用な情報などを得るため，テキスト解析技術に関する研究が行われている．文書分類の研究として，単語の係り受け関係を用いて分類を行う研究や文書中に現れる語の共起関係を用いたもの [6] などがあり，また，テキストデータのソースにマイクロブログ扱う研究 [7] やテキストデータの時間表現を活用する Temporal Information Retrieval という分野の研究 [8] なども行われている．それらテキスト解析に関する多くの研究が，分類精度の向上にチャレンジしており，分類に関するさまざまな学習法を提案している．Wang らは，語の重要度の決定に，RageRank アルゴリズムを用いる手法を提案し，ランク付けが文書分類に有効であることを示した [9]．また，単語の共起行列を作成するために，文書に現れた単語同士の共起頻度を利用した手法がある．単に単語同士の共起頻度を取るだけでは，意味的に近い単語であっても，別の共起頻度としてカウントされ，その特徴ベクトル間の距離が離れてしまう問題があった [10]．別所らは，単語同士の共起頻度ではなく，単語とコーパスにおける単語に付随する意味属性との共起頻度を取る手法を提案し，共起ベクトルの質が向上することを示した [11]．

3. 共起行列の作成

3.1 共起行列

1 つの文中に現れた単語は，単語の位置が近いため，意味的に近い関係であろうという仮定のもと，これらの単語は，共起関係にあると言う．その共起関係に基づき，ある単語と別の単語の共起関係の頻度を成分にした行列が共起行列である．単語間の共起頻度を利用した共起行列では，テキストデータの対象となったすべての単語が含まれてしまうため，行列の大きさが巨大になってしまう．行列が大きくなると次のような問題が

ある．

- 行列の次元数の増大に伴い，計算コストも増大する
- 行列がスパース（疎）になってしまう
- 本来，近い関係にあるべき特徴ベクトルが離れた状態になってしまう

そこで，単語行列を属性行列に変換する手法が提案されており，笠原らは，国語辞典を用いる手法を提案している [12].

3.2 単語による特徴ベクトル

全テキストデータに含まれる単語を w_i とし，N 個の単語が含まれているとすると，単語 w_i の特徴ベクトルは次のように表される．

$$w_i = (v_{i1}, v_{i2}, \cdots, v_{iN}) \tag{1}$$

ただし，v_{ij} は w_i における重みである．単語特徴ベクトル w_i を要素とした列ベクトルは，次のような行列で表される．

$$F_w = \begin{pmatrix} w_1 \\ \vdots \\ w_N \end{pmatrix} = \begin{pmatrix} v_{11} & \cdots & v_{1N} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{NN} \end{pmatrix} \tag{2}$$

この単語特徴行列から属性行列を生成する．属性数（行列の列数）を m とすると，単語の属性ベクトル $\acute{w}_1, \cdots, \acute{w}_N$ および，その列ベクトルは次の行列で表される．

$$F_p = \begin{pmatrix} \acute{w}_1 \\ \vdots \\ \acute{w}_N \end{pmatrix} = \begin{pmatrix} \acute{v}_{11} & \cdots & \acute{v}_{1m} \\ \vdots & \ddots & \vdots \\ \acute{v}_{N1} & \cdots & \acute{v}_{Nm} \end{pmatrix} \tag{3}$$

ただし，$\acute{v}_{ij}$ は $\acute{w}_i$ における重みである．

3.3 分類語による共起行列

共起行列の作成において，単語間の共起頻度を利用した共起行列の生成では，似た意味の単語であるのに，単語間の共起ベクトルの距離が離れてしまう問題が指摘されている [11].

図 1 は，左端の列が記事中に現れた単語の例を表しており，上段の共起語は，同一文章中に現れた共起語を表している．また，表中の数字は，その出現頻度を表している．単語間の共起に基づいた共起行列の作成手法では，出現語 A と B が，意味の近い単語であったとして，その共起語 a と b が別々にカウントされる．(a と b も意味の近い単語）そうなると，A と B のベクトルは，意味の近い単語であるにもかかわらず離れてしまう．そこで，共起語に現れた a と b を同じ意味を表す 1 つの単語にまとめることが出来れば，A と B のベクトルは離れない．

分類語
意味属性　図書

出現語＼共起語	---	a: 小説	b: 雑誌	---
---	---	---	---	---
A: 本棚	---	3	80	---
B: 書棚	---	73	5	---
---	---	---	---	---

図 1　共 起 行 列

表 1　分類語彙表の項目

1	レコード ID 番号	9	段落番号
2	見出し番号	10	小段落番号
3	レコード種別	11	語番号
4	類	12	見出し
5	部門	13	見出し本体
6	中項目	14	読み
7	分類項目	15	逆読み
8	分類番号		

図 1 で具体的に見てみると，本棚と書棚は，意味の近い単語である．その共起行列を見たときに，本棚は雑誌の出現頻度が高く，書棚は小説の出現頻度が高いと，それぞれの特徴ベクトルは，離れた状態となる．これを小説と雑誌の分類語である「図書」にまとめることができれば，それぞれの特徴ベクトルの向きは離れずに済む．

本研究では，意味の似ている語をまとめると共起ベクトルの距離は近くなるという仮定を前提に，単語間の共起頻度を用いるのではなく，単語に付随する意味属性を利用する．単語の意味属性には，単語を意味によって分類整理したシソーラスである分類語彙表を利用し分類語に適用する．分類語彙表を構成する項目は，表 1 のようになっており，共起行列に用いる意味属性には，その中の「分類項目」を用いた．共起行列の 1 列目には，形態素解析の結果得られた単語のうち，名詞のみを取り出し入力し，数字の部分は，1 文中に共起する頻度をカウントした数が入った行列となっている．また，1 行目には，意味属性として分類語彙表の分類項目の語を入れる [13].

このようにして得られた共起行列は，式 (3) に相当し，単語間の共起行列である式 (2) から式 (3) を導き出す作業は，次式で表される変換行列 K を求めることに等しい [14].

$$F_p = F_w K \tag{4}$$

ただし，K は，N 行 m 列の行列である．

4. 機 械 学 習

並列分散環境での機械学習が数多く開発されており，並列分散処理システムがオープンソースで提供され，パソコンなどの安価なハードを用いることで，大規模データの分析が可能になりつつある．ここでは，バッチ処理方式とオンライン方式の代表的な機械学習である Mahout と Jubatus について説明する．

4.1 Mahout

蓄えておいた全データをまとめて学習することをバッチ処理といい，その機械学習の 1 つに Mahout がある [15]．Mahout は，Hadoop [16] ベースの機械学習プラットフォームとして開発された．Mahout で実装されるアルゴリズムの多くは，Hadoop の MapReduce を利用しており，それは，複数のマシンに対し

て部分データ集合の処理を割り当てる Map とその結果を出力する Reduce に分かれるフレームワークである．

4.2 Jubatus

バッチ処理は，蓄積されたデータに対する機械学習であり，一方，リアルタイムで学習・予測を行う機械学習が開発されている．ストリーム型のデータに対応可能な新たな機械学習フレームワークとして，Jubatus が提供されている [5]．Jubatus の処理は，Update, Analyze, Mix の 3 段階に分かれる．Update は，学習処理に相当し，Analyze は予測処理，Mix は，全マシンからローカルモデルの重みを集め，その平均を取る処理を行う [17]．Jubatus は，機械学習やデータマイニングによるデータ分析に特化した大規模データ処理基盤であり，ビッグデータ解析に必要なリアルタイム・ストリーム処理，分散並列処理，機械学習やマイニングなどの深い分析といった特徴を持つ [18]．

本報告では，将来的な大規模データのリアルタイムでのデータ処理を考慮し，Jubatus を使用することとした．Jubatus 上でいくつかの学習アルゴリズムを動かすことによりテキスト分類を行う．文書分類に必要な操作は，多クラス分類であり，線形識別器を用いて，これを実現する．学習アルゴリズムとして，次のアルゴリズムを用いる．

(i) Perceptron [19]
(ii) Passive Aggressive (PA) [20]
(iii) Confidence Weighted Learning (CW) [21]
(iv) Adaptive Regularization Of Weight vectors (AROW) [22]

(i) の Perceptron は，Rosenblatt により提案されたニューラルネットワークアーキテクチャで，学習させたいデータを入力したときに，出力が間違っていた場合に重みを更新する．(ii) の PA は，カテゴリ分類に使われることの多いアルゴリズムで，新しいデータの損失がゼロであるように重みを更新する．(iii) の CW は，オンライン分類学習で注目されているアルゴリズムであり，自然言語処理などさまざまな分野で応用が示されている．このアルゴリズムでは，パラメータへの信頼度を導入し，重みベクトルの更新において，信頼度の低いパラメータを大きく更新し，信頼度の高いパラメータはあまり変化させない．(iv) の AROW は，CW の制約条件を緩和し，訓練例を正しく分類することを重要としていた CW の問題点を改良したアルゴリズムである．4 つの学習アルゴリズムは，重みの更新において，表 2 のような特徴を持つ．

表 2　学習アルゴリズムの特徴

アルゴリズム	特徴
Perceptron	・ 分類器で正しく分類出来なかった場合，重みを更新する
PA	・ Perceptron よりも学習効率が高い ・ 学習データが正しく分類できたら，重みを更新しない
CW	・ Perceptron, PA と比べて学習効率は高いが，計算量は大きい ・ 出現頻度を考慮し，重みベクトルにガウス分布を導入して更新する
AROW	・ 学習データにノイズが含まれる場合に他の手法と比べ優れた学習効率を示す ・ CW と同様の手法を実現しつつ，複数の条件を同時に考慮しながら最適化を行う

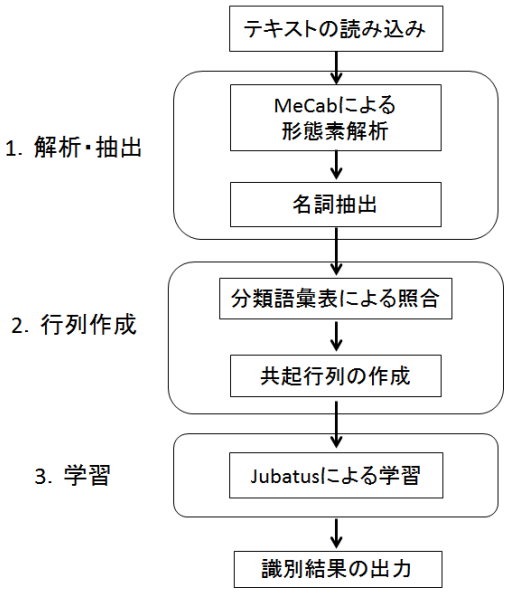

図 2　処理の流れ

今回我々は，上記 4 つの学習アルゴリズムを用いて文書分類を試みた．共起行列の生成と Jubatus の学習器による文書識別までの処理の流れを図 2 に示す．

5. 実　　験

実験では，ニュース記事のカテゴリ分類を行った．毎日新聞社のサイト [23] より記事を収集し，それを，政治，経済，社会，スポーツ，エンターテイメントの 5 つのカテゴリに分類する．分類に用いた記事の数は，1,500 である．収集した記事を，MeCab [24] を用いて形態素解析し，その中から名詞の単語を共起行列作成のための出現単語として用いる．抽出された名詞の数は，重複を除いて 16,999 個であった．単語同士による共起行列では，この出現語を用いて，共起行列を生成していた．ここでは，シソーラスの分類語彙表を用いて共起行列を生成した．そのため，共起行列の列の数は，510 個と大幅に削減された．

生成された共起行列で，学習器に Jubatus を用いた実験を行った．実験環境の OS は CentOS 6.5，Jubatus のバージョンは 0.6.0 である．Jubatus の学習アルゴリズムとして，Perceptron，Passive Aggressive (PA)，Confidence Weighted Learning (CW)，Adaptive Regularization Of Weight vectors (AROW) の 4 つの学習アルゴリズムを用いて文書識別を行った．各学習アルゴリズムごとの識別結果を表 3 と図 3 に示す．4 つの学習アルゴリズムの中では，AROW による識別率が最

表 3　学習アルゴリズムによる比較

	Perceptron	PA	CW	AROW
正識別率 (%)	82.6	84.8	84.2	85.2

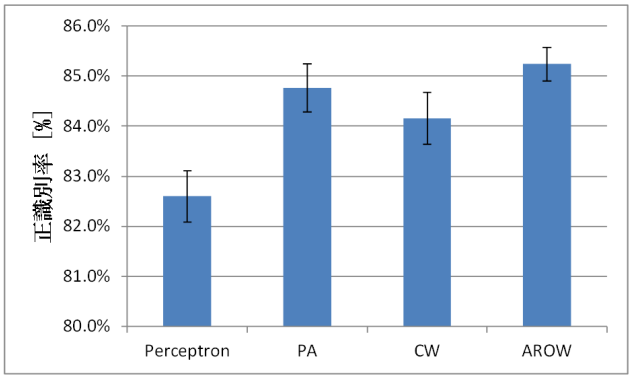

図 3　各学習アルゴリズムの識別率

も高くなったが，PA，CW，AROW の 3 つのアルゴリズムに有意差は見られなかった．ただし，3 手法と Perceptron では，$p < 0.05$ で有意差が見られた．

6. おわりに

本研究では，分類語彙表を用いて単語の意味属性に基づき，共起行列を生成する方法について述べた．これにより，得られた共起行列を学習データとして学習器に与えた．学習器には，機械学習フレームワークの Jubatus を使用した．Jubatus 上で，4 つの学習アルゴリズムを用いて文書分類を行い，識別率の比較を行った．実験の結果，AROW を用いた場合に最も高い識別率が得られたが，PA，CW，AROW では，識別率の結果に有意差は見られなかった．3 手法と Perceptron については，有意差が見られた．

今後の課題として，Jubatus 上で他の学習アルゴリズムを用いた場合や Jubatus 以外の学習器を用いた場合などの識別を行う予定である．

謝　辞

本研究の一部は，JSPS 科研費 (24500120) の助成を受けて実施された．ここに記して謝意を表す．

文　献

[1] R. M. Samer Hassan and C. Banea: “Random-walk term weighting for improved text classification”, Proc. of the First Workshop on Graph Based Methods for Natural Language Processing (2006).
[2] F. Sebastiani: “Machine learning in automated text categorization”, Proc. ACM Computing Surveys, **34**, 1, pp. 1–47 (2002).
[3] 国立国語研究所：“分類語彙表－増補改訂版”, 大日本図書刊 (2004).
[4] 有村博紀：“テキストマイニング基盤技術”, 人工知能誌, **16**, 2, pp. 201–211 (2001).
[5] Jubatus, http://jubat.us/ja/.
[6] 渡部広一, 奥村紀之, 河岡司：“概念の意味属性と共起情報を用いた関連度計算方式”, 自然言語処理, **13**, 1, pp. 53–74 (2006).
[7] M. Pennacchiotti and A.-M. Popescu: “Democrats, republicans and starbucks afficionados: user classification in twitter”, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data miningACM, pp. 430–438 (2011).
[8] O. Alonso, J. Strötgen, R. A. Baeza-Yates and M. Gertz: “Temporal information retrieval: Challenges and opportunities.”, TWAW, **11**, pp. 1–8 (2011).
[9] D. B. D. Wei Wang and X. Lin: “Term graph model for text classification”, Springer-Verlag Berlin Heidelberg 2005, pp. 19–30 (2005).
[10] 片岡 良治：“単語と意味属性との共起に基づく概念ベクトル生成手法”, 人工知能学会第 20 会全国大会論文集, **3C3-1**, pp. 1–3 (2006).
[11] 別所克人, 内山俊郎, 内山匡, 片岡良治, 奥雅博：“単語・意味属性間共起に基づくコーパス概念ベースの生成方式”, 情報処理学会論文誌, **49**, 12, pp. 3997–4006 (2008).
[12] 笠原要, 松澤和光, 石川勉：“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, **38**, 7, pp. 1272–1283 (1997).
[13] 尾脇拓朗, 福元伸也：“単語の意味を考慮した共起ベクトルによるテキスト分類”, DEIM Forum 2014, **C6-2**, (2014).
[14] 笠原要, 稲子希望, 加藤恒昭：“単語の属性空間の表現方法”, 人工知能学会論文誌, **17**, pp. 539–547 (2002).
(1996).
[15] Mahout, http://mahout.apache.org/.
[16] Hadoop, http://hadoop.apache.org/.
[17] 比戸将平：“並列分散環境における機械学習技術の最新動向”, 信学誌, **98**, pp. 54–58 (2015).
[18] 岡野原大輔：“大規模データ分析基盤 jubatus によるリアルタイム機械学習”, 人工知能学会誌, **28**, 1, pp. 98–103 (2013).
[19] F. Rosenblatt: “The perception: a probabilistic model for information storage and organization in the brain”, Neurocomputing: foundations of research MIT Press, pp. 89–114 (1988).
[20] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz and Y. Singer: “Online passive-aggressive algorithms”, The Journal of Machine Learning Research, **7**, pp. 551–585 (2006).
[21] M. Dredze, K. Crammer and F. Pereira: “Confidence-weighted linear classification”, Proceedings of the 25th international conference on Machine learningACM, pp. 264–271 (2008).
[22] K. Crammer, A. Kulesza and M. Dredze: “Adaptive regularization of weight vectors”, Advances in Neural Information Processing Systems, pp. 414–422 (2009).
[23] 毎日新聞, http://mainichi.jp/.
[24] MeCab, http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html.