文書の意味空間へのマッピング*

1T－8

宮崎 哲夫† 田中 栄治‡ 古城 則道§
学習情報通信システム研究所¶

1 はじめに

最近、パソコンやネットワークの普及により、電子化された情報が大量に出回ってきている。現状ではこれらの情報は整理されているとは言い難い状態にある。このような未整理な情報源から何かの知識を取得したい場合、見当違いの方面を探索してしまい、目的である知識になかなか到達することができないことが多い。そこで、電子化された情報の効率的な検索支援のために、何らかの方法による整理・分類機能が必要になる。

文書の自動分類には、文書に出現する単語パターンの類似性に基づく方法がある。通常は単語間の関係を考慮せずに、単語を含むか含まないかなどの情報のみで分類していることが多い。

そこで本稿では、単語間の関連を考慮した文書分類のために、単語の共起関係データに対する主成分分析に基づく意味空間の生成、および、文書を意味空間へマッピングする方法について述べる。

2 意味空間

単語の出現の有無を要素とした文書ベクトルの類似性に基づく文書分類では、文書の特徴がそこに含まれる単語集合で表現されると仮定されている。しかし、この方法では、単語が意味的に独立であることを暗黙の前提とし、単語間の関係は考慮されていない。通常の文書においては、それぞれの単語に冗長性や意味的なノイズが含まれる。したがって、単語の出現の有無だけに基づく文書分類では、高い精度の分類は期待できないと考えられる。そこで、単語間の関係を取り入れるために、サンプル文書集合の単語出現状況について主成分分析を行ない、意味的要素を抽出する。この意味的要素が作る空間を意味空間と呼び、意味空間にマッピングされた文書ベクトルが文書意味ベクトルである。文書分類は文書意味ベクトルの類似性を基に行なわれる。

*A Method of Mapping Documents to Semantic Space
†Tetsuo Miyazaki
‡Eiji Tnaka
§Norimichi Kojo
¶Software Research Laboratory

3 マッピングアルゴリズム

文書ベクトルを意味空間へマッピングするアルゴリズムは以下のとおりである。(図1参照)

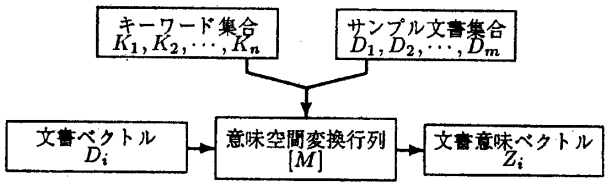

図 1: 意味空間生成アルゴリズム

(1) 対象分野においてキーワードとなりうる i 個の単語 $(K_1, K_2, \cdots, K_n)$ を定義する。

(2) 単語の意味的要素を抽出するために、m 個のサンプル文書 $(D_1, D_2, \cdots, D_m)$ を用意する。

(3) 用意されたサンプル文書について単語の出現状況を調べ、各文書の文書ベクトル $\vec{D_i}$

$$\vec{D_i} = (d_{i1}, d_{i2}, \cdots, d_{in})$$

を作成する。ただし、

$$d_{ij} = \begin{cases} 0 & (\text{キーワード } K_j \text{を含まない}) \\ 1 & (\text{キーワード } K_j \text{を含む}) \end{cases}$$

である。

(4) サンプル文書の文書ベクトルの要素データに対して主成分分析を適用し、主成分を軸とする意味空間を作成する。このとき、第 i 主成分 Y_i は元の基底 $\vec{k_1}$、$\vec{k_2}$、$\cdots$、$\vec{k_n}$ を用い、

$$\vec{Y_i} = y_{i1}\vec{k_1} + y_{i2}\vec{k_2} + \cdots + y_{in}\vec{k_n}$$

のように表現される。ここで、係数 y_{ij} は主成分 Y_i に対するキーワード K_j の寄与率を表す。

(5) l 番目の意味キーワードに対応する第 l 主成分までを使うと、意味空間への変換行列

$$[Y] = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{l1} & y_{l2} & \cdots & y_{ln} \end{bmatrix}$$

が得られる。ここで、主成分の個数 l は、主成分の

累積寄与率で決まる。

しかし、この変換行列を用いて意味空間へのマッピングを行なうと、意味空間の各軸のスケールは一定ではなく、サンプル文書集合の分散の大きい軸方向に広がってしまう。この意味的異方性を除去するために、各主成分の要素を、それぞれの分散の平方根で割ったものを、意味空間への変換行列 $[M]$ とする。すなわち、

$$[M] = \begin{bmatrix} \frac{y_{11}}{\sqrt{\lambda_1}} & \frac{y_{12}}{\sqrt{\lambda_1}} & \cdots & \frac{y_{1n}}{\sqrt{\lambda_1}} \\ \frac{y_{21}}{\sqrt{\lambda_2}} & \frac{y_{22}}{\sqrt{\lambda_2}} & \cdots & \frac{y_{2n}}{\sqrt{\lambda_2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{y_{l1}}{\sqrt{\lambda_l}} & \frac{y_{l2}}{\sqrt{\lambda_l}} & \cdots & \frac{y_{ln}}{\sqrt{\lambda_l}} \end{bmatrix}$$

とする。ここで、λ_iは第i主成分の分散の値である。

(6) サンプル文書の文書ベクトルの作成と同様に、分類する文書について単語の出現状況を調べ、各文書の文書ベクトル

$$\vec{D_i} = (d_{i1}, d_{i2}, \cdots, d_{im})$$

を作成する。

(7) 文書ベクトル$\vec{D_i}$を変換行列 $[M]$ を使って文書意味ベクトルを$\vec{Z_i}$に変換する。すなわち、

$$\vec{Z_i^T} = [M]\ \vec{D_i^T}$$

となる。

4 評価実験

意味空間の有効性を確認するため、文書の文書意味ベクトル表現を基に類似度を計算し、クラスタ分析による文書分類を試みた。

実験に用いる文書は、ネットニュース (fj.ibm.sys) の投稿記事を利用した。まず、意味空間を生成するために、キーワードには、fj.ibm.sys のニュースグループに関連のありそうな単語を、予め人手によって 200 単語選定した。サンプル文書は、fj.sys.ibm のニュースグループから 150 文書を無作為に抽出した。

自動分類の評価に使うテスト文書はサンプル文書とは別に、fj.sys.ibm のニュースグループから 3 種類の話題 (サブジェクト A、B、C) に関する 26 文書を選び出した。3 つのグループに属する文書番号で表現すると

A = { 1 2 3 4 5 6 7 8 }
B = { 9 10 11 12 13 14 15 16 17 18 }
C = { 19 20 21 22 23 24 25 26 }

となっている。

これらの文書についてキーワードの出現の有無を調べ、文書ベクトルを作成したのち意味空間へマッピングする。マッピングされた文書意味ベクトルの類似度を計算しクラスタ分析を行う。第 60 主成分までによる意味空間を用いた解析結果を樹形図で表すと図 2 のようになり、サブジェクトによって分類された 3 つのグループ構成を再現していることがわかる。

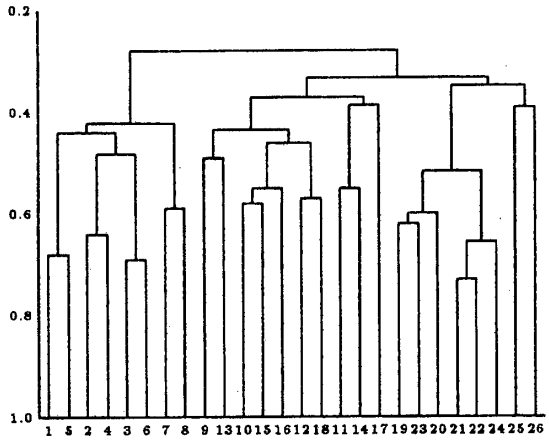

図 2: 文書自動分類のクラスタ分析樹形図

5 おわりに

本稿では、文書を自動分類するときに、単語の表層的な出現パターンだけではなく、単語間の関係も考慮する手法を検討した。すなわち、サンプル文書を用いて単語の共起関係から意味空間を作成し、各文書をその意味空間における表現を基に分類する手法である。

ネットニュースの記事を用いた評価実験の結果から、意味空間を導入することにより、予め用意した 200 キーワードに対して、60 次元程度まで縮約可能であることを確認した。

今後の課題として、サンプル文書からのキーワードの自動抽出、教示による変換行列のチューニング機構の検討を考えている。

参考文献

[1] 湯浅 夏樹, 上田 徹, 外川 文雄 : 大量文書データ中の単語間共起を利用した文書分類, 情報処理学会論文誌,Vol.36,No.8,pp.1819-1827

[2] 本間 直人, 石川 眞澄 : 情報検索におけるキーワードと文献の空間表現, 信学技報,NC95-143 (1995-03),pp.211-218

[3] Scott Deerwester, Susan T.Dumais, Richard Harshman : Indexing by Latent Semantic Analysis, Journal of the american society for infomation science, 41(6):391-407,1990