大量文書向けのクラスタリング手法の評価

3Y－6

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．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)によって決定する手法を提案し[3]，GAによる計算量の削減を報告した[2]．

本稿では，同手法の定量評価に先立ち，クラスタリング精度をタスク達成度により測定し，本手法のパラメータとタスク達成度との関係について報告する．

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符号長及びGAについて述べる．

2.1 処理手順

procedure clustering()
　全文書をルートクラスタ (C_{root}) に割り当てる;
　C_{root}をキュー Q に登録する;
　while (Q が空になる) {
　　C_p = Q の先頭のクラスタ;
　　if (C_p の文書数　MAX 個以下) {
　　　sub_clustering(C_d); /* C_d のクラスタリング */
　　　break ;
　　}
　　D_d = Select(D);
　　/* C_pに割り当てられた文書集合 D の中から最適と
　　思われる MAX 個の文書集合 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 (2.2).
　　解の探索には遺伝的アルゴリズムを
　　用いる (2.3). */
　　sub_clustering(C_d);

/* D_dをクラスタ C_dに割り当て，クラスタ化する */
残りの文書集合 ($D - D_d$) を最も距離の近いリーフ ($L_i \in C_d$) に割り当てる; /* (図 2) */
$D_i = L_i$に割り当てられた文書集合;
$|D_i| > 0$ となった D_iをクラスタ C_iとしてをQに追加する;

procedure sub_clustering(C_d)
　do {
　　$C_m = C_a \cup C_b$; /* ($C_a, C_b \in C_d$) */
　　$P(C_m \mid C_a, C_b)$ が最大となる (C_a, C_b) から C_mを求める.
　　C_a, C_bを子とし，C_mを親とするツリーを作成する.
　　$C_d = C_d - (C_a, C_b) + C_m$;
　} **while** (C_d != C_m)

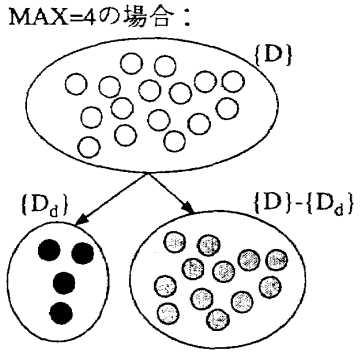

図 1: MAX 個の最適な文書を抽出

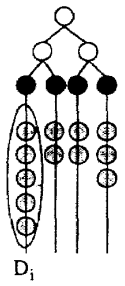

図 2: リーフ L_iに割り当てられた文書集合

An Evaluation of 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

2.2 最適文書集合を求めるための符号長

文書集合Dの中から最適と思われるMAX個の文書集合D_dを抽出する際の基準として，クラスタの符号長を最小にすることを以前提案している[3].

クラスタの符号長は，木の記述自体に必要な情報量と，木が表すモデルの情報量との和として計算する[4][5]. ただし，計算量を削減するため，$D-D_d$中の全文書を用いるのではなく，その中の一定数Lの文書を用いて，木が表わすモデルの情報量を計算する (図3).

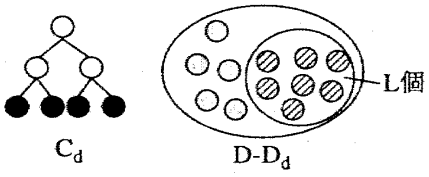

図 3: 残り文書の割り当て

2.3 最適文書集合を求めるためのGA

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，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 (R_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

3. クラスタリング精度の測定

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. データとして，ac.jpドメインのHTMLファイル100個を用いた.

最適化の際のパラメータは

- M：抽出文書数
- L：部分クラスタに割り当てる残り文書数
- N_g：世代数
- N_{pg}：世代あたりの遺伝子数
- R_g：世代ギャップ

とした. $R_g = 0.3$ として，タスク達成率を求めた.

ここで，タスク達成率とは同一文書に類似する文書を検索し，類似度の高いものから上位10文書を見たときに，総当たりでのクラスタリング結果と同じ文書が含まれている割合と定義する.

3.2 実験結果

Mを十分大きくした最適なクラスタで検索した文書集合をリファレンスと見なし，GAを用いて検索した文書集合との共通部分の割合をタスク達成率と考え，類似文書を上位10個もしくは20個出力する場合のタスク達成率 (T_1, T_2) を10回測定した. (表1)

M	L	N_g	N_{pg}	T_1	T_2
∞	—	—	—	1.0	1.0
32	16	5	5	0.39	—
32	16	5	10	0.33	—
32	16	10	5	0.42	—
32	16	10	10	0.35	—
32	32	5	5	0.47	0.54
32	32	5	10	0.39	0.63
32	32	10	5	0.33	0.53
32	32	10	10	0.33	0.62

表 1: タスク達成率

T_1では世代あたりの遺伝子数を増やしたときにタスク達成率が下がる等の問題があった. 但し，T_2では世代数を増やすのはタスク達成率は向上しなかったが，世代あたりの遺伝子数を増やすことによりタスク達成率が向上した.

4. おわりに

本稿では，提案手法のクラスタリングのタスク達成率を実験で求めた. 実験結果は満足なものではなかったが，使用した文書数が少なかったために統計的な性質をうまく利用できなかった可能性がある. また，クラスタリングの精度の評価にタスク達成率を利用したが，本来は木構造の類似度を測定する方が好ましい.

今後は，検証規模の拡大や木構造の類似度を用いた評価を行っていく予定である.

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, "大量文書向けのクラスタリング手法の評価", 情報処理学会第55回全国大会 (平成9年後期), 1997.

[3] 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.

[4] 中茎洋一郎, 古関義幸, 田中みどり, "確率モデルの学習方式と診断への応用", 人口知能 74-3, pp.19-24, 1991.

[5] 伊藤秀一, "MDLのパターン認識への応用", 人口知能学会誌 Vol.7 No.4, pp.608-614, 1992.