Vol.2009-MPS-76 No.50
Vol.2009-BIO-19 No.50
2009/12/18

文書間の関連強度閾値を活用した再帰的括り出しによる情報分類

佐々木靖彦† 門田和也††

文書情報の分類を行う際に，文書間の共通キーワード数を用いて相互間の関連強度を表現し、その強度における閾値基準を用いた接続関係上の到達可能性からグルーピングを行うことが可能である。さらに，このようなグルーピングにおいて、複数の関連強度閾値を設定する操作を繰り返して再帰的に強関連の大きなグループを括り出すステップを繰り返すことで段階的なグルーピングも可能となる。このような手法は、分類の多様性や文書間の関連に対しての人による着目の遷移を表現可能である。さらに，分類の経過を段階的に把握することが可能となるため，人の理解に向いた分類として期待できる。本稿では、技術文献に対する適用例を通して、文書群が地形構造に類似したフラクタル的構造として分類されることについて、および，その可視化を通しての人の理解への有効性について説明する。

Information Classification by Recursive Clustering Method Utilizing Multiple Thresholds of Relationship Strength between Documents

Yasuhiko Sasaki† and Kazuya Monden††

The strength of mutual relationships between documents can be represented by the number of commonly-included keywords. Also, document groups can be made on the basis of the reachability on the connection graph where their connectivity is decided by the relationship strength against a set of threshold values. Recursively bundling and extracting strongly-related documents by the method above, documents can be gradually grouped on the basis of relationship strength. The proposed method can represent the variety of classification and the transfer of attention from one document to others, and help gradual understanding of the grouping process. Therefore, the method is expected to be a promising candidate for an easy-to-use clustering method. In this paper, we also explain the fractal-like characteristics of the clustered results obtained through actually applying the method to a set of technical documents, and we show the effectiveness of the visualized results in helping understanding.

1. はじめに

情報の適切な分類は，その人による理解という目的の観点から、古くより多くの研究がなされてきた。ここで，分類には，教師つき分類であるカテゴリゼーションと，教師なし分類であるクラスタリングの２つが存在するが、本稿では，このうち，後者を目的とした技術に関して，一つの方法論を提示する。

一般的に，クラスタリングの手法としては幾つかの類型が存在するが、その中でも特に２つがよく知られている。一つは，k-means 法などに代表される分割最適化手法であり、もう一つは、最短距離法などに代表される階層的手法である[1]。前者は，分割の良さを表す評価関数を定め，それが最適値となるように繰り返し配置を行いながら分割を改善するものであり、形式的に数で分けられた分割に向く。一方、後者は，その空間内の距離で近いものを併合しながら段階的にグループ化を進めるものであり、全体の構成把握に向く。

このような一般的なクラスタリングの手法は，各要素を表現する情報が数値属性量であるとき，簡便に適用できる方法であり有用である。しかし，これらの方法は、テキストからなる文書情報のような各要素を表現する情報が数値属性量で表されないような場合には、そのまま単純に適用することができないといった問題が存在する。

そこで、テキストを対象とした文書などの情報に対して適用できるクラスタリングの手法として，文書をなんらかの数値属性量で表現した上で、上述の一般的なクラスタリングの手法を適用することが考えられる[2]。例えば、文書を語の集合として捉えた上で、それら語の重みづけベクトルとして表現した後に、分割最適化手法や階層的手法を適用することができる。重みづけには語の頻度や tf・idf などを用いることができる。

他方、上述したような方法とは幾分異なる類型に属する方法として、対象間のなんらかの関係をグラフ上に表現し，それらのつながりをもとに分類を行うといったものも存在する。むしろ、このようなグラフベースのクラスタリングは、文書はもちろん、それに限らず、人のネットワークなど多様な分野でのクラスタリング手法として、様々なものが提案されている[3-4]。

本稿では，類型上はグラフ活用型の分類方法と言える手法の一つとして、文書相互間の関連強度閾値を複数用いて、その値を操作しながらグループ化と部分分離とを段階的、かつ、再帰的に繰り返す方法を提案する。これは、文書群全体の構造を、地形構造になぞらえて表現するものであり、その抽出結果もまた、地形構造に類似したフラクタル的な性質を持つものとなることを実験結果を踏まえて説明する。さらに，その可視化を通して，人の視覚的理解への有効性にも言及する。

本手法の特徴は、その前提として、人が求める分類にはただ一つだけの正解となる分類が存在するというよりも、むしろ、見方や判断によって複数の分類解があっても

Vol.2009-MPS-76 No.50
Vol.2009-BIO-19 No.50
2009/12/18

よいとする考えに基づいているところにある。ただし、そのような複数の分類結果は、見方や判断をでたらめにとりかえることにより生まれているというものではなく、一貫性を持ったプロセスの下でも必然的に生成されるはずのものである。本手法は、そのような目的を実現するための一貫性あるプロセスの一つを示すものである。

なお本手法では、文書をその代表的な分類の対象としているものの、そこでの文書に対する“キーワード”の部分を、その対象の特徴を表現するなんらかの“特徴属性”と読み替えることにより、人、組織、物などのあらゆるつながりを持つもののクラスタリングや分類に応用することができる。

2. 基本プロセス

まず，提案する方法の基本プロセスを示す。大きくは，以下の三つのステップにより構成される。

（1）文書からのキーワード抽出
（2）文書間のキーワード関連強度計測
（3）関連強度閾値を用いた再帰的括り出し

以下，それぞれのステップを説明するが，3つのステップの中で特に重要な 3 番目のステップを詳述する。

2.1 文書からのキーワード抽出

本ステップでは，まず，文書を構成している語を形態素解析により分離する。その後、全ての語から一般用語を排除した上で，その出現頻度と前後の単語への連接度をベースとして重要度の高いキーワードを抽出する 。例えば，[5-6]に示されるような方法を用いることができる。

一つの文書からの抽出キーワード数はパラメータとなりうるものであるが，その文書をうまく代表したものとなるよう適正な値を決定する必要がある。すなわち，抽出キーワード数の増加は，表現力向上（正の要素）とノイズ混入の問題（負の要素）とのトレードオフの関係を有する。そこで，パラメータの決定では，ノイズ混入が起こらない条件でできるだけ多くのキーワードが入るような値に設定する。

2.2 文書間のキーワード関連強度計測

2つの文書が，キーワードによりどの程度関連しているかを計測する。ここでは，ある一つのキーワードが2つの文書で共有されているかどうかをもとに，離散的に有/無（1/0）を判断するものとする。そして，ある文書ペア内の全ての抽出キーワードに関してその離散値を調べ，それらの総和をとることで文書ペアの関連強度とする[a]。これを，分類の対象としている全ての文書における全ペアに対して実施し，文書間の関連強度結果をマトリクス化しておく。

2.3 関連強度閾値を用いた再帰的括り出し

2.3.1 ステップの概略

本ステップでは，上記で計測したキーワード関連強度に関しての閾値を用いて文書間の接続性の有無を決定した上で，接続を介した到達可能性をもとに文書群をグルーピングする。さらに，分類の終了基準に合致する部分と合致しない部分を分離する。

分類の終了基準に合致した部分は，それでグループ化が完了したものとして終了であるが，他方，分類の終了基準に合致しない部分は，その部分を括り出した上で，それに対して，関連強度閾値を再設定しなおし，本ステップを再帰的に繰り返す。

2.3.2 地形構造概念とのリンク

本ステップは，図1に示されるような陸地と海（水）で表されるような地形構造の概念とリンクして考えると理解がしやすい。

まず，基本となる考えは，2つの文書があったとき，それらの間がより多くのキーワードで関連するものほど同一グループとしての性格が強い，というものである。これは，図1に示すような地形構造で考えると，文書という“島”が，キーワードという“土”で地続きになっている状況である。同図では“垂直に立つ棒状部”が“文書”を、また,“2つの棒状部を水平に接続する平板部”が“文書間の共通キーワード”を表現している。ここで，島は無限の高さを持つものとする。そして，水面がキーワード関連強度の閾値である。

そもそも文書どうしがなんのキーワードによる関連もなければ（島をつなぐ土がなければ），それらは完全に独立していると言える。一方，文書どうしがキーワード関連を持っていて（土でつながっており），水面をある一定の高さに設定すれば，文書群はその閾値基準となるキーワードの関連強度にもとづき分離される。同図（b）では，C1，C2 が一つの島続きとなり，C3，C4，C5 が別の一つの島続きとなり，2つのグループに分離されることがわかる。これは，実際の地形構造でも水面以上の土がなければ,島どうしは互いに行き来ができなくなるため分離されてしまう状況と同じである。すなわち，到達可能性によりグループ分けしていることを意味する。

このように、より多くのキーワード（多くの土）で接続される文書群（島々）ほど，

a Jaccard 係数等と同様の概念である

Vol.2009-MPS-76 No.50
Vol.2009-BIO-19 No.50
2009/12/18

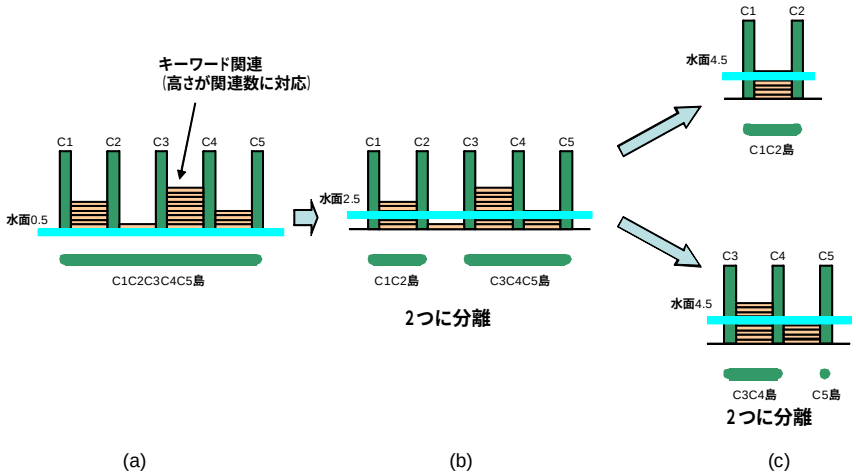

図 1　複数の関連強度閾値を用いた再帰的括り出しによる分類

水面が高くなっても，なお一つの塊（列島）としての性格を残しやすくなる。したがって，このような連結した文書群は互いに関連も深く，本来，同じグループに属するものとして分類することが妥当であると言える。逆に言えば，ある特定のキーワード関連強度閾値（水面）を設定すれば，関連の薄い文書群どうしを互いに分離することが可能になる。

なお，このような接続関係は必ずしもそうする必要はないが，グラフ化して表現することが可能である。文書をノードとして表現し，関連強度閾値より大きい関連がある場合だけ，ノード間のエッジとして接続することができる。

2.3.3　単一の関連強度閾値による分類

上述のような方法を用いて分類を行う際に，最も容易に考えつく方法は，一つの適切な関連強度閾値を設定することである（図 1(b)）。もし，分類の目的から分類結果のグループ数が予め与えられている場合には，グループ数が目的の値に近い閾値を採用する。このようにして文書を分類すれば，関連の強弱を一義的に決めることが-できるために，ある意味便利である。

しかしそれは，単に文書集合を一つの視点で分離できたというだけで終わってしまっており，文書群をいろいろな関連強度の視点から見たような、全体として俯瞰するといったものになっていない。すなわち，閾値以下の強度が低い関連情報は欠落してしまっており、分類に生かされなくなってしまっている。また、グループ数が所望の値に近いという観点でも、複数通りの結果が存在する可能性があり、どの結果を採用するのがよいかは必ずしも簡単には判断できない。

2.3.4 複数の関連強度閾値による分類（関連強度の閾値を用いた再帰的括り出し）

そこで、無理に一つの閾値でのみ分類するというアプローチを見直し、複数の閾値を用いて分類するというアプローチを試みる。このような視点を持つ方が，文書群全体が持つ複雑な関係を捉える際に，できるだけ多くの関連情報を取り込んで分類していくことが可能になるからである。

具体的には，関連強度閾値を段階的に変化させながら，文書群全体を少しずつ分離していく。まず，低い関連強度閾値に初期基準を設定する。次に，以下のサブステップを再帰的に繰り返す（図 1 (c)）。

（S1）与えられた関連強度閾値により互いに接続される（到達可能な）文書塊をグループ化する
（S2）別途設定したグループ当たり上限文書数以下の数の文書を有するグループは分類終了であるが, グループ当たり上限文書数より大きい数の文書を有するグループは, それを全体から括り出し分離する
（S3）関連強度閾値を現在値より高い値に再設定する
（S4）(S2)で括り出されたグループのそれぞれに対し，再度(S3)で決めた閾値を用いて(S1)からの手続きを実施する

このように，抽出された文書グループに対して再帰的にグループ化と分離抽出を繰り返していくことで，最初は大きな塊として分類された文書群が，少しずつほぐされるように分離されていく。最終的に，ある関連強度閾値でグループ化された全てのグループの文書数がグループ当たり上限文書数より小さくなった時点で終了する。

2.3.5 グループ当たり上限文書数

前節において，文書群の括り出しを行い，それに対して再帰的に処理を継続するかどうかの決定を行う際に，グループ当たり上限文書数という基準を用いた。一般的にこの基準は、分類の結果を利用する際の目的に合わせて適切に設定することがよいと考えられる。しかしながら、このような基準を設けることの典型的な理由の一つに、1グループあたりの文書数が大きくなり過ぎると人の理解に向かなくなるため，その大きさを制限したいということがある。

Vol.2009-MPS-76 No.50
Vol.2009-BIO-19 No.50
2009/12/18

例えば，分類された結果の各グループが持つ意味を人が理解したり考察したりするといった利用形態を考えた場合，その数を 5〜9 といった値に設定することができる。これは，人の短期記憶が 7±2 と言われることからも，人の認識の過程でグループとしてまとめて記憶できる範囲の数だからである[7]。

2.3.6 関連強度閾値の段階的設定

2.3.4 節で述べたように，本方法では，関連強度閾値は複数の値をとることになる。今，n 回目の閾値を T_n とすることにすれば，その値の設定方法には幾つかの方法が考えられる。最初の閾値 T_1 は，そもそもグループ化する際の最低基準を与えるようなものであるから，幾つかの値を試行して極端に類似性の低い文書どうしがグループにならないような高さに設定する。次に，２回目以降の T_n の設定については，以下のような方法が考えられる。

（a） $T_n = T_1 * n$
（b） $T_n - T_{n-1} = k * (T_{n-1} - T_{n-2})$, ただし $T_0 = 0$

(a)は第 1 回目の閾値の定数倍を順に使っていくものであり最も単純な方法である。一方，（b）は，T_n に関してもう少し一般化した形であって、括り出されたグループに対して，より厳しい類似性，あるいは，より緩い類似性を要求して分類を進める方法である。k が１より大きければより厳しい類似性を，k が 1 より小さければより緩い類似性を求めることになる。k が１に等しいときは、(a)に相当する。

なお、T_n の設定をもっと柔軟に行うことも可能である。例えば、後述するような強関連の塊が小さいクラスのグループ数が、１回あたりの操作で一定の数だけ抽出されるように、T_n の値を順次増加させながら選択していくといったことができる。さらには、上述の方法とは逆に、関連強度閾値を大きい側から減少させていくといったことも可能である。これは、関連の強い部分だけを抽出すれば分類の目的が達成されるといったケースに有効な手法である。

以上のように、関連強度閾値の設定方法としていろいろな方法を用いることができるが、やはり、あいまいな基準ではなく、分類者の目的にそった一貫性あるポリシーの下に設定方法を選択することが望ましい。

2.3.7 分類多様性の表現

複数の関連強度閾値をどのような値に設定するか、また、グループ当たり上限対象数（文書数）をどのような値に設定するかで分類の結果が異なる。これは、明示化された基準を用いることにより、分類の多様性を表現できることを意味している。

関連強度閾値は、どの程度のつながりを有効とみなすかを与える判断基準であり、例えば、人の着目がどの程度のつながりの強さがあれば対象間を遷移することができるかといったことに対応する。一方、グループ当たり上限対象数は、グループとしてどの程度の大きさまでを許容するかを与える判断基準であり、例えば、人の認知可能な数の限界を与えるといったことに対応する。

3. 実験

3.1 技術文献に対する分類実験

提案手法を技術文書群に対する分類に適用実験した。具体的には，ACM 学会（Association of Computing Machinery）誌に掲載された技術文献を取り出し，これらに対して同手法を適用した。対象文書群の諸元を表１に示す。

表 1　分類の実験対象

項目	内容
対象文書の種類	技術文献 ACM magazines
文献数	2345
単文書当たり抽出キーワード数	50
延べｷｰﾜｰﾄﾞ数	117250

3.2 ３つのクラス

分類の結果は，各階層ごとに以下の３つのクラスのいずれかになる。ここで、便宜上，名称は地形構造へのアナロジから理解しやすいものとしている。

１）強関連の塊が大きいクラス（大陸クラス）
２）強関連の塊が小さいクラス（列島クラス）
３）関連が弱く孤立するクラス（孤島クラス）

大陸クラスは，ある階層において，グループ化された後のグループ内文書数がグループ当たり上限文書数を超えているものを意味する。また，列島クラスは，グループ化された後のグループ内文書数がグループ当たり上限文書数以下のものを意味する。孤島クラスは，他の文書との関連が弱いために関連強度が閾値以下となって孤立した

Vol.2009-MPS-76 No.50
Vol.2009-BIO-19 No.50
2009/12/18

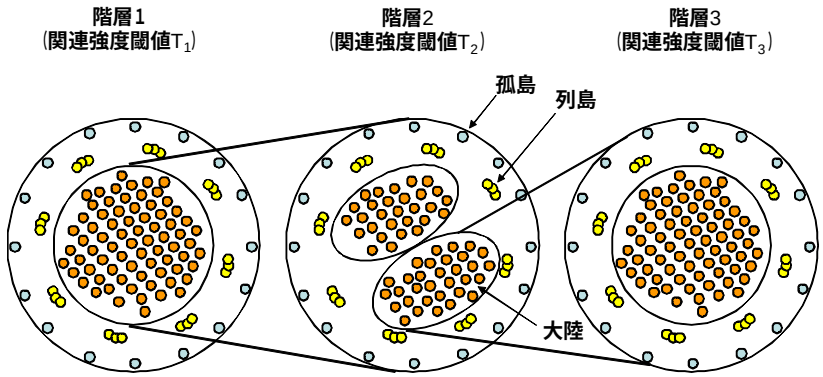

図 2　地形構造に類似した分類結果

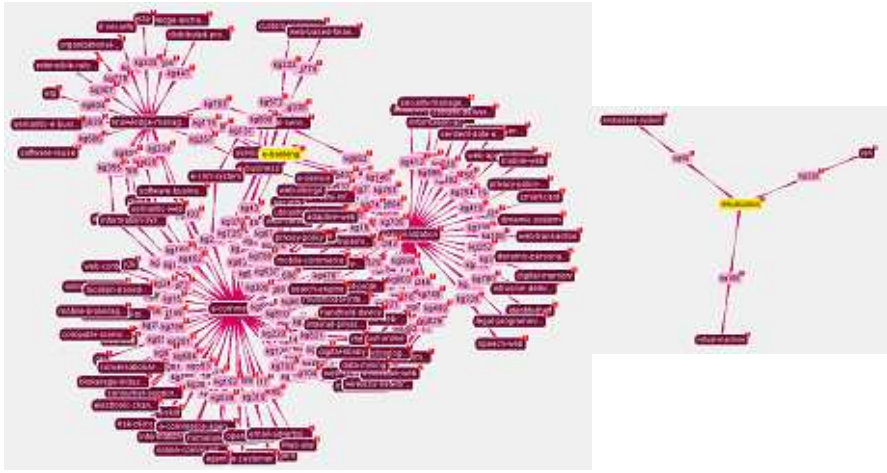

図 3　分類結果の可視化

ものを意味する。

表２に，分類の結果例を示す。ここでは、関連強度閾値の段階的設定の方法として、最も単純な方法として 2.3.6 節における（a）の方法を用いている。また，図２に，分類結果の概念表示を示す。文書情報空間は，あたかも地形構造に類似したフラクタル的な構造となる。本例においては、閾値が低い第１階層においては、相互に関連の強い大きいグループ（大陸クラス）が一つ、相互に関連が強い小さいグループ（列島クラス）が数十個以下、他との関連が弱いグループ（孤島クラス）が約千存在することがわかる。閾値基準を上昇させて階層を上げるにつれて，下位階層側で大陸クラス中にあった文書（群）のある部分が分離されて上位階層側での列島クラスや孤島クラスへと分類されていく。上位の階層においても、大陸、列島、孤島といった構造は依然存在し、それらの数の間の大小関係も類似したものとなる。

3.3 分類結果の可視化

各グループの接続関係を可視化するために，文書どうしが関連している様子をトピックマップ[8]を用いて記述し，その表示を行った。トピックマップは、“トピック”、“関連”、“出現”を要素とする知識表現手段であり、トピック間の関係をわかりやすく表現することが可能である。また、ISO の標準としても採用されていることもあり、いったん同表記に対応させれば、その後のいろいろな表示ツールへの展開が可能である。

大陸クラスと列島クラスに属する文書グループの例を図３に示す。本ケースからもわかるように，各文書に対応する技術どうしがどのように関連しながら塊を成すかについて、その形状から視覚的に理解することが可能である。例えば、大陸クラスの中でもどの文書が広いつながりを有するかといったことや、列島クラスの中心にあるものはどれかといったことを容易に判断できる。また，各エッジには，対応するキーワード群が付加されているため，文書群がどのようなキーワードを介して相互関連しているのかをすぐに把握することも可能である。これは、分類結果の意味を考える際に非常に有効な手助けとなる。

Vol.2009-MPS-76 No.50
Vol.2009-BIO-19 No.50
2009/12/18

表 2　階層別に見た３クラスの諸元

階層	第 1	第 2	第 3
関連強度閾値	6	12	18
大陸クラス グループ数 （内部文書数）	1(400)	2(69)	1(11)
列島クラス グループ数 （内部文書数）	26(63)	11(32)	6(13)
孤島クラス グループ数	1018	299	45

4. おわりに

情報分類の手段として，文書間の関連強度閾値を複数用いて再帰的括り出しを行う分類手法を提案した。同手法を用いれば，基準や判断に応じて変化する分類の多様性を表現することが可能となる。また、キーワード視点による人の着目の分類対象間の遷移を考慮したグルーピングなども可能になる。段階的なグルーピングを通した地形構造類似の様子を可視化することにより，全体構造の把握も容易化されるため，人の理解に向いた分類方法として期待できるものと考える。

参考文献

1) A. Hinneburg, D. Keim, “Clustering techniques for large data sets,” Tutorial notes, Principles and practice of knowledge discovery in databases 2000.

2) 岸田和明, “大規模文書集合に対して階層的クラスタ分析法を適用するための単連結法アルゴリズム”, Library and information science, No. 47, pp27-37, 2002.

3) A. Barabasi, “LINKED: The new science of networks,” Basic Books, 2002.

4) V. Ganti, J. Gehrke, R. Ramakrishnan, “CACTUS—clustering categorical data using summaries,” Proceedings of the fifth international conference on knowledge discovery and data mining, pp73-83, 1999.

5) E. Brill, “A simple rule-based part of speech tagger,” Proceedings of the third conference on applied natural language processing, 1992.

6) 中川裕志、森辰則、湯本紘彰 :"出現頻度と連接頻頻度に基づく専門用語抽出", 自然言語処理, Vol.10 No.1, pp.27-45, 2003.

7) 田中 啓治編, “認識と行動の脳科学”, 東京大学出版会, 2008.

8) Topic Map, ISO/IEC 13250:2002.

※1ページ目　著者所属先については、下記のとおりとなります。

†日立製作所　中央研究所
Hitachi, Ltd., Central Research Laboratory
††日立製作所　システム開発研究所
Hitachi, Ltd., Systems Development Laboratory