大量文書向けのクラスタリング手法の評価

4Q－3

青木 圭子　　松本 一則　　橋本 和夫
国際電信電話株式会社 研究所

1. はじめに

近年，電子化文書の流通が増大し，大量の文書情報の中から必要なものを検索する必要性が増してきており，類似性を基準に大量の文書をクラスタリングする技術が重要となってきた．既に，文書中の語の出現確率を用い，文書集合をベイジアンクラスタリングする手法[1]が提案されているが，同手法の場合，生成中の全クラスタ対においてクラスタどうしを一旦マージする必要があるため，大量の文書集合を処理することが困難になる．そこで筆者らは，計算量を削減することを目的とした大量文書向けのクラスタリング手法[2]を提案・実装して，提案手法と従来手法によるマージ回数の比較を行った．

本稿では，提案手法のマージ回数の推定方法について考察し，実測値との比較結果を報告する．

2. 提案するクラスタリング手法

ここでは，提案するクラスタリング手法の処理手順を説明する．そして，同手法で最適な文書集合を求める際に使用する遺伝的アルゴリズム[3]について述べる．

2.1 処理手順

(1) 全文書をルートクラスタ (C_{root}) に割り当てる．

(2) C_{root}を“処理待ちクラスタ”キュー Q に登録する．

(3) Q が空になるまで以下の処理を行う．

(i) Q の先頭のクラスタを C_p とする．

(ii) C_pに割り当てられた文書 {D} の中から最適と思われる M 個の文書 $\{D_d\}$ を抽出する (図 1)．最適化は MDL 基準に基づき，分類結果の符号長が最小になるようにする．そして，解の探索には遺伝的アルゴリズム (以下，GA) を用いる．(2.2)

(iii) C_p に割り当てられた文書数が M 以下の場合，その全てをクラスタ化する．文書数が M 以上の場合，選ばれた M 個の文書をクラスタ化する．クラスタ化の方法は以下の通りである．このクラスタを C_dとする．

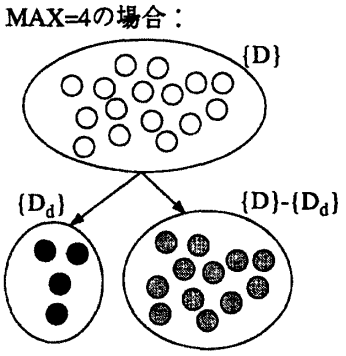

図 1: M 個の最適な文書を抽出

(a) C_dにおいて，すべての組み合わせ ($_MC_2$通り) についてクラスタ C_aと $C_b(C_a, C_b \in C_d)$ を一旦マージし，クラスタ $C_m = C_a \cup C_b$を生成する．C_mの語分布のもとで C_aと C_bの語分布が起こる事後確率 $P(C_m \mid C_a, C_b)$ を求め，この値が最大となるクラスタペアー (C_a, C_b) からマージされたクラスタ C_mを採用する．

(b) C_a, C_bを子とし，C_mを親とするツリーを作成する．

(c) $C_d = C_d - (C_a, C_b) + C_m$とし，$C_d$が C_m に等しくなるまで (3)(iii)(a) 以下を繰り返す (図 2)．

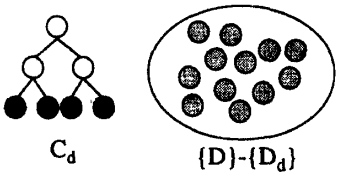

図 2: クラスタリング

(d) 残りの文書 ($\{D\} - \{D_d\}$) を最も距離の近いリーフクラスタ ($C_i \in C_d$) に割り当てる．クラスタ C_iに割り当てられた文書群を $\{D_i\}$ とする (図 3)．

(e) $|D_i|$($\{D_i\}$ の要素数)> 0 となったクラスタ C_iについて，クラスタ C_iを Q に追加する．

A Clustering Algorithm Suited for Large Document Set
Keiko AOKI, Kazunori MATSUMOTO, Kazuo HASHIMOTO
KDD R&D Laboratories
2-1-15 Ohara, Kamifukuoka, Saitama 356, Japan

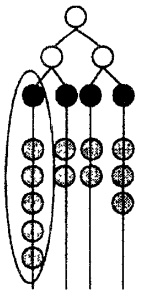

図 3: 残り文書の割り当て

2.2 最適文書集合を求めるための GA

特定の初期値から始めて探索を行う方法では，通常，数多くの異なる初期値で探索を繰り返す必要がある．本稿の最適な文書集合を求める問題の場合，次の探索のための適切な初期値を定める方法がないため，初期値をランダムに選んで次の探索を再開しなければならない．このため，最初に多点サンプルを行い，探索を並行して行う GA が適していると思われる．

ここでは，次のようなモデルを用いた．

- スケーリング べき乗スケーリング ($f' = f^2$)
- 選択交配 適応度比例戦略及びエリート保存戦略
- 交叉，突然変異 2つの親を掛け合せるのではなく，世代ギャップ数 (N_g) の親のある一定割合のビット分をランダムなビットに置き換える方法をとった．
- 世代モデル 連続世代モデル

2.3 マージ回数の推定

クラスタリング処理の大半はクラスタどうしのマージに費やされるため，処理時間を推定するためにはマージ回数の正確な推定が必要である．

従来手法のマージ回数は ${}_NC_2 + \sum_{k=1}^{N-2} k$ となり，$(N-1)^2$ となる．一方，提案手法のマージ回数は，

- 2.1(3)(iii)(a) におけるマージ回数 $= (M-1)^2$
- マージするドキュメント数 $= c_n$
- 2.1(3)(iii)(d) における 1 ドキュメントあたりのマージ回数 $= \alpha(logM - 1)$
- 評価した遺伝子の数 $= (N_{pg} + N_{pg}R_g(N_g - 1))$

を用いて，

$$((M-1)^2 + 2(c_n - M)(\alpha(logM - 1))) \times (N_{pg} + N_{pg}R_g(N_g - 1))$$

と評価できる．

ただし，$\alpha(> 0)$ はクラスタの形に依存した数で，クラスタのバランスが良いと 1 前後となる．また，

N_{pg}：世代あたりの遺伝子数

R_g：世代ギャップ

N_g：世代数

である．

3. マージ回数測定実験

3.1 実験環境と測定パラメータ

計算機は Sun Netra140E (SunOS 2.5.1, 64MB) を用いた．最適化の際のパラメータは抽出文書数 (M)：16，世代数 (N_g)：5，世代あたりの遺伝子数 (N_{pg})：8，世代ギャップ (R_g)：0.25 とした．

3.2 実験結果

結果を表 1，表 2に示す．表中の語の意味は以下の通りである．

- N：ドキュメント数
- M_c：マージの回数
- M_t：マージに要した時間 (秒)
- P_t：全処理時間 (秒)

表 1: 従来手法

N	M_c	M_t	P_t
100	9,811	357.7	387
200	39,601	3,440.9	3,996

表 2: 提案手法

N	c_n	M_c		M_t	P_t
		最適化処理部分	その他		
100	100	14,182	821	1,845	1,932

$\alpha = 1.3$ と仮定して，2.3の推定方法でマージ回数を推定すると，

$$\{(16-1)^2 + 2 \times (100-16)(1.3 \times 3)\} \times (8 + 8 \times 0.25 \times (5-1)) \approx 14,083$$

となる．

4. おわりに

本稿では，提案手法のマージ回数の推定方法について考察し，実測値との比較結果を報告した．

参考文献

[1] Makoto IWAYAMA, Takenobu TOKUNAGA, "Hierarchical Bayesian Clustering for Automatic Text Classification", Proceedings of IJCAI-95, pp.1322-1327, 1995.

[2] 青木, 松本, 橋本, "類似ドキュメントの発見手法の検討", 情報処理学会第５４回全国大会 (平成９年前期), 3-39, 1997.

[3] 北野宏明, "遺伝的アルゴリズム", 産業図書, 1993.