類似性の不明なデータを手がかりとして与えるクラスタリング手法

佐々木　稔†　松本　良太†　新納　浩幸†

† 茨城大学工学部情報工学科
〒 316-8511 茨城県日立市中成沢町 4-12-1

A Clustering Method Based on an Additional Feature Obtained from Another Document Set

Minoru SASAKI†, Ryota MATSUMOTO†, and Hiroyuki SHINNOU†

† Department of Computer and Information Sciences, Faculty of Engineering, Ibaraki University
4-12-1 Nakanarusawa, Hitachi, Ibaraki 316-8511, JAPAN

1. はじめに

文書クラスタリングは，与えられた文書データの集合に対して，内容の類似している文書をまとめて，クラスタと呼ばれるいくつかの文書集合に自動的に分割するものである．これは文書集合にどのような内容が含まれているのか，いくつのトピックから成り立っているのかなどといった特徴を抽出することができ，新しい情報を発見するテキストマイニングにつながる重要な構成要素となっている．そのため，文書クラスタリングはWeb データ，新聞記事，特許や論文など，様々な文書集合に対して応用されている．

文書クラスタリングを行う処理でまず最初に行うことは，文書の中からクラスタリングに必要な特徴を抽出する処理である．この抽出した特徴を利用して，文書間における特徴の分布が似ているペアをまとめるボトムアップクラスタリングや文書集合全体の二分割を繰り返し行うトップダウンクラスタリングが行われる．従来の文書クラスタリングにおいては，手がかりとなる特徴として，文書に含まれる単語が最もよく利用される．単語を抽出するためには，一般的に形態素解析などを利用して文書内から名詞や動詞が抽出される．このようにして抽出された特徴に対して統計的に類似性を比較するために，特徴の頻度やIDF などの重み付けを考慮した重要度などが計算され，特徴を素性とするベクトルを形成し，それらを索引語-文書行列として表現し，ベクトル空間モデルとして文書間での比較が行われる．

このように，従来の文書クラスタリングでは単語を素性として行っていたが，単語だけではなく，他の素性も利用したベクトル空間モデルに基づくクラスタリング手法や類似度計算手法も提案されている．例えば，Co-Clustering のように，はじめに素性として使う単語をまずクラスタリングして，いくつかの単語をまとめたクラスタを形成し，それを素性とするクラスタリング手法が存在する．この手法を用いることにより，文書クラスタリングの精度が向上したという報告がされている [1] [2]．また，類似度の計算で使われるカーネルトリックは実際に単語をまとめることはしないが，単語間のなんらかの組合せを素性として大きな次元での内積計算が行われている [3]．この方法はサポートベクターマシンで一般的に利用され，最も性能の良い機械学習手法のうちのひとつとしてよく利用されている．さらに，Adaptive Sprinkling のように索引語-文書行列の主成分ベクトルを元の行列に追加することで，精度の高い特徴ベクトルが抽出できることが報告されている [4]．このように単語以外の素性を利用して，文書のクラスタリングや分類などの目的に対する性能を向上する研究が盛んに進められている．

このように， Co-Clustering やサポートベクターマシンは単語素性間の関連性を利用した素性の抽出を行っているが，これまでの研究においてクラスタリングをしたい文書集合と他の文書集合との関連性を素性として利用する手法は提案されていない．クラスタリング対象外の文書を素性として利用すると，関連のある文書間では関連度が高くなるので，対象となる文書をクラスタリングする際の有効な手がかりとなると考えられる．そのため，我々はクラスタリング精度向上を目的として文書間関連度を文書の特徴として追加する手法を提案した [5]．従来の単語を手がかりとしてクラスタリングを行っていたものを他の文書関連度を単語の重要度に加えて文書クラスタリングを行う．その際，実験では新聞記事 120 件に対して，別の新聞記事 120 件を追加して精度比較を行い，性能が改善されたことが示されたが，文書数 120 件と少量のデータに対しての実験であっ

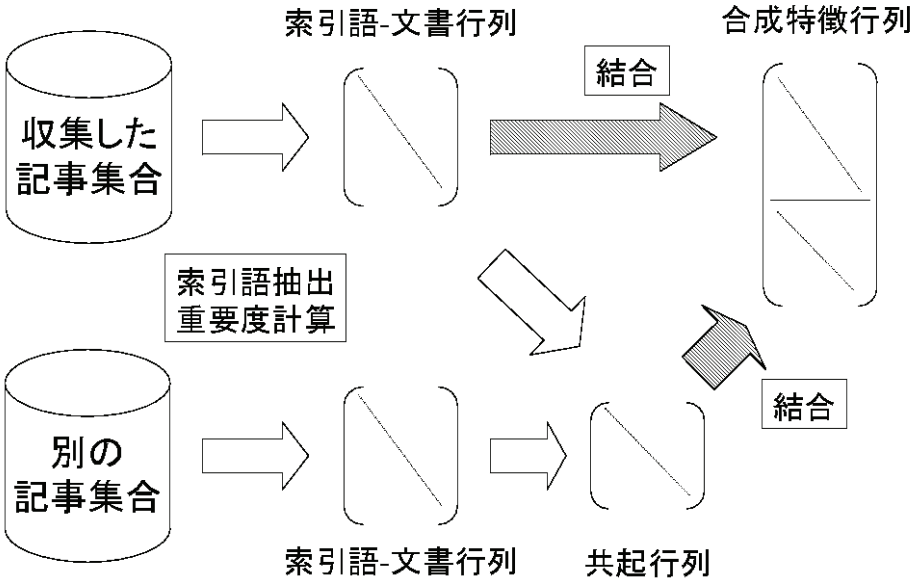

図 1　文書関連度を追加した特徴行列作成の流れ

たため，1000 件ほどのより大きなデータ集合に対する実験を行うことで提案手法の有効性がより確かなものとなると考えられる．また，特徴抽出手法として有効な LSI(Latent Semantic Indexing) などの次元を縮退させてクラスタリングをした場合の精度との比較することも有効性を示す上で重要な点である．本稿では，大規模なデータ集合を利用した場合の精度を比較する実験と次元縮退を行ってクラスタリングした場合の精度を比較する実験をそれぞれ行い，次元伸張を行ったクラスタリング手法の有効性を検証する．

2.　素性の追加によるクラスタリング

2.1　素性の抽出

クラスタリングを行う際に手がかりとなる素性として，文書から単語を抽出する．抽出された単語集合に対して，各素性単語が文書内容についてどの程度の重要度を持っているか計算する．重要度計算には数多くの手法が提案されているが，本稿における実験では TFIDF と呼ばれる尺度を利用した [6]．TFIDF の他にも重要度の計算手法には単語頻度や Log-Entropy などが存在する．本稿では，素性を追加した場合の性能を比較評価することが目的であるため，元の索引語-文書行列で使う重要度計算はどのようなものを利用しても構わないこととした．そのため，単語素性の重み付け手法は重要度計算手法の中で一般的に広く使われている TFIDF を利用することとする．

2.2　手がかりとなる素性の追加方法

我々の研究する文書クラスタリング手法は，上述の素性抽出方法により索引語-文書行列を作成し，その行列に対して文書間の類似度行列を素性として追加した合成行列に対してクラスタリングを行うものである．この類似度行列を追加するために，クラスタリング対象となる文書集合以外の文書集合を用意する．用意した他の文書集合に対して元の行列と同様にして索引語-文書行列を計算する．元行列の各列と追加文書行列の各列に対して類似度計算を行い，追加文書を行ベクトルに，元の文書を列ベクトルに対応するような類似度行列を求める．このとき，類似度計算にはコサイン尺度を利用した．ただし，ベクトル間で内積計算を行う必要があるので，行列の要素がどの単語に対応するのかを一致するように対応付けておく．このようにして求めた類似度行列を元の索引語-文書行列の下に追加し，単語と文書間類似度を要素として持つ合成文書ベクトルを作成する．これまでの一連の処理の流れを図 1. に示す．

このとき，追加する類似度行列としてすべての類似度の値を要素に含んだ行列を利用した場合，類似度の低い値がノイズとなり，クラスタリング精度が低下する傾向がある．そのため，クラスタリング対象の各文書に対して類似度の高い値から指定された数のデータを残し，それ以外は 0 とする [7]．類似度の高い要素のみを残すことでノイズの少ない類似度行列を追加することができると考えられる．

2.3　クラスタリング方法

クラスタリングの対象となる文書数 n とし，そこから抽出される単語数を m，関連度を追加するために用いる他の文書集合に含まれる文書数を l とすると，関連度を追加した行列は $(m+l)\times n$ の合成特徴行列となる．この行列に対してクラスタリングを行い，内容の似ている文書をまとめる．

索引語-文書行列に対してクラスタリングを行う手法はこれまでに数多くの研究が進められている．このクラスタリング手法には分割型と凝集型があり，分割型は文書集合からいくつかの内容のまとまりであるクラスタを求め，これをクラスタの変化がなくなるまで繰り返す処理を行う．また，凝集型はひとつの文書をひとつのクラスタに対応させ，似ているふたつの文書をひとつにまとめることを，与えられたクラスタ数になるまで繰り返す処理を行う．

本稿では，以上のように数多く提案されているクラスタリング手法の中で性能の高いツールである cluto を利用した [8]．今回の実験では文書関連性を示す素性を追加した場合のクラスタリング性能を評価することが目的であるため，クラスタリング

手法はひとつに限定して評価することとした．しかし，提案手法が利用するクラスタリング手法と関係があり，高精度な文書クラスタリングが可能となることも考えられるため，いくつかのクラスタリング手法を利用した実験については今後の課題とする．

3. 評価実験

提案した手法の有効性を検証するため，ニュースグループへの投稿記事からなるデータ集合を利用してクラスタリング実験を行った．本稿では，500 件と 1000 件についてクラスタリング実験を行い，どの程度の記事数を手がかりとして追加すると効果的であるのか分析を行う．また，特徴抽出手法としてよく知られる特異値分解を行い，次元数を縮約した場合でのクラスタリング結果との比較を行う．

3.1 データ

本稿では，データ集合として 20 Newsgroups を利用した．これはネットニュースに投稿された 20 ニュースグループの記事を各 1000 件ずつ合計 20000 件の文書を集めたデータ集合である．そのデータ集合から 10 グループを抜き出し，その中から各 50 件を抽出した合計 500 件の文書集合と各 100 件を抽出した合計 1000 件の文書集合をクラスタリング対象とする．追加するデータ集合は全 20 グループから各 50 件ずつ抽出した合計 1000 件とし，類似度行列の各対象文書ベクトルに対して残す類似度の数を上位 100 件から類似度を取り除かない 1000 件まで、100 件ごとに利用する類似度数を変化させてクラスタリング実験を行った．

3.2 評価尺度

クラスタリングの評価尺度には，一般的に用いられているエントロピー (Entoropy) と純度 (Purity) を利用する．エントロピーと純度は以下の式により表される [9].

$$Entorcpy = -\sum_{i=1}^{C} \frac{n_i}{N} \sum_{j=1}^{K} \frac{n_{ij}}{n_i} \log \frac{n_{ij}}{n_i} \tag{1}$$

$$Purity = \sum_{i=1}^{C} \frac{1}{N} \times \max_{j}(n_{ij}) \tag{2}$$

ここで，N はクラスタリング対象となる全データ数，C はクラスタ数，K は評価データのカテゴリ数，n_i はクラスタ C_i に存在するデータ数，n_{ij} はクラスタ C_i に存在するカテゴリ j のデータ数をそれぞれ表す．

4. 実験結果・考察

4.1 文書数 500 でのクラスタリング結果

文書数 500 件の異なる 2 つの文書集合に対してクラスタリング実験を行った結果を表 1 と表 2 に示す．表 1 においては，精度が若干下がっているが，表 2 では精度が向上している．追加して与えるデータに対して関連性が少ない場合は，ノイズの多さによる大幅な精度の減少は少なく，ほぼ同程度のクラスタリング性能となっている．しかし，関連性のある文書が追加された場合は，元の文書集合のクラスタリング精度は追加なしと比較して，純度について最大 7.2% 向上した．この結果により，大規模なデータ集合であっても，他の文書集合との類似度を手がかりとして追加することでクラスタリング性能が向上することを確認することができた．

表 1　追加するデータ数による精度の変化 1(対象 500 記事)

追加数	エントロピー	純度
なし	0.400	0.686
1000	0.415	0.648
900	0.415	0.620
800	0.415	0.620
700	0.415	0.620
600	0.406	0.676
500	0.406	0.676
400	0.415	0.620
300	0.415	0.620
200	0.410	0.672
100	0.410	0.672

表 2　追加するデータ数による精度の変化 2(対象 500 記事)

追加数	エントロピー	純度
なし	0.371	0.698
1000	0.402	0.664
900	0.360	0.690
800	0.306	0.768
700	0.306	0.768
600	0.312	0.760
500	0.312	0.760
400	0.312	0.760
300	0.312	0.760
200	0.312	0.760
100	0.306	0.768

また，クラスタリング対象の各文書に対して，追加する文書集合との類似度が高い上位 100 件から，類似度を取り除かない 1000 件まで，手がかりとして採用する類似度を 100 件ごとに変化させた場合，表 1 では類似度を上位 500〜600 件程度，表 2 では上位 700〜800 件程度を利用することで，クラスタリング精度が最も良くなっている．追加する類似度の数は少なくても，対象文書と関連のある文書が含まれていると精度が良くなるが，文書間類似度をすべて追加すると低い類似度がノイズとなって大きな影響を与えることが分かる．

4.2 文書数 1000 でのクラスタリング結果

対象文書数 1000 件の異なる 2 つの文書集合に対して実験を行った結果を表 3 と表 4 に示す．500 件での結果よりも全体的に低くなるが，500 件での実験同様に手がかりを追加した方が精度の高いクラスタリング結果となった．対象データの数が変化をしても，関連度を手がかりとして追加することでクラスタリング精度が向上することが分かる．

また，各文書に対して追加する類似度の数を変化させた場合についても，500 件での結果と同様に，高いクラスタリング精度が得るためには，上位 500 件程度の類似度を残すことが良い結果となった．すべての類似度を利用した場合，若干精度は良くなってはいるが，ノイズの影響が大きいために，他の追加数と比較すると精度が悪くなっている．

表 3　追加するデータ数による精度の変化 1(対象 1000 記事)

追加数	エントロピー	純度
なし	0.379	0.755
1000	0.378	0.696
900	0.327	0.779
800	0.328	0.774
700	0.363	0.762
600	0.328	0.774
500	0.328	0.774
400	0.328	0.774
300	0.328	0.774
200	0.356	0.731
100	0.327	0.779

表 4　追加するデータ数による精度の変化 2(対象 1000 記事)

追加数	エントロピー	純度
なし	0.315	0.750
1000	0.300	0.764
900	0.308	0.750
800	0.308	0.750
700	0.295	0.767
600	0.284	0.780
500	0.284	0.780
400	0.295	0.767
300	0.322	0.752
200	0.284	0.780
100	0.293	0.777

4.3　次元縮退によるクラスタリングとの比較

提案手法の有効性を幅広く確認するために，特異値分解により次元数を縮退させた場合のクラスタリング実験を行った．10 グループにおける各 100 件の合計 1000 記事を対象に索引語-文書行列を求め，その行列に対して特異値分解を行い，特異値の高い左特異値ベクトルを指定した数だけ抽出し，縮退した次元数の要素を持つ文書ベクトルに変換する．その文書ベクトルに対してクラスタリングを行った結果を表 5 に示す．

表 5 より，索引語の次元数を縮退させてクラスタリングを行った場合，クラスタリング精度は低くなる傾向にある．手がかりとなる索引語の間に存在していた異なり部分が，特異値分解を利用して特異値ベクトルを抽出することにより合成される．このため，文書間において明確な特徴の違いを捉えることが難しくなり，クラスタリング精度が低下する結果となったと考えられる．クラスタリングにおいて，特異値ベクトルのような手がかり間の関連性を利用する場合は，索引語といった元の手がかり情報を保持したまま，手がかり間の関連性は追加をする方が高い精度のクラスタリングを行うことができると考えられる．

5.　おわりに

本稿では，半教師ありクラスタリングよりも制約の緩いクラスタリング手法の開発を目指して，クラスタリング対象となる文書集合の特徴に他の文書集合との文書間関連度も特徴として追加する手法を提案し，大規模な文書集合に対してクラスタリング精度の比較を行った．その結果，大規模なデータ集合であっても，関連性のある文書集合を手がかりとして追加することでクラスタリング性能が向上することを確認することができた．また，類似度行列内の類似度が低い要素はノイズとして悪影響を及ぼすことから，各文書について手がかりとして採用する類似度の数を制限して，類似度の小さい要素は 0 として実験を行った．その結果，追加したデータに含まれるノイズを抑えることができ，クラスタリング精度が向上することが分かった．

今後の課題としては，索引語-文書行列に対して関連度を追加する際，効果的なクラスタリングを行うことができるための数値的なバランスを考慮する必要がある．追加する関連度の範囲が小さい場合は，追加する効果があまりなく，大きすぎると影響が強く出ることが考えられるため，有効な追加方法についてより詳細な分析を行いたい．

表 5　縮退する次元数による精度の変化 (対象 1000 記事)

次元数	エントロピー	純度
1000	0.235	0.791
900	0.314	0.730
800	0.309	0.728
700	0.252	0.786
600	0.326	0.693
500	0.264	0.769
400	0.325	0.729
300	0.314	0.730
200	0.310	0.736
100	0.328	0.712
50	0.355	0.675
10	0.389	0.626
5	0.478	0.544

文　献

[1] H. Cho, I. Dhillon, Y. Guan and S. Sra: “Minimum sum squared residue co-clustering of gene expression data”, Proceedings of the Fourth SIAM International Conference on Data Mining (2004).

[2] I. S. Dhillon, S. Mallela and D. S. Modha: “Information-theoretic co-clustering”, KDD ’03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM, pp. 89–98 (2003).

[3] C. Cortes and V. Vapnik: “Support-vector networks”, Machine Learning, **20**, 3, pp. 273–297 (1995).

[4] S. Chakraborti, R. Mukras, R. Lothian, NirmalieWiratunga, S. Watt and D. Harper: “Supervised latent semantic indexing using adaptive sprinkling”, Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI07), pp. 1582–1587 (2007).

[5] 佐々木, 新納：“文書関連性を素性として追加した文書クラスタリング”, 言語処理学会第 14 回年次大会, pp. 580–583 (2008).

[6] 北, 津田, 獅々堀：“情報検索アルゴリズム”, 共立出版 (2002).

[7] 戸田, 北川, 藤村, 片岡, 奥：“グラフ分析を利用した文書集合からの話題構造マイニング”, 電子情報通信学会論文誌 D, **J90-D(2)**, pp. 292–310 (2007).

[8] G. Karypis: “CLUTO: A Clustering Toolkit, Release 2.1.1”, http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview (2003).

[9] 新納：“R で学ぶクラスタ解析”, オーム社 (2007).