文書集合のトップダウンクラスタリングへのMDL基準の適用

2V－2

松本 一則　青木 圭子　帆足 啓一郎　橋本 和夫
KDD研究所

1. はじめに

大量の文書をクラスタリングすることを目的に，対象の文書集合から選びだした一定数の文書から部分クラスタを生成し，残りの文書を葉ノードとの類似性で分類し，分類された各文書集合に対してクラスタ生成処理を繰り返す「トップダウンクラスタリング」が提案されている．文献[1]のトップダウンクラスタリングでは，文書選択の基準は無く，ランダムに選んだ文書から部分クラスタを生成している．一方，筆者らが提案した手法[2]では，いったん部分クラスタリングを複数回試行的に作成し，残りの文書を分類する時のエントロピーを最大化する（葉ノードに分類される残り文書の数ができるだけ均等する）部分クラスタを採用する方法をとっている．

本稿では，まず，エントロピーが大きい部分クラスタリングが検索精度が良いか検証する．また，部分クラスタをMDL基準で評価可能か検討する．

2. トップダウンクラスタリングの概要

ここでは，本稿で使用するトップダウンクラスタリング手法について説明する．

類似尺度

入力文書 d_{test} と文書集合 c_i との類似度 $P(c_i|d_{test})$ は以下の式[3]で計算する．

$$P(c_i|d_{test}) = P(c_i)\sum_t \frac{P(T=t|c_i)P(T=t|d_{test})}{P(T=t)}$$

- $P(T=t|d_{test})$: d_{test} での単語 t の出現確率．
- $P(T=t|c_i)$: c_i での単語 t の出現確率．
- $P(T=t)$: 全検索対象文書中での単語 t の出現確率．
- $P(c_i)$: 検索結果に c_i 中の文書が含まれる確率（クラスタ生成時には，1とする）．

マージするノードの選択基準

マージ前のノードの各文書がマージ後のノードで検索される確率を最大化する．このため，クラスタ生成の各過程でマージすべきノード対 c_x, c_y としては，以下の式を最大化するものを選択する．

$$\frac{\prod_{d_i \in c_x \cup c_y} P(c_x \cup c_y|d_i)}{\prod_{d_i \in c_x} P(c_x|d_i) \prod_{d_i \in c_y} P(c_y|d_i)}$$

部分クラスタ選択基準

葉ノードの数が同じ部分クラスタ c を比較する際，葉ノードで分類された文書のエントロピー $E(c)$ が大きいほど，分類精度が良いと考え，$E(c)$ が最大になる部分クラスタを採用している．

$$E(c) = \sum_i p_i \log p_i$$

ただし，葉ノード i に分類される文書数を n_i とすると，$p_i = n_i / \sum_k n_k$ である．

3. エントロピーと分類精度の関係

エントロピーが高い部分クラスタが最終的に検索精度が高いかどうかを確認するため，以下の実験を行った．

実験条件

クラスタリング対象として使用した文書は，1993年～1997年に公開された特許文書250件を用いた．

トップダウン検索の場合，検索精度に最も影響を与えそうなのは，最終的なクラスタの最上位にあり，一番最初に決定される部分クラスタである．そこで最上位のクラスタを構築するために選択する文書数 M をM=8, 16, 32, 64と変えて，検索精度を測定した．その際，最上位の部分クラスタ以外が検索精度に影響を与えないよう，部分クラスタリングではなく，厳密なボトムアップクラスタリングを行った．

部分クラスの評価に使用する文書の数は128である．評価用文書集合は部分クラスタの構築に先立って選び出してあり，常に同一の評価用文書集合で部分クラスタの評価を行うようにした．

部分クラスタ試行構築は50回で，種々のエントロピーを得るため合えて試行は盲目的に行った．

実験結果

M=8,16,32,64のそれぞれの結果を図1,2,3,4に示す．

各図の横軸は，部分クラスタの葉ノードに分類された評価用文書のエントロピーである．

図の縦軸は，平均Precisoin[5]と呼ばれるもので，情報検索で使用されるRecall-Precision曲線を積分したも

"Top-down Clustering of Document Set Based on MDL Criteria", Kazunori MATSUMOTO, Keiko AOKI, Keiichirou HOASHI and Kazuo HASHIMOTO: KDD R&D Laboratories Inc..

のであり，値が大きい程，検索精度が高い．

考察

クラスタ生成用文書とテスト用文書が異なること，クラスタ構築用文書の数が比較的小さいこと，盲目的な試行のため評価値が高い部分クラスタを見のがしている可能性があることから，必ずしもエントロピー最大の部分クラスタが最高の検索精度を示していない．しかし，図1〜3では，評価値と精度に比較的良い相関が見られる．

部分クラスタを大きく(Mを大きく)すると，最大エントロピーは大きくなる．しかし，異なるMで構築した部分クラスタを評価するためには，部分クラスタを表現するモデル自体の記述長で比較しなければならない．モデルの記述長は，部分クラスタの木自体の符号長 L_1 と 分類状況を表す符号長(エントロピー) L_2 の和で表現できる．

木の中間ノードを0，葉ノードを1で符号化[4]すると，部分クラスタは葉ノードの数が Mなので，$L_1 = 2M - 1$ となる．よって，M=8の場合はエントロピーに15を，M=16では31を加えれば，部分クラスタのモデルの記述長が得られる．

MDL基準に従うと，モデルの記述長が最小のモデルを選択することになる．このため，M=64の場合，エントロピーが小さい部分クラスタでもモデル全体の記述長はかなり大きくなる．このことは，限られた文書数で部分クラスタを評価する場合，あまりMを大きくしない方が良いことを意味しており，実際，M=64の場合，エントロピーと精度の相関はあまり良くない．

部分クラスタ評価用文書数が128だとエントロピーが小さく，木の記述長が小さいものが有利になる．実際，M=8では，最大エントロピーの部分クラスタは，ほぼ検索精度も最大である．

4. おわりに

トップダウンクラスタリングで部分クラスタをエントロピーで選択することの妥当性を検証するため，複数サイズの部分クラスタのエントロピーを計測した．実験の結果，クラスタ評価用の文書数に応じた適切なサイズの部分クラスタで，エントロピー最大の部分クラスタを選ぶの方法は妥当であることが分かった．

参考文献

[1] 岩山, 徳永, 桜井, "文書検索のための大規模クラスタリング", 言語処理学会第3回年次大会 (1997年3月), pp.245-248, 1997

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

[4] 伊藤, 川端, "パラメータ分散推定量を用いたユニバーサル・データ圧縮アルゴリズム", 第8回情報理論とその応用研究会, p.239-244, 1985

[5] Vorhees, Harman: "The Fifth Text Retrieval Conferenc (TREC-5)", NIST SP 500-238, 1997

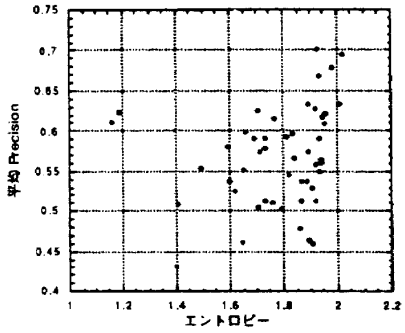

図 1: 部分クラスタの評価値と検索精度 (M=8)

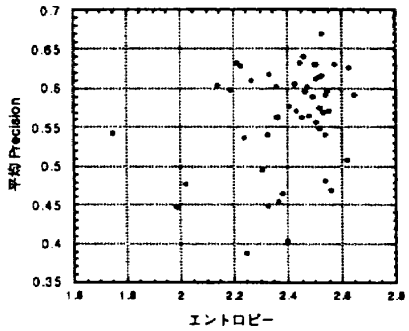

図 2: 部分クラスタの評価値と検索精度 (M=16)

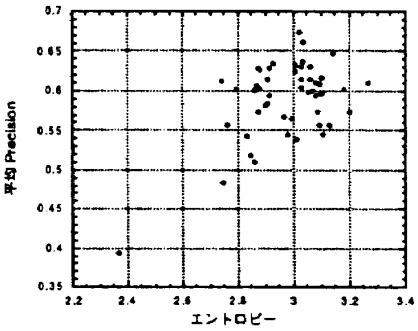

図 3: 部分クラスタの評価値と検索精度 (M=32)

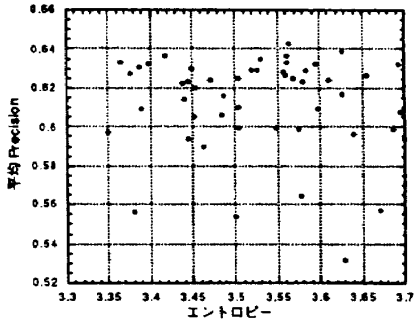

図 4: 部分クラスタの評価値と検索精度 (M=64)