2002-15-01

文書の自動クラスタリング手法の提案と開発

80122329 鈴木 直明　　　指導教員 青山 藤詞郎 教授，矢向 高弘 専任講師

1　緒論

近年ではコンピュータを使った文書作成が一般的になっており，大量の文書を整理して利用者の目的とする情報を迅速に検索できることが望まれている．本論文では，人手によらずにコンピュータを利用して文書を自動的に分類する汎用的な手法の提案を目的とし，精度の低下要因として正規化による文書ベクトルの均質化に注目した．提案手法の特徴は，文書ベクトルの正規化を行わず，ノルムに応じた重みを与えて重心ベクトルを計算することで文書ベクトルを差別して扱っていることである．

2　提案手法とその評価

本論文の提案手法は，E.Han と G.Karypis による重心ベースの分類手法 [1] を基礎としている．この既存の手法はベクトルで文書を表現し，分類分野をその分野に含まれる全文書ベクトルの重心で表現する．未分類の文書は，文書ベクトルと分野ベクトルの余弦を類似度として計算し，最も類似度の高い分野に分類する．

この手法は分類分野に関する訓練データを必要とするため，大規模な文書分類には適さない．そこで本論文ではこの手法を訓練データを使わないクラスタリング手法に応用し，その手法をもとに分類実験を行った．実験内容は，5080 件の新聞記事 [2] [3] を文書ベクトルで表現し，文書ベクトル間の類似度が高いものから階層的にクラスタリングするものである．

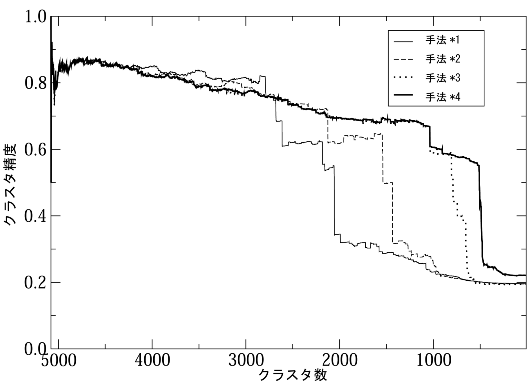

図 1: クラスタリングの結果

図１に４種類の手法による実験結果を示す．横軸はクラスタリングの推移，縦軸は横軸の時点でのクラスタ精度の平均値を示している．クラスタ精度はクラスタ内のすべての文書の組み合わせの内，UDC 分類コード (3 桁)[4] が一致する文書の組の割合である．

図中の*1 は既存の手法でクラスタリングを行った結果である．それに対して*2 は文書ベクトルを正規化しないでクラスタリングする提案手法による結果であり，評価値が高いままの状態で推移している．その要因としては重心ベクトルの角度がノルムの大きな文書ベクトルの影響を大きく受けていることが考えられる．この点に着目して*2 の手法に改善を施した結果が*3 である．*3 の手法では χ^2 値を文書ベクトルの重量として重心計算の際の重みに利用し，重心ベクトルの位置決定に反映させている．*4 は*3 の手法に加えて，特定のクラスタのサイズが特異的に大きくなる現象を抑制する規則を導入した改善手法である．この手法ではクラスタの重心ベクトルとそのクラスタに含まれる文書ベクトルとの間の類似度の平均値が最大になるクラスタをクラスタリング対象にしない規則を導入している．

以上の 4 種類の実験において，推移していくクラスタ精度の平均値の値をすべて足し合わせ，足し合わせた数で平均をとった全体的な評価値を表 1 に示す．本研究ではこの評価値において，手法*1 に比較して最大 22.3%の向上を実現することができた．

表 1: クラスタリング手法の比較

	手法*1	手法*2	手法*3	手法*4
評価値	0.570	0.617	0.676	0.697
向上率	—%	8.3%	18.7%	22.3%

3　結言

正規化による文書ベクトルの均質化に注目し，正規化を行わず，文書ベクトルのノルムに応じた重みを与えて重心ベクトルを計算する提案手法によってクラスタ精度を向上させることができた．今後の課題は他の文書集合に対して検証実験を行うことである．

参考文献

[1] Eui-Hong(Sam)Han, George Karypis: "*Centroid-Based Document Classification: Analysis & Experimental Results*", PKDD 2000, pp.424-431, 2000

[2] 木本強 他: 日本語情報検索システム評価用テストコレクション BMIR-J2, 情報処理学会研究報告 データベースシステム, No.114-3, pp.15-22, 1997

[3] CD-毎日新聞'94 データ集, 毎日新聞社

[4] RWC-DB-TEXT-95-3 http://www.rwcp.or.jp/wswg/rwcdb/text/