ラベルに基づく文書クラスタリング手法の提案と評価

田之上 和誠† 岡部 正幸†† 梅村 恭司‡

†‡豊橋技術科学大学 情報工学系 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

††豊橋技術科学大学 情報メディア基盤センター 〒441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1

E-mail: †tanoue@ss.cs.tut.ac.jp, ††okabe@imc.tut.ac.jp, ‡umemura@tut.jp

あらまし 本稿では，文書ベクトルを構成せずラベルに基づいてソフトクラスタを作成する文書クラスタリング手法を提案し，提案手法により作成されたクラスタの妥当性の評価を行う．ラベルは，文書集合から抽出したキーワード群の中から出現回数に基づいて選んだ 2 つのキーワードで構成し，この 2 つのキーワードが共に出現する文書をまとめることでクラスタリングを行う．この方法にはクラスタのラベルが当初より決まっているという利点がある．評価は，単連結法と呼ばれる文書間の類似度に基づいたクラスタリング手法により作成されたクラスタと妥当性を比較することで行い, 妥当性の尺度には同じクラスタ内の 2 つの文書間の類似度の重み付き期待値を使用した. その結果，クラスタのラベルは存在するものの，提案手法は単連結法に比べてクラスタの妥当性は低かった．

キーワード 文書クラスタリング，キーワード抽出

Evaluation of Proposed Document Clustering based on Labels

Kazumasa TANOUE† Masayuki OKABE†† Kyoji UMEMURA‡

†‡ Department of Information and Computer Sciences, Toyohashi University of Technology

†† Information and Media Center, Toyohashi University of Technology

1-1 Hibarigaoka, Tempaku, Toyohashi, Aichi, 441-8580 Japan

E-mail: †tanoue@ss.cs.tut.ac.jp, ††okabe@imc.tut.ac.jp, ‡umemura@tut.jp

Abstract In this paper, we propose a new soft clustering method for documents based on labels without document vectors, and evaluate adequacy of clusters created by our method. Labels consist of two keywords selected by their frequency, and clustering can be carried out by gathering documents which include both keywords of label. Our method has an advantage that the labels of clusters are initially obvious. We carry out evaluation by comparing the adequacy of the clusters created by our method and single linkage method which is based on the similarity between two documents. Adopted adequacy criterion is weighted expectation value of similarity between two documents in same cluster. Experimental evaluation shows that our method gets lower adequacy than single linkage method, but our method has the initially obvious label advantage.

Keyword Document Clustering, Keyword Extraction

1. はじめに

情報検索の分野では，図書や雑誌論文などの文書集合を，内容的に均質ないくつかの群に分けるための文書クラスタリングの研究が長年にわたって試みられてきた．文書クラスタリングの手法は階層型と非階層型に大別でき，代表的なクラスタリング手法として，階層型では単連結法，非階層型では k-means 法がある．いずれにしろ，一般的な文書クラスタリング手法では文書を各語の重みから構成されるベクトルとして表現し，ある類似尺度をもって 2 つの文書ベクトル間の類似度を定義した後，この類似度に基づいてクラスタリングを行う．しかし，作成されたクラスタに属する文書の内容を一通り読まなければ，そのクラスタがどのような内容の文書群なのかが把握できないため，クラスタリング結果の概要を一目で確認することができず不便である．

本稿では，あるクラスタがどのような内容の文書群なのかを端的に表すものとして，クラスタに共通な主題(ラベル)が分かっていることは重要であると考え，

クラスタリングに先立ってまずラベルを作成し，ラベルに即する文書をまとめてソフトクラスタを作成する手法を提案する．あるクラスタについてそのラベルが分かっていれば表示の際に便利であると考えられ，検索効率の更なる向上およびより良い検索支援になることが期待できる．

実験では，武田らのキーワード抽出プログラム[1]を利用して NTCIR1 からキーワードを抽出し，抽出されたキーワード群の中から，出現回数に基づいて選んだ 2 つのキーワードを用いてラベルを構成し，ラベルのキーワードが共に出現する文書をまとめることでソフトクラスタを作成した．そして提案手法により作成されたクラスタと，単連結法により作成されたクラスタとの妥当性の比較を行った．その結果，同じクラスタ内の 2 つの文書間の類似度の重み付き期待値という評価尺度の下に，我々の提案する手法は単連結法よりもクラスタの妥当性が低い結果となった．しかしその後の改善の試みの中で，出現回数別にキーワードを組合せていくことで妥当性が改善されるのではないかという見込みがあったことを報告する．

2. 単連結法

2 つの文書 d_i と d_j の類似度を s_{ij} と表記する．文書集合に対して単連結法を適用する場合，2 つのクラスタ C_m と C_n との類似度 S_{mn} は

$$S_{mn} = \max\{s_{ij} \mid d_i \in C_m, d_j \in C_n\} \qquad (1)$$

で定義される．つまり，それぞれのクラスタに含まれる文書のうち，内容的に最も近い文書間の類似度をクラスタ間の類似度として採用する．アルゴリズム 1 に単連結法の手順を示す．

3. ラベルに基づくクラスタリング

単連結法や k-means 法などの一般的な文書クラスタリング手法では，文書を各語の重みから構成されるベクトルとして表現し，2 つの文書ベクトル間の類似度に基づいてクラスタリングを行うが，提案手法では文書をベクトルとして表現せず，その代わりに，文書集合から抽出されたキーワードに基づいてクラスタリングを行う．文書集合に含まれる文書件数を N とすると，多くの場合，抽出されたキーワードの総数 K は $K<N$ となると考えられる．また，抽出されたキーワードは文書集合において重要であると判断された語句であるため，このキーワードを本文中に含む文書をまとめることで妥当なクラスタが作成されると考えられる．しかし，単一のキーワードに基づいて文書をクラスタリングした場合，作成されたクラスタ内の文書の内容に大きなばらつきがでるのではないかと考え，2 つのキーワード組み合わせたものをラベルとし，本文中にラベルのキーワードを共に含む文書をまとめることでより類似した内容のクラスタが作成されるのではないかと考えた．提案手法ではラベルとして採用するキーワードのペアをどのように選択するかが重要なポイントであり，この部分がクラスタリングの結果を大きく左右する．アルゴリズム 2 にラベルに基づくクラスタリングの手順を示す．

アルゴリズム 1　単連結法

1. 文書集合 D の全ての文書に対するベクトルを構成
2. D の全ての文書ベクトル間の類似度を計算
3. 各文書 $d_i \in D$ を，自分自身のみを持つ集合 $\{d_i\}$ とする
4. G を，全ての $\{d_i\}$ からなる集合とする
5. ***while*** G の要素数 > 1
6. $S_{mn} = \max\{s_{ij} \mid d_i \in C_m, d_j \in C_n\}$ となるような $C_m, C_n \in G$ を選択
7. $C_{new} = C_m \cup C_n$ とする
8. 現在の類似度 S_{mn} と併合した C_{new} の情報をファイルに書き出す
9. G から C_m と C_n を削除
10. G に C_{new} を挿入する
11. ***end***

アルゴリズム 2　ラベルに基づくクラスタリング

1. 文書集合 D からキーワード群を抽出し，キーワードのリストを作成
2. リストから特定の条件を満たすキーワード群を選択し，これをラベルの候補とする
3. ラベル候補のキーワード群の中から規則に従って 2 つ選択し，このキーワードのペアをラベルとする
4. ラベルのキーワードを共に含む文書をまとめてクラスタを作成
5. ラベルとして使用されたキーワードをリストから除外する
6. 手順 3～手順 5 をリストが空になるまで繰り返す

4. 実験

4.1. 実験の概要

本稿では約33万件の論文抄録であるNTCIR1に対して単連結法と提案手法によるクラスタリングを行い，両手法において作成されたクラスタに対して，同じクラスタ内の2つの文書間の類似度の重み付き期待値という評価尺度の下で比較・評価を行った．

単連結法は，与えられた N 件の文書に対して全ての文書間の類似度を計算するため，計算量は $O(N^2)$ に比例する．そのため，実際には単連結法の手順通りに全ての文書間の類似度は計算せず，山本らの DP マッチングによる類似度算出プログラム[2]を使用し，各文書に対する類似文書上位 30 件とその類似度のリストを用意した．そして文書1と文書2，文書2と文書3の類似度がある一定の値よりも高ければ，文書1と文書3が類似している可能性が高いという仮定の下，リスト 1 に示す手順によりクラスタリングを行った．

ラベルに基づくクラスタリングでは，ラベルの数＝クラスタの数となることに加え，ラベルとして採用するキーワードがクラスタリングの結果を大きく左右する．実験では，小規模なクラスタができることを期待して，抽出されたキーワード群の中から NTCIR1 における出現回数が 10~20 回のキーワードをラベルの候補とし，また大規模なクラスタができることを期待して出現回数が100から500回のキーワードもラベルの候補として，出現回数が 10~20 回のキーワードと100~500 回のキーワードを組合せてラベルを構成し，リスト2に示す手順によりクラスタリングを行った．

4.2. 実験結果

両手法により作成されたクラスタの数を規模別に示したグラフを図1に示す．単連結法では，クラスタの数が NTCIR1 の文書数の 1／10 程度になるように類似度の閾値 63 におけるクラスタを取得した．両手法において規模が 0~10 のクラスタが非常に多くなった．提案手法では規模が0~10のクラスタが12,737個得られたが，そのうち 10,115 個が規模 0 のクラスタであり，ほとんどのラベルでクラスタが作成されない結果となった．しかし，規模が 20~40 のクラスタに関しては両手法においてほぼ同じ数のクラスタを得ることができた．そのため，評価対象は規模が 20~40 のクラスタに限定し，それ以外の規模のクラスタは両手法でクラスタ数に大きな差があるため比較には不適切であると考え評価の対象外とした．

リスト1　実験での単連結法の手順

1. 山本らの DP マッチングプログラムを用いて，NTCIR1 の各文書 d_i に対する類似文書上位30件とその類似度のリストを作成し，この 30 件を文書 d_i に対するクラスタ C_i とする
2. クラスタ数が NTCIR1 の文書数の 1／10 程度になるように類似度の閾値を 63 に設定
3. 手順1で作成した各クラスタ C_i から，設定した閾値以下の類似度の文書を除外する
4. もし，文書 d_1 に対するクラスタ C_1 に文書 d_2 が含まれていれば，文書 d_2 に対するクラスタ C_2 と C_1 を併合して新たに C_1 とする
5. 手順4を全てのクラスタに対して行う

リスト2　実験での提案手法の手順

1. 武田らのキーワード抽出プログラムを使用して NTCIR1 からキーワード群を抽出
2. NTCIR1 における出現回数が10～20回のキーワード 191 個と出現回数が 100～500 回のキーワード 71 個をラベルに使用するキーワードの候補とする
3. 出現回数が 10~20 回のキーワード 191 個と，出現回数が100~500個のキーワード71個を組み合わせて 191×71＝13,561 個のラベルを用意
4. 手順3で用意した全てのラベルに対して，ラベルのキーワードを共に含む文書をまとめてクラスタを作成

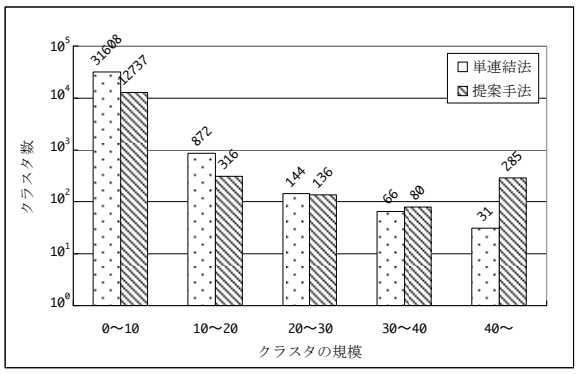

図1　作成されたクラスタ

$$E=\sum_{k=1}^{L} n_k \left(\frac{1}{n_k^{\,2}} \sum_{\substack{d_i \in C_k, d_j \in C_k \\ rank(d_i,d_j) \le 30}} similarity\left(d_i, d_j\right) \right) \Bigg/ \sum_{k=1}^{L} n_k \quad (2)$$

4.3. 評価と考察

文書クラスタリングの結果の妥当性の評価には，直接的な評価と間接的な評価の 2 つが考えられる[3]. NTCIR1 では正解を利用することができないため，クラスタリング結果の妥当性の評価尺度には, (2)式に示す同じクラスタ内の 2 つの文書間の類似度の重み付き期待値を用いた. (2)式では分子のシグマの直後にある n_k がペナルティ項となっており, n_k の大小によって不公平が生じないようになっている．L は評価対象となるクラスタ数，n_k はクラスタ C_k に属する文書数，$similarity(d_i, d_j)$は d_i, d_j間の類似度，$rank(d_i, d_j)$は d_i をもとに検索したときに d_jが出てくる順位である．また(2)式において $similarity(d_i, d_j)$を求める際に，事前に用意した類似文書上位 30 件のリストに類似度が存在する場合にはその値を使用し，存在しない場合には類似度を 0 としたため，得られた期待値は実際には近似的な値である.

表 1 にクラスタの妥当性の評価結果を示す．提案手法のクラスタの妥当性は単連結法の約 1／10 と低く，良い結果が得られなかった．クラスタ数は，規模が 20~40 の間に限定すれば単連結法とほぼ同じ数が得られたので，内容的にもほぼ同じクラスタが生成されているのではないかと予想したが，予想を裏切る結果となった.

提案手法により作成したクラスタの妥当性が低い結果となった原因として，ラベルとして採用したキーワードがクラスタを構成するのに不適切であったのではないかと考え，参考のためにラベルを単一のキーワードとした場合と比較した．表 2 に比較結果を示す．出現回数が 10~20 回のキーワード 191 個をそれぞれラベルとしてクラスタを作成した場合と，出現回数が 100~500 回のキーワード 71 個をそれぞれラベルとしてクラスタを作成した場合の両方において，提案手法よりも妥当性は改善された．しかし評価対象となるクラスタ数は減少してまった．ラベルを単一のキーワードとした場合，NTCIR1 から抽出した単語をキーワードとしているので，用意したキーワードの数だけクラスタが作成されるが，比較のために評価対象となるクラスタを規模が 20～40 の間に限定しているのでこのような結果になってしまったと考えられる．しかし出現回数が 10~20 回のキーワード単体のラベルによりクラスタを作成した場合，評価対象となるクラスタ数は 68 で，提案手法の約 1／3 に現象したのに対し，期待値は提案手法の約 3 倍に伸びた. そのため出現回数が 10~20 回のキーワードはラベルとして有効なキーワード群であると考えられ，組み合わせるキーワード次第ではクラスタの妥当性を向上させることが期待できる.

表 1　妥当性の評価結果

	評価対象クラスタ数	類似度の期待値
単連結法	214	8.706×10^{-2}
提案手法	223	0.911×10^{-2}

表 2　単一キーワードのラベルとの比較

ラベルの種類	評価対象クラスタ数	類似度の期待値
出現回数が 10~20 回のキーワード単体	68	2.686×10^{-2}
出現回数が 100~500 回のキーワード単体	6	1.261×10^{-2}
提案手法	223	0.911×10^{-2}

5. まとめ

本稿では，ラベルに基づく文書クラスタリング手法を提案し，提案手法により作成したクラスタの妥当性の評価を行った．実際には，NTCIR1 に対して単連結法と提案手法によるクラスタリングを行い，両手法において作成されたクラスタに対して，同じクラスタ内の 2 つの文書間の類似度の重み付き期待値という評価尺度の下で比較・評価を行った．その結果，提案手法により作成したクラスタの妥当性は，単連結法のそれよりも約 1／10 程度低い結果となった.

しかし，提案手法はクラスタのラベルが分かっているため，クラスタリングの結果の概要を一目で確認することができる．また単一のキーワードをラベルとしてクラスタリングを試みたところ，出現回数が 10~20 回のキーワード群がラベルとして有効ではないかということが分かった.

参　考　文　献

[1] 武田 善行, 梅村 恭司, "キーワード抽出を実現する文書頻度分析", Mathematical Linguistics vol.23 no.2, pp.65-90, 2001.

[2] E. Yamamoto and M. Kishida and Y. Takenami and Y. Takeda and K. Umemura, "Dynamic programming matching for large scale information retrieval", Proceedings of the sixth international workshop on Information retrieval with Asian languages, pp.100-108, 2003.

[3] 岸田 和明, "文書クラスタリングの技法：文献レビュー", Library and Information Science (49), pp.33-75, 2003.