遺伝子情報の共通部分に注目したGAの高速化手法の検討 ～大量文書のクラスタリングにおいて～

2T－1

青木 圭子　松本 一則　帆足 啓一郎　橋本 和夫
KDD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．以前，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]の計算量を削減するため，部分クラスタの評価にMDL基準を用い，準最適なクラスタを遺伝アルゴリズム(以下，GA)[3]を用いて求めることにより高速化する手法を提案し，評価した[2]．その結果，評価値の計算部分の高速化が必要であることが分かった．

本稿では，遺伝子を評価する際，評価すべき遺伝子と評価済み遺伝子の共通部分に注目し，評価計算処理の重複をなくし，GAの処理を高速化する工夫について述べる．

2. クラスタリング手順の概要

2.1 再帰的クラスタリング

本クラスタリングアルゴリズムでは，トップダウン的に一定量の文書を選択，クラスタリングし，残り文書をそのクラスタの類似するノードに割り当て，ノード毎に割り当てられた文書を再帰的にクラスタリングする．

(1) 対象文書集合 Dの文書数が M個以上であれば (2) 以下を実行し，M個未満であれば，厳密クラスタリング[1]を行う．
(2) M個の最適な文書 Sを選択．
(3) M個の文書を分類し，クラスタを生成する．(図1)
(4) 選択されなかった文書 $D-S$を (3) のクラスタの葉ノードに割り当てる．(図2)
(5) 各葉ノードに割り当てられた文書を Dとして，(1) 以下を繰り返す．

クラスタリングは

- $P(T=t|D)$: 対象文書 D中の単語 t の相対頻度
- $P(T=t|c_i)$: クラスタ c_i中の単語 t の相対頻度

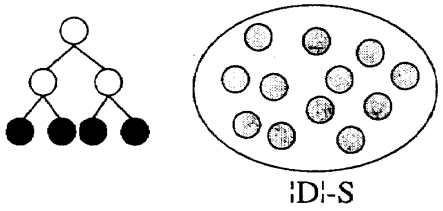

図 1: M 個の文書を選択してクラスタリング

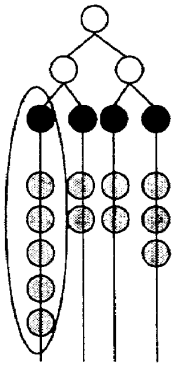

図 2: 残り文書の割り当て

- $P(T=t)$: 全文書中の単語 t の相対頻度
- $P(c_i)$: 全文書中の文書のクラスタ c_iに属する文書の相対頻度

として，全てのノード間の類似度を求め，以下の類似度を最大化するクラスタ対をマージすることにより行う．

$$P(c_i|D) = P(c_i)\sum_t \frac{P(T=t|c_i)P(T=t|D)}{P(T=t)} \quad (1)$$

2.2 文書選択の最適化

上記クラスタリング手順における最適な文書を選択するための評価関数としてMDLを用いる．最適解の探索方法としては，GAを用いる．

本手法の場合，遺伝子型は対象文書集合を表す．遺伝子型の各遺伝子は，選択した文書の遺伝子の値を1，選択されなかった文書の遺伝子の値を0とする．(図3)

次世代の遺伝子型は，1の値を持つ遺伝子の数が変わらない様にして世代ギャップ分の遺伝子の値をランダムに置き換えることにより，決定する．これら全ての遺伝子型に対応する M個の文書をクラスタリングして評価値を求め，評価値が最大となる文書集合を求める．

Fast Genetic Algorithm based on Common Part of Genotypes. -For Clustering of Large Document Set-
Keiko AOKI, Kazunori MATSUMOTO, Kazuo HASHIMOTO
KDD R&D Laboratories Inc.
2-1-15 Ohara, Kamifukuoka, Saitama 356-8502, Japan

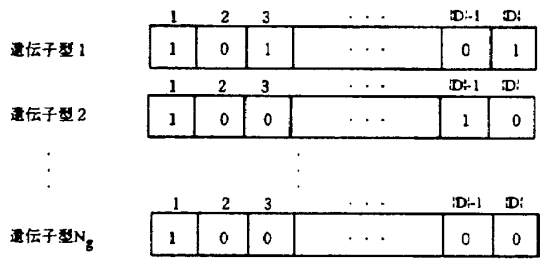

図 3: GA の適用

このため，文書数が増えるとマージ回数が増え，事後確率の計算量が膨大となり，高速化の必要がある．

3. 遺伝子情報の共通部分に注目した GA の高速化

全ての遺伝子型の評価において，全文書間の類似度を求めることにより，クラスタリングを行う必要があるが，ノード対が同じものに対して，類似度の共有化を提案する．例えば，図 4では，1，2，5 の遺伝子が共通であるため，3 つの遺伝子間の類似度の計算を共有化出来る．また，葉ノード間の類似度だけでなく，図 5の様にマージされた中間ノードと他の中間ノードまたは葉ノードとの間の類似度も共有化出来る．

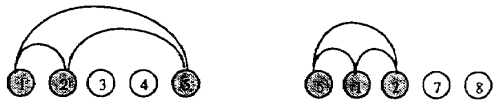

図 4: 重複するノード対

図 5: 重複する中間ノード

4. マージ回数測定実験

これらの遺伝子情報の共通部分が多数存在することを確認するため，重複ノード対の測定を行った．

4.1 実験環境と測定パラメータ

計算機は Sun UltraEnterprise450 (Solaris 2.6, 512MB) を用いた．データには，1993 年〜1997 年に公開された特許文書 250 件を用いた．

パラメータは

N：文書数

M：抽出文書数

L：ぶら下げコンテンツ数

N_g：世代数

N_{pg}：世代あたりの遺伝子型数

R_g：世代ギャップ

である．このとき，遺伝子型の評価回数 E_gは，

$$E_g = N_{pg} + N_{pg} R_g (N_g - 1) \qquad (2)$$

となり，式 (1) の計算回数 M_gは

$$M_g = ({}_M C_2 + \sum_{k=1}^{M-2} k) E_g = (M-1)^2 \cdot E_g \qquad (3)$$

となる．

$N = 250$, $M = 32$, $L = 128$, $R_g = 0.3$ として，N_g, R_gの値を変えて測定を行った．

4.2 実験結果

結果を表 1に示す．

表 1: 共通ノード対の数

N_g	N_{pg}	E_g	M_g	重複ノード対の数
5	10	22	21,142	11,772
10	10	37	35,557	24,427
5	20	44	42,284	19,172

$N_g = 5$, $N_{pg} = 10$ の場合，遺伝子型の評価回数は 22回，マージ回数は 21,142 回であり，重複ノード対の数は 11,772 であった．1 回のマージに要する時間は約 76.4msec である．この場合，約 899〜1,866sec の計算時間が短縮されることが分かった．

5. おわりに

本稿では，提案手法のクラスタリングの処理時間を実験で求めた．最適化における評価値計算の中で，重複するノード対が多数存在することが確認できたため，これらの評価値を共有化することにより，最適化における評価値計算の処理速度の向上が可能と判断した．

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 5 4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.