共起情報を利用した文書の自動分類について

4K－10

藤井洋一、鈴木克志、今村誠、高山泰博

三菱電機株式会社　情報技術総合研究所　音声・言語インタフェース技術部

1.はじめに

近年大量のテキスト情報がインターネットなどを通じてアクセス可能となるにつれて、文書の自動ファイリングやユーザの要求を満足させる自動配信のための文書自動分類技術への要求が高まっている。

そのための技術として、ベクトル空間モデルがある。ベクトル空間モデルによる自動分類は自動学習可能であり、大量かつ幅広い分野のテキストデータを扱う場合に適している。しかし、分類精度が良くないため、精度向上の様々な手法が取り入れられている。例えば、河合[1]は単語の意味属性を用いて分類精度が向上することを示し、湯浅[2]は単語共起による単語のベクトルを利用した分類方法を提案した。

我々は、[1]で指摘されているように単語の多義性を解消すれば、精度が向上するであろうと予測した。ただし、従来の言語学で言われる「多義語」を多義解消するのではなく、複数分類項目で頻繁に出現する単語（例えば「首相」という単語は「政治」や「経済」といった複数の分類項目で頻繁に出現する）を「分類多義語」と定義し、「分類多義語」の多義性解消を試みた。すなわち、「分類多義語」に対して分類項目別に学習情報の頻度と重み付けを再学習し、分類対象文書の「分類多義語」の頻度を共起情報を利用して分類項目別に計算して分類先を決定した。

2.従来の自動分類精度について

従来の分類精度確認のため、「朝日新聞」記事(朝日新聞社提供)[3]を使い河合[1]の方法で精度と分類項目数の関係を確認した。

2.1.河合の方法

河合の方法は、学習セットから分類先の頻度を学習し、χ^2統計を応用した方式で重み付けしている。

分類項目を$C_i\,(i=1,\ldots,N)$、単語を$w_j\,(j=1,\ldots,M)$、出現頻度をF_{ij}とすると、理論頻度を(1)、重み付けを(2)で計算する。

$$M_{ij}=\sum_{i=1}^{N}F_{ij}\cdot\sum_{j=1}^{M}F_{ij}\Big/\sum_{j=1}^{M}\Big(\sum_{i=1}^{N}F_{ij}\Big) \qquad (1)$$

$$Y_{ij}=(F_{ij}-M_{ij})\cdot\left|F_{ij}-M_{ij}\right|\Big/M_{ij} \qquad (2)$$

文書の単語出現頻度を$D=(Dw_1,\ldots,Dw_M)$とし、$S_j=(Y_{1j},\ldots,Y_{Nj})$とした時に、

$$S=\sum_{j=1}^{M}(S_j\cdot Dw_j)=(s_1,\ldots,s_N) \qquad (3)$$

に対して、$S/(\sum_{i=1}^{N}s_i)$を類似度ベクトルとし、ベクトル値の大きい分類項目を分類先とする。

2.2.実験方法

全文書数(各200, 1000, 10000)から75%を学習セット、残りの25%をテストセットとしてランダムに抽出し実験した。形態素解析には、JUMAN 2.0[4]+EDR日本語単語辞書1.5版[5]を利用した。

また、新聞記事[3]の分類は4階層(大分類、中分類、小分類、小小分類)で、かつ実際の新聞記事に付与されている分類は極力小分類および、小小分類であることから、小分類、小小分類を大分類、中分類に変換したものを分類先の個数に合わせて、大分類、中分類の出現頻度に変換し、複数分類先があるものは1文書の学習データを分類先の数で、割ったものを出現頻度として各分類項目に学習させた。(表1)

表1　分類先マッピング方法の例

分類項目		大分類	頻度にかける割合
052	図書	0	0.67
054	読書		
412	日本文学	4	0.33

処理に用いる単語は名詞および、固有名詞、未知語、サ変名詞とした。ただし、記号のみからなる未知語、1文字の単語は削除した。文書セットは、[1]とほぼ同じになるよう、文書サイズが305bytesから1010bytesの記事を古い順番に取り出した。

評価は、分類先類似度を閾値として再現率と適合率を計算し、再現率と適合率が等しい点を分類精度として比較した。

An Automatic Document Classification Using Lexical Co-occurrences
Youichi Fujii, Katsushi Suzuki, Makoto Imamura, Hiroyasu Takayama
Human Media Technology Dept. Information Technology R&D Center, Mitsubishi Electric Corp.
5-1-1 Ofuna, Kamakura, Kanagawa 247, Japan

2.3.実験結果

2.2.に従って実験した結果を表2に示す。

表2　従来の手法による分類精度

分類（分類項目数）	記事数		
	200	1000	10000
大分類(10)	42.0%	63.5%	69.0%
中分類(92)	—	50.0%	58.5%
小分類(735)	—	35.0%	45.0%

ただし、小分類は小小分類を含む

このように、分類精度と、学習文書数および分類項目数の関係は大きく依存していることがわかる。従って、共起情報を用いた分類多義語解消による効果を、特に分類項目数が多い場合(記事数10000、小分類)で検討した。

3.分類多義語の解消による効果について

分類多義語の選定方法を以下のように定め、理想的な解消がされた場合を想定して実験を行った。

3.1.分類多義語の決定方法

分類多義語の選定方法として、上記従来の方法にて学習した重み付け情報の中で、以下の(4)の条件を満たす単語 $w_j\,(j=1,\ldots,M)$ を分類多義語とした。

$$\#\{Y_{ij} \mid \max_{1\le k\le N}(Y_{kj})\cdot V \le Y_{ij}\} \ge 2 \quad (V:\text{閾値}) \qquad (4)$$

3.2.理想実験の方法と結果

学習セットに対しては、表3の様に頻度分配し、式(1)、(2)によって重み付けを再計算する。一方、実験セットに関してもあらかじめ分類先が分かっているので、実験セットの文書に対して理想的に分類多義が解消された場合を同様に計算し、その解消された分類多義の頻度を使って実験セットを分類した。

表3　理想的頻度分配の例(単語「随筆」)

分類多義語 (頻度5)	正解分類先			分配後の頻度
	図書	読書	日本文学	
随筆(外国文学)	0.83	0.83	0.0	1.66
随筆(日本文学)	0.83	0.83	1.67	3.33

結果は表4の通りである。

表4　理想条件での分類精度

閾値(V)	(なし)	0.9	0.7	0.5	0.3	0.1
精度(%)	45.0	45.1	45.8	47.4	51.4	61.7
分類多義語数	—	2229	7176	12399	18099	24637

4.共起を用いた自動分類について

3.の結果に基づき、閾値 V=0.1 の場合について共起を用いた分類多義の解消実験を行った。

4.1.共起による頻度の補正

共起による頻度補正は図1のステップで行った。

- 分類先文書の単語の出現した段落から共起単語ベクトルを取り出す。(I)
- 分類文書も共起単語ベクトルを取り出す。(II)
- (I)，(II)の内積をとり、頻度を比例分配する。

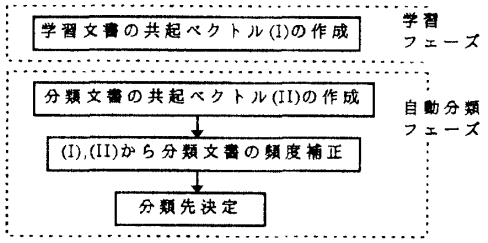

図1　共起による分類方法

4.2.実験結果

共起による頻度補正の方法として表5の2つの場合を実験した。

表5　共起による頻度補正での分類精度

頻度補正方法	精度(%)
(a)出現頻度で内積計算し分配した場合	46.4
(b)出現の有無のみで内積計算し分配した場合	47.0

(a)は共起ベクトルを頻度で作成し内積をとることで頻度分配した場合である。(b)は、共起ベクトルを単語の出現の有無のみの0,1ベクトルで作成し内積をとることで頻度分配した場合である。

5.考察

4.の実験により、単語の共起情報利用によって精度が45.0%から多少ではあるが分類精度が向上することがわかった。しかし、上記の共起による分類多義の解消は、単純に共起単語の頻度ベクトルおよび、出現の有無によるベクトルの内積で行っているため、解消後の頻度にあまり差が現れなかったと考える。

6.まとめ

今後は、学習文書の共起ベクトルがより特徴量を表現する方法を検討するとともに、大量データによる検証を行いたい。

[参考文献]
[1] 河合:意味属性の学習結果に基づく文書自動分類方式，情報処理学会論文誌，Vol.33，No.9(1992).
[2] 湯浅：大量文書データ中の単語共起を利用した文書分類，情報処理学会論文誌，Vol.36，No.8(1995).
[3] 朝日新聞記事データベース(1991年9月～1992年8月).
[4] 松本他：日本語形態素解析システムJUMAN使用説明書 version 2.0(1994).
[5] EDR電子化辞書 日本語単語辞書1.5版，（株）日本電子化辞書研究所(1995).