大規模文書集合の高速クラスタリング

田中英輝
NHK 放送技術研究所
tanakah@strl.nhk.or.jp

1 はじめに

著者らは現在，英日のニュース記事の対訳データベースを利用した翻訳支援システムの研究開発を行なっている [3]。これには，日本語記事そのものを検索キーとして対訳データベースから類似の日本語記事を検索する機能を持たせる予定である。

このシステムが対象とするデータベースは大きい。また，ユーザができるだけ多くの記事を「斜め読み」できるシステムを念頭に置いている [3]。多くの記事を扱うには高速な検索が必須であり，また「斜め読み」を実現するには記事集合に構造を持たせておくと便利である。このため，文献 [1] のようにあらかじめ記事集合を階層構造にクラスタリングしておく手法を考えている。

文書集合のクラスタリングには「組み合わせ的な凝集型階層クラスタリング」を使うことが多い。しかしこの手法は分類対象の数 N の二乗に比例する大きさの距離行列を使うため，分類対象の数が増えると実行が困難になる問題がある。

また，基本的にクラスタリング（分類）は判別と無関係であるため [2]，検索戦略は別途考えなくてはならない問題がある。

本稿ではこのような問題を考慮したクラスタリングアルゴリズムを提案する。本手法は変量の観測値だけを使うため記憶領域が少なくて済む。また事例の判別木の形でクラスタ階層を出力するため判別に直接利用可能である。

以下では本アルゴリズムの基本型を説明し，これを文書集合に適用する場合の変形に付いて述べる。また，日本語記事約 33,000 件を使ったクラスタリング実験の結果について述べる。

2 諸定義

2.1 データ形式

本アルゴリズムへ入力するのは，事例の集合である。事例は複数の変量の観測値からなる。観測値は整数である。表 1 に事例集合の例を示す。本稿では事例の数を N で，変量の数を M で表す。

表 1: 入力データ

事例	変量 1	変量 2
a	1	1
b	1	3
c	2	1
d	2	1
e	3	3
f	3	3

ここで配列の要素の参照の仕方を次のように規定する[1]。配列 A の j 番目の要素を $A[j]$ で参照する。配列 A の s 番目の要素から e 番目 ($s \leq e$) までの要素を $A[s,e]$ で参照する。$A[s,e]$ の平均を $\overline{A}[s,e]$ で表す。

n 個の事例がある場合，変量 i は n 個の観測値を持つ。この観測値の集合を n 次元の配列に格納する。この配列を $X^{(i)}$ で表す。先の規定に従うと変量 i の全観測値は $X^{(i)}[1,n]$ となる。

2.2 観測値集合の変動

観測値集合 $X^{(i)}[1,n]$ の変動は次式で計算できる。

$$t(X^{(i)}[1,n]) = \sum_{j=1}^{n}(X^{(i)}[j] - \overline{X}^{(i)}[1,n])^2 \qquad (1)$$

観測値集合が平均値の周りに集まっていれば変動は小さくなる。すなわち変動は観測値集合のまとまり具合いの指標である。

2.3 観測値集合の二分割

観測値集合 $X^{(i)}[1,n]$ を p 番目の要素で二分割するとは次の操作を指す。

まず，配列 $X^{(i)}[1,n]$ の要素を昇順に整列する。次に $X^{(i)}[1,n]$ を $X^{(i)}[1,p]$ と $X^{(i)}[p+1,n]$ の二つの部分集合に分割する（$1 \leq p < n$）。

[1]本稿では配列と集合を同じ意味で使う。

2.4 変動の減少

観測値集合 $X^{(i)}[1,n]$ を $X^{(i)}[1,p]$ と $X^{(i)}[p+1,n]$ に二分割する。この二分割で発生する変動減少 $b(X^{(i)}[1,n],p)$ は次式で計算できる。

$$b(X^{(i)}[1,n],p) = \quad (2)$$
$$t(X^{(i)}[1,n]) - t(X^{(i)}[1,p]) - t(X^{(i)}[p+1,n])$$

この値の大きな分割ほど，得られた二つの部分集合のまとまりが良くなったことを示す。

3 アルゴリズム

提案アルゴリズムの基本型は以下の通りである。

初期事例集合（大きさ N）を入力する。

初期事例集合全体を一つのクラスタとし，これを持つルートノードを作成する。

$n = N$ とする。

1. 終了条件の確認
   事例集合の大きさ n が規定の数以下であれば終了
2. 最適変量選択
   (a) 各変量 $i, (1 \leq i \leq M)$ を対象に以下の処理を行なう。
   (b) $X^{(i)}[1,n]$ の要素を昇順に整列する。
   (c) $X^{(i)}[1,n]$ のすべての可能な二分割を行ない，そのときの変動減少 $b(X^{(i)}[1,n],p)$ を記録する $(1 \leq p < n)$。
   記録した結果から，最大の変動減少を与える変量 i_b とその分割点 p_b を求める。これらを最適変量，および最適分割点と呼ぶ。もし最大変動減少がゼロならば終了する。そうでなければ分割閾値を $\frac{X^{(i_b)}[p_b]+X^{(i_b)}[p_b+1]}{2}$ で計算する。
3. 事例集合の分割と樹形図の作成
   入力された各事例の最適変量の値が，分割閾値より大きければその事例をグループ 1 に分類する。そうでなければグループ 2 に分類する。現在のノードから子ノード二つを作成し，グループ 1 の事例集合を左の子ノードに，事例集合 2 の事例集合を右の子ノードに割り当てる。
4. 再帰処理
   グループ 1 の事例集合に対してクラスタリングを再帰的に実行する。グループ 2 の事例集合に対してクラスタリングを再帰的に実行する。

図 1 に表 1 をクラスタリングした結果を示す。本手法の主な特徴は以下の通りである。

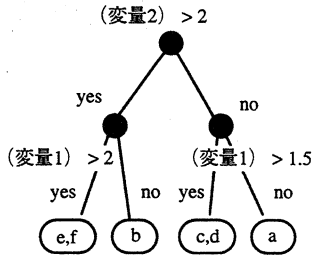

図 1: 樹形図

- 距離の不使用
  本手法は事例間の距離を使わず事例の観測値のデータに基づいてクラスタリングを行なう。通常のクラスタリングの問題では M は定数であるため，本手法の必要とする記憶領域の大きさは $O(N)$ である。
- クラスタの特徴を説明可能
  本手法で得られる階層は二分木であり，各ノードには変量の不等式が対応している。このため，事例があるクラスタに分類された理由を，ルートからリーフまでの変量の不等式の集合で表現できる。

4 文書集合クラスタリングへの適用

提案手法で文書集合をクラスタリングするには，事例は文書に対応し，変量に単語を使い，観測値にその生起頻度を使う。このとき全文書に出現した単語すべてを変量に使うと，変量数 M が N に依存する巨大な数になる。また表 1 形式のデータは極めて巨大かつスパースな行列となるので前章の手法はそのままでは非効率的になる。ここではこの問題を解決する手法を示す。

4.1 頻度間の分割

文書集合のクラスタリングの場合 $X^{(i)}[1,n]$ の要素には同じ値 (頻度) が多数出現する。大半が 0 で 1，2 と続く。また最適分割点は異った頻度の間となるので，$X^{(i)}[1,n]$ を一つづつ分割評価するのは無駄が多い。

そこで，表の各 $X^{(i)}[1,n]$ を頻度のヒストグラムの形で管理して異る頻度間で分割を評価する。そうすると分割評価の回数を大幅に減らすことができる。また最適変量選択処理の要素の整列が不要になる。

例えば表 1 の変量 2 は，頻度 1 が 3 回，頻度 3 が 3 回と管理すると，分割作業に使う要素数は 2 個となり，分割は 1 回で済む。

4.2 ターム飛び越し条件

変量数 M が大きくなると二分割評価の数が増大し，処理時間が長くなる。ここでは一部の変量についてのみ二分割を評価するだけで最適な観測項目と分割点を求める手法を示す。尚，変量（単語）の集合をタームリストと呼ぶ。

初期事例集合を $X^{(i)}[1,N]$, $(1 \leq i \leq M)$, $(n \leq N)$ とする。まずタームリストの単語をその初期変動 $t(X^{(i)}[1,N])$ （i が単語に対応する）の大きな順に並び替える。すなわち次の関係が成り立つようにしておく。

$$t(X^{(i)}[1,N]) \geq t(X^{(i+1)}[1,N]) \qquad (3)$$

「最適変量選択処理」では並び替えた変量の 1 番から順番に最適分割点を求めていく。ここで次の規則が成り立つことが証明できる。

> 観測値集合 $X^{(i)}[1,n]$ を対象に二分割を行なった結果，最適分割点が $p_b^{(i)}$ となったとする。もし
>
> $$t(X^{(i+1)}[1,N]) \leq b(X^{(i)}[1,n], p_b^{(i)}) \quad (4)$$
>
> が成立するなら $i+1$ 以降の変量は二分割の評価を行なう必要がない。そこまでで得られている最適変量とその最適分割点が求める解である。

以上の条件を最適変量処理に加えると実行時間を大幅に短縮できる。

この条件の証明は省略するが，次式で示す「変動」の性質を使うと簡単に証明できる。

$$0 \leq b(X^{(i)}[1,n], p) \leq t(X^{(i)}[1,n]) \qquad (5)$$

この不等式は，集合を分割すれば非負の変動減少があること，変動減少の最大値は高々元の集合の持っている変動となることを示している。

5 クラスタリング実験

5.1 対象文書集合

NHK の放送原稿データベースから抽出した二つの記事集合を対象とした。各記事はニュースの一項目（話題）を単位としており，一記事の平均文数は 5.2，一文の平均文字数は 88.9 である [4]。また，二言語放送や海外向けラジオ放送のため，一部の記事は英訳されている。

本実験で使用した日本語記事集合は以下の通りである。

表 2: 計算量に関する結果

文書集合	1	2
記事数	33,617	8,538
ターム数	60,115	26,771
実効ターム数	14,549	5,200
分割テスト数	2.54×10^4	6.44×10^3
評価セル数 (Ward)	6.33×10^{12}	1.04×10^{11}
実行時間	14h 50m	1h 20m

- 記事集合 1
  1995 年 3 月から 1996 年 2 月までの 33,617 記事
- 記事集合 2
  1995 年 3 月から 1997 年 6 月までの記事のうち，英訳された 8,538 記事

クラスタリングは各記事集合に出現したすべての名詞を変量に使って行なった。使った計算機の記憶容量は 256 MB であり，処理速度は $SPECint92 = 202.9$, $SPECfp92 = 259.5$ である。

5.2 計算量評価

表 2 に計算量に関する結果を示す。この実験で採用した「終了条件」は（記事数 ≤ 2）である。

二行目の「ターム数」はタームリストが含んだ名詞の数である。

三行目の「実効ターム数」は，最適な変量（ターム）を選択するのに評価したタームの数の平均である。この数は「ターム飛び越し条件」がなければ二行目のターム数に一致するので，両者の差はターム飛び越し条件の効果を表している。文書集合 1 でタームリストの数は実効的に $\frac{1}{4}$ に減少しており，文書集合 2 では $\frac{1}{5}$ に減少している。

四行目の「分割テスト数」は発生した分割テスト数の総計である。五行目の「評価セル (Ward)」は Ward 法で必要となる距離の比較回数である[2]。この数にくらべて「分割テスト」の数は桁違いに小さい[3]。尚，記事数 6,225 の文書集合に対して Ward 法を適用したところ実行時間は 8 時間となった。またこの記事数がメモリー上で実行できる上限であった。

両方のプログラムは同じレベルでコーディングしている。以上の結果より，扱える文書集合の規模に関しては本手法が優位であったと考える。

5.3 クラスタの評価

図 2 に文書集合 2 で得られたクラスタ階層の一部を示す。これは終了条件（記事数 ≤ 20）で得られた

[2] 基本的な手法を想定して算出した。事例数を $N+1$ とすると $\sum_{i=1}^{N} \frac{i(i+1)}{2} = \frac{1}{6}N(N+1)(N+2)$ となる。

[3] 全体の計算量比較は，本手法の初期変動計算部分，Ward 法の距離行列計算部分を考慮する必要がある。

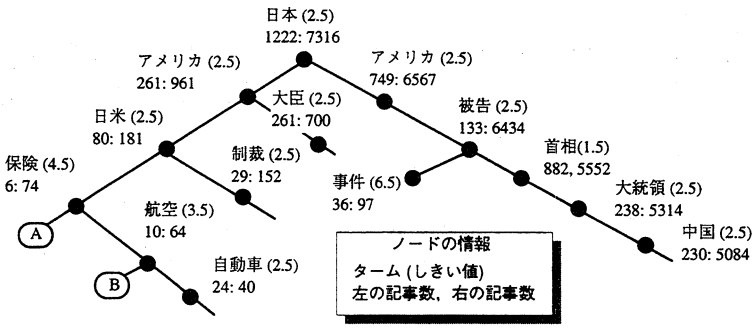

図 2: 得られた階層の一部

表 3: 同一クラスタの記事タイトル

リーフ A
カンター、保険合意の完全実施求める
日米保険協議難航
日米保険協議、シャピロ大使悲観的な
日米保険協議決着先送り
米政府高官、要求無視すれば制裁
保険協議、早急な解決で一致
リーフ B
日米首脳会談を前に、フィルム、航空
日米旅客協議再開に向けて非公式協議
日米航空協議、物別れ
日米航空交渉物別れ
日米航空問題・結局 物別れ
フェデラル副社長会見
制裁回避めざし日米航空交渉始まる
朝用・日米航空交渉
夜用・日米航空交渉決着
大阪・関空フェデラル第一便

ものである。

この階層では，タームを閾値以上含んだ記事を左の枝に分類し，残りを右の枝に分類している。このため，階層は右方向に深くなる傾向がある。この図では左下の白抜きの部分がリーフであり，記事を格納している。ここに格納された記事タイトルを表 3 に示す。

同じリーフには類似した話題の記事タイトルが並んでおり妥当な結果であろう。階層全体を眺めた場合，右の深い部分は関係の薄い記事群が分類される傾向があり，一度の分割で一つの記事が分割されることもある。右の枝の最も深い場所の記事群は極端に短い事件の情報であり通常の記事ではなかった。

図 2 の各ノードの単語は分割の各局面で最大の変動を持つ，すなわち記事集合の分離力が最大の単語である。これはある種のキーワードと見なすことができる。そうすると本手法の階層作成の手順から，階層の最右辺の単語群は全記事集合を大まかに分割するキーワード（主キーワード）となる。一方，それ以外の単語は主キーワードで分割された記事をさらに分割するキーワード（副キーワード）とみなすことができる。

文書集合 1 の主キーワードは「日本, 被告, 選挙, 容疑者, 地震, 選手, 事件…」であった。図 2 の主キーワードには国名や政治家の肩書きが多く出現しており，文書集合 1 に比べて国際関係，政治関係のニュースの話題が中心であることが推察できる。これは，英語放送の性格からして妥当な推察であろう。

6 おわりに

距離行列を使わない階層クラスタリング手法を提案した。また，本手法を文書クラスタリングに応用して，従来より大規模な文書集合を対象にできることを示した。今後さらに大きな記事集合を対象にクラスタリング実験を行なう予定である。また，本手法を応用した類似記事検索システムの構築を進めており，これらについては今後報告したい。

参考文献

[1] M. Iwayama and T. Tokunaga. Cluster-based text categorization: A comparison of category search strategies. In *Proc. SIGIR 95*, pp. 273–280, 1995.

[2] 竹内啓. 統計学辞典. 東洋経済新報社, 1989.

[3] 熊野正, 田中英輝, 浦谷則好, 江原暉将. 日英放送原稿翻訳支援のための類似用例提示システム. 言語処理学会第 3 回年次大会, D5–1, 1997.

[4] 熊野正, 田中英輝, 金淵培, 浦谷則好. 日英ニュース原稿の対訳コーパス化に関する基礎検討. 言語処理学会第 2 回年次大会, pp. 41–44, 1996.