6N－7

階層分割型クラスタリングを使った文書ブラウザ

田中 英輝† 熊野 正† 松田 伸洋‡

†NHK 放送技術研究所 ‡(株)漢字情報サービス

1 はじめに

著者らは明確な意図を持った検索に加えて文書ブラウザを実現することを目的とした研究を行なっている。そしてこの目的に田中ら [4] の提案した逐次二分割の階層クラスタリング手法を使っている。この種のクラスタリング手法は一般的な用途でも "largely ignored[1] (pp. 253)" とされ，また単項的 (monothetic) な階層木を出力するため検索にも向かないとされていた [3]。

しかし，後述するように文書集合の構造を明示でき，文書ブラウザを構成するのに適した特徴を持つ。本稿ではこのクラスタリング手法の概要と試作したブラウザについて報告する。

2 クラスタリング手法

同一話題の文書集合には「それらのみ」に限って出現する特徴的な単語がある。逆に文書集合の部分集合に際立って出現する単語は話題を限定する力を持つ。田中らのクラスタリングアルゴリズム [4] はこのような仮定に基づいている。

本稿では文書の部分集合に際立って出現する単語を話題限定語と呼ぶ。例えば，1997 年前半のニュース記事の集合の中で「セルパ[1]」という単語を 2 回以上含む記事は「ペルー事件」関連に違いない。ここで使うアルゴリズムは「セルパ」のような話題限定語とその閾値頻度を求め，これを使って逐次二分割クラスタリングを行なう。すなわち，最初に文書集合全体を一つのクラスタとする。そして最良の話題限定語とその閾値頻度を求め，この語を閾値以上含む文書集合とそれ以外に二分割する。以後，2 つの部分集合を対象にして二分割を再帰的に繰り返す。

最良な話題限定語は文書集合における単語の出現頻度分布を使って求めている。すなわち，このクラスタリング手法は単語の頻度表だけを使うため必要な記憶容量が小さい。また，最良の話題限定語を高速に求める手法があることから，従来手法で処理困難な大規模な文書集合にも適用可能である。

"An application of hierarchic divisive clustering to a document browser"

TANAKA Hideki(tanakah@strl.nhk.or.jp),
KUMANO Tadashi, MATSUDA Shin'yo,
NHK Science and Technical Research Laboratories
1-10-11 Kinuta, Setagaya-ku, Tokyo, JAPAN 157

[1]1996 年 12 月から 6ヶ月にわたってペルーの日本大使館を占拠したゲリラグループのリーダの名前。

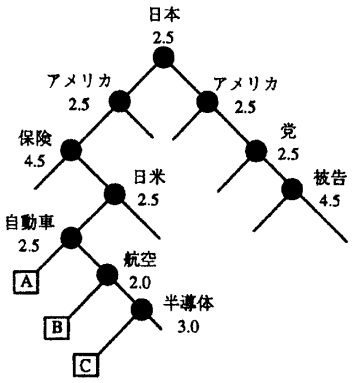

図 1: 階層木

3 ブラウザへの応用

本稿のクラスタリング手法を使って文書集合を分類した階層木の一部を図 1 に示す[2]。利用した文書集合は 1996 年 3 月から 1997 年 2 月までの NHK の日本語ニュース記事 (英訳をもつ記事) 14,701 件である。この図では，各ノードに話題限定語と閾値頻度を表示している。また各ノードでは話題限定語を閾値以上含む文書を左下の子ノードに，それ以外を右下の子ノードに分類している。階層木の特徴を以下に示す。

- 各クラスタ特徴の明示
  図 1 のリーフ A の文書集合は「日本，アメリカ，日米，自動車，という話題限定語を閾値以上含んでいる」とその特徴を明示できる
- 文書集合の構造の明示
  最初に全文書群をルートノードの話題限定語「日本」を 2.5 以上含むかのテストをする。すると大半の文書は右下のノードに分類される。このように最右辺上の話題限定語は基本的に大きな文書群を分類する語になっており文書集合全体の特徴を示す
- 検索へ応用可能
  記事を入力するとその類似記事を出力できる。すなわち，入力を解析して単語の頻度分布を得るとそれを使ってリーフの類似記事を検索できる

以上の特徴を生かしたニュース記事のブラウザを作成した。このシステムには 1) 類似記事検索機能と 2) 記事のブラウジング機能がある。

[2]尚，ここではルートからリーフに向かう経路上に同一単語が出現しないような制限をクラスタリングアルゴリズムに追加している。

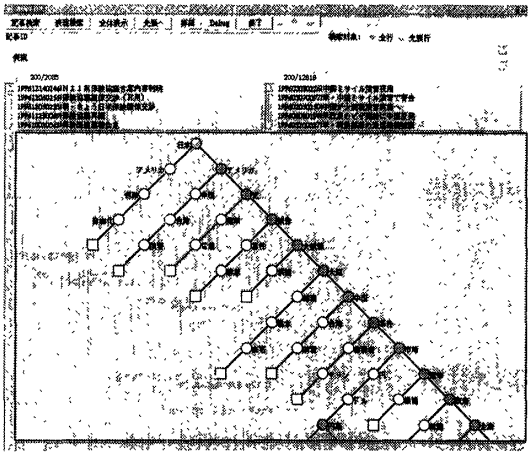

図 2: 開始画面

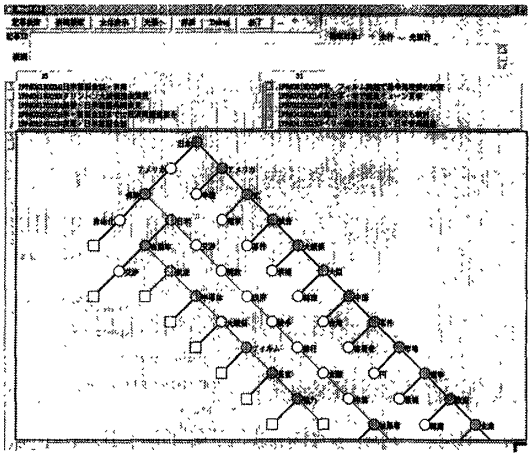

図 3: 展開画面

初期状態では図 2 に示す折り畳んだ階層木を示している。ここでは主キーワード[3]を表示してその下の話題限定語を可能限り表示する。すなわちユーザは文書集合全体の話題を把握することができる。

ユーザは任意のノードとリーフをマウスクリックで指定でき，その下に含まれる記事のタイトルを中程の二つのウインドウで閲覧できる。二つのウインドウはノードの下の二本の枝に対応している。またタイトルをクリックすると本文を閲覧できる。

ユーザが折り畳まれたノードをクリックした場合には，その下のノードを可能な限り展開して表示する。これによってユーザは興味ある話題に関する詳細な分類を見ることができる（図 3）。

また記事の類似検索を階層木を使って行なう事ができ，この時検索した階層木上の経路を表示する（図 4）。このためユーザは文書群が検索された根拠を形路上の話題限定語から知ることができる。さらにここからブラウジングを開始することが可能である。

[3]最右辺上の話題限定語

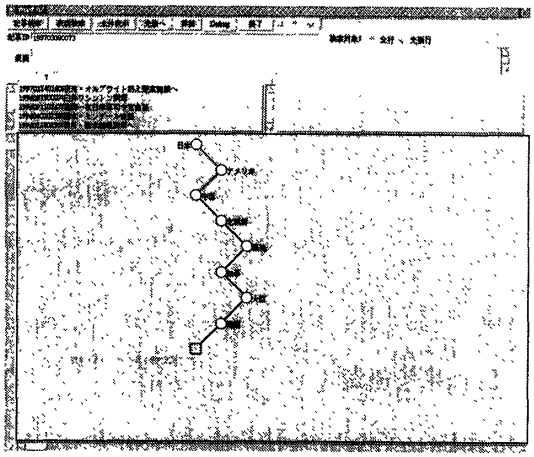

図 4: 検索画面

4 おわりに

本稿では逐次二分割型階層クラスタリング手法を使った文書ブラウザの概要を報告した。現在このブラウザは用例提示型翻訳支援システム [5] の一部として利用しており性能を評価中である。

今回の階層木の主キーワードの数は 416 で一見するとかなり多い。しかし最右辺上の深さ 100 以上ノードは 1,141 件の記事しか分類していない。すなわち 13,560 件の記事は 100 の主キーワードで分類できており，深さ 100 以上の部分では雑音的な記事を延々と分類した可能性が高い。これを防止するにはクラスタリングの停止条件を工夫する必要がある。

今回はすべての名詞を使って階層木を作成したが，あらかじめある観点で選択した名詞群だけを使う手法も考えられる。これについては今後実験したい。

参考文献

[1] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data.* A Wiley-Interscience Publication, 1990.

[2] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval.* McGraw-Hill, 1983.

[3] P. Willet. Recent trends in hierarchic document clustering: A critical review. *Information Processing and Management*, Vol. 24, No. 5, pp. 577–597, 1988.

[4] 田中英輝. 大規模文書集合の高速クラスタリング. 言語処理学会第 3 回年次大会, pp. 249–252, 1997.

[5] 熊野正, 田中英輝, 松田伸洋, 浦谷則好, 江原暉将. 日英放送原稿翻訳者のための類似用例提示型翻訳支援システム. 情報処理学会第 55 回全国大会, 1997.