様式 C-19

科学研究費補助金研究成果報告書

平成 23 年 5 月 30 日現在

機関番号：１２１０１
研究種目：基盤研究(C)
研究期間：2008　～　2010
課題番号：２０５００１２４
研究課題名（和文）　縮約類似度行列を用いた大規模文書データに対するスペクトラルクラスタリング
研究課題名（英文）　Spectral clustering for large document data using the reduced similarity matrix
研究代表者
新納　浩幸（SHINNOU　HIROYUKI）
茨城大学・工学部・准教授
研究者番号：１０２５０９８７

研究成果の概要（和文）：本研究では大規模文書クラスタリングにスペクトラルクラスタリングを用いる手法を開発した。基本的には大規模データを k-means で予め小規模クラスタに分割し、そこから信頼度の高いデータを抽出し、それらデータに対して類似度行列を作る。作成された類似度行列は縮約されているので、スペクトラルクラスタリングが実行できる。クラスタリングの更なる精度向上のために、精緻な名詞間距離の測定法や、文書間の距離学習法の開発も行った。

研究成果の概要(英文)：In this research, I proposed the spectral clustering method for large document data. First, large document data is divided into small clusters by k-means. then some reliable data are picked up each clusters. We construct a similarity matrix from these reliable data. This matrix is reduced, so we can use the spectral clustering for it. Furthermore, in order to improve the precision of clustering, I researched the distance measurement of two nouns, and distance learning for documents.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	900,000	270,000	1,170,000
２００９年度	1,300,000	390,000	1,690,000
２０１０年度	1,100,000	330,000	1,430,000
年度			
年度			
総　計	3,300,000	990,000	4,290,000

研究分野：自然言語処理
科研費の分科・細目：情報学・知能情報学
キーワード：縮約類似度行列、スペクトラルクラスタリング、文書クラスタリング、距離学習、最大マージン化最近傍法

１．研究開始当初の背景

文書クラスタリングとは、入力される文書のセットを、トピックの類似性からいくつかのグループに分割する処理である。情報検索やテキストマイニングで利用される要素技術であり、その重要性は高い。近年、データが大規模化しており、文書クラスタリングにおいても大量の文書を対象としなくてはならない場面が多い。このような背景から、大規模文書データセットに対する精度の高い文書クラスタリング手法が望まれている。

一方、スペクトラルクラスタリングはグラフ理論を応用したクラスタリング手法であり、その精度の高さから近年活発に研究が行われている。ただしスペクトラルクラスタリングはデータ数の自乗のサイズの固有値問題を解く必要があり、大規模データセットに

対しては直接適用することが不可能である。この問題の解決のために、本研究では、データセットに対する類似度行列を縮約する手法を開発する。

２．研究の目的

本研究の目的は、スペクトラルクラスタリングを大規模文書データセットに対して適用することで精度の高い文書クラスタリング結果を得ることである。

本研究のアプローチは小規模クラスタ生成の一種である。ただし小規模クラスタを生成した後に、単純にそれらクラスタを１点で代表させて、クラスタリングを行うというアプローチではなく、各小規模クラスタに対してそのクラスタの代表点を作成する。次にそこで中規模の個数からなる代表点に対して、k-means 等の簡易なクラスタリングを行う。次に得られた各クラスタから信頼性の高いデータを取り出し、それらを各クラスタの Committee とする。

ここで各 Committee を １ 点で表し、Committee に属さない代表点と合わせて、スペクトラルクラスタリングの用途に特化した形の類似度行列を作成する。これが本研究での縮約類似度行列である。これを基にスペクトラルクラスタリングを実行し、得られたクラスタ内にある代表点をもとのデータに復元することで、最終的なクラスタリング結果を得る。

以上より、本研究では以下の 4 点を解決することが目的となる。(1) 大規模データを小規模クラスタに分割する、(2) 小規模クラスタのクラスタリング方法、(3) 各クラスタからの Committee の作成方法、(4) Committee 群からの縮約類似度行列の作成方法。

３．研究の方法

本研究では以下の 4 点の研究開発を行う。(1) 大規模データを小規模クラスタに分割する方法。(2) 小規模クラスタのクラスタリング方法。(3) 各クラスタからの Committee の作成方法。(4) Committee 群からの縮約類似度行列の作成方法。

まず (3)と(4)を解決する。具体的には、既存のデータセットの各データが小規模クラスタの代表点だと考える。これによって (1) の処理が仮想的に行えたと見なせる。次に既存のデータセットを k-means 等でクラスタリングすることで(2) の処理結果も得ることができる。ただし (2) を k-means 等で単純に処理するのは効率が悪いので、これは (3) と (4) の研究に取りかかるための処置であることを注記しておく。

(3) に関しては 2 つのアプローチを試みる。1 つは各クラスタに対してその重心を求め、クラスタ内の各データとその重心までの距離を測り、この距離に基づいて Committee を作成するアプローチである。距離によって Committee に属するか属さないかを判定するが、その際の閾値の設定が問題である。この設定には様々な統計的手法を取り入れることができるので、各手法を試しながら最も効果的な方法を探ってゆく。もう 1 つのアプローチは各クラスタのデータを訓練データと考えて、帰納学習の手法を用いて分類器を作成し、その分類器によって Committee を作成するアプローチである。具体的にはそのクラスタに真に属する確率を調べ、ある確率以上のデータを選出することで Committee を作成する。この場合の問題は計算の効率である。SVM などでは分類器の学習に多大な時間を要する。ここでは Naive Bayes の利用を計画している。これは文書データに対して親和性が高い、分類器学習の計算コストが低い、分類器は確率を算出できるなどの点で、本手法に適していると考えられる。また確率の閾値の設定の問題も残る。当初は経験的に決めて、実験を通して知見を得る。

本手法では Committee 内のデータが 1 点に代表されるので、Committee 内の誤りは以後の処理で回復できない。このため Committee に属するか属さないかの判断には高い精度が求められる。しかし精度の高い判断を要求しすぎると Committee のサイズが小さくなる。この場合、データの縮約の度合いが小さくなり、計算の負荷が高まる。つまり計算の負荷と精度とのトレードオフの関係があるので、その点での理論的な解析も進める。

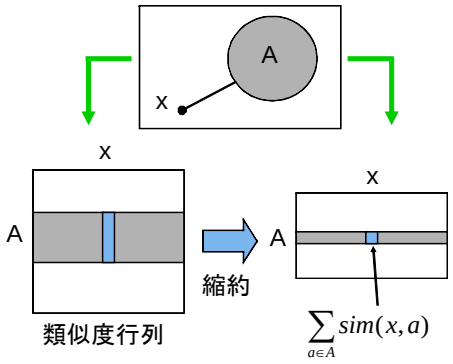

図 1

(4) が本研究の核と言える。まず理想的に Committee 内のデータが真に同じクラスタに属する場合、Committee A と任意の点 x との類似度は A 内の各点 a と x との類似度の総和によって定義し、縮約類似度行列を作成すればよい（図 1 参照）。この場合、もとの全てのデータに対する類似度行列を使っ

たスペクトラルクラスタリングの結果と、提案する縮約類似度行列を使ったスペクトラルクラスタリングの結果は同一になる。この点の証明も行う。更に縮約類似度行列の作成に要する計算時間も調査する。

クラスタリングシステムへ入力されるデータの形式は、通常、ベクトル表現されたデータの集合であり、上記の縮約類似度行列の作成コストは高い。そのために、上記の縮約類似度行列を近似する。近似式はアドホックに探ってゆくしかないが、タスクが文書クラスタリングであるため、ベクトルが高次元かつスパースとなっている。そこで次元数と非ゼロ要素の割合を利用して近似式を作成する。具体的には次元数と非ゼロ要素の割合から、データの degree （データとその他のデータとの類似度の和）の期待値が推定できる。縮約類似度行列はこの degree を基に作成できるので、degree の推定により近似が可能である。

また現実的には Committee 内のデータが真に同じクラスタに属するという仮定は成り立たないので、誤りが多少含まれていることを前提として近似式を作る方が精度が上がる。その点を考慮した近似式を提案する。ここにはリスク最小化理論が応用できる。

(1) に関しては、転置索引語ファイルを利用する。これはデータの各次元の非ゼロ要素の集合に注目する手法である。各次元に対してその次元が非ゼロとなるデータだけに注目すれば、データ間の類似度が 0 になる計算を避けることができる。大規模文書クラスタリングでは大きな効率化になる。ここでは転置索引語ファイルを利用して、データ間の類似度が 0 より大きくなるデータ対に対して類似度を測る。またデータ対の重複した計算は避ける。得られたデータ対と類似度の集合を類似度でソートし、適当なクラスタ数になるまで順次データを統合させてゆく。これによって大規模データを小規模クラスタに分割することができる。ただしこの方法ではデータがどのクラスタにも属さない場合が生じる。しかし提案する縮約類似度行列は、(1) の処理でこのようなケースが生じても問題ないことを注記しておく。

(2) に関しては、(1) のデータの統合処理を継続すれば、そのまま得られる。ただし (1) の処理は粗く、類似度の大きな部分では妥当な結果となるが、類似度があまり大きくない部分まで処理を継続すると、精度が下がってしまう。そのため、ある程度精度の高いクラスタリングを行う必要がある。ここではクラスタリングツールの CLUTO で提供されている API を利用する。CLUTO は k-way clustering と呼ばれる手法を用いており、k-means よりも経験的に精度が高い。しかも CLUTO は類似度行列を入力としたクラスタリングが可能である。クラスタ間の類似度をクラスタ内のデータ間の最大の類似度で定義した場合、(1) の処理を通して、小規模クラスタに対する類似度行列が作成できていると見なせる。それをそのまま CLUTO の API に渡すことで小規模クラスタに対するクラスタリング結果が得られる。

４．研究成果

本研究の目的である (1) と (2) に関しては既存の手法とツールを用いることで実現できた。(3) の各クラスタからの Committee の作成方法については各クラスタの重心からの距離を利用した。重心から近いデータは、そのクラスタに真に属すると考え、Committee のメンバと見なすことにした。どの程度近ければ Committee と見なすかは、最終的な精度と計算時間の兼ね合いとなる。実験によって、クラスター内のデータを重心から近い順に 2 割取り出し、これをクラスターの Committee とすれば、比較的妥当な結果が得られることを確認した。また (4) の Committee 群からの縮約類似度行列の作成方法については以下の方法を確立した。まず上記処理によって作成された Committee は K 個（小規模クラスターのクラスタ数）あるが、それぞれを 1 点に縮約し、新たなデータセットを作成し、Committee 群が作成した後、各 Committee を縮約した縮約類似度行列を作成するための式を示した。Committee 内に誤りが含まれなければ、導出した縮約類似度行列と、縮約しない完全な類似度行列は、スペクトラルクラスタリングに同じクラスタリング結果が得られることを示した。ただし現実的には Committee 内に誤りが含まれるために、示した式を使うと誤りが増大してしまう。そのために現実的に利用可能な近似式も示した。

実験は CLUTO のサイトで提供されている 7 つのデータセット(tr12, tr31, mm, la12, sports, ohscal, cacmcisi)を用いた。作成された縮約類似度行列を用いて、スペクトラルクラスタリングを行い、最終的に得られたクラスタリング結果をエントロピーと純度で評価した結果、オリジナルのスペクトラルクラスタリングの結果よりも精度は悪いが、k-means よりも精度は改善されることを示した。

クラスタリングの更なる精度向上のために、文書間距離の正確な測定法がある。文書クラスタリングの場合、クラスタリング手法よりも文書間距離の設定が精度に影響する。ここでは 2 つのアプローチを試した。1 つは Web ディレクトリを用いて名詞間距離を精緻に求め、それらを利用して文書間距離をより適切に設定する手法であり、もう 1 つは少

量の教師データを与える手法である。後者の手法では、複数のペアの文書間が同じカテゴリに属する文書かどうかのラベルを与え、それを教師データとすることで、文書間の距離をクラスタリングにとって最適になるように学習する距離学習の手法を利用した。いくつかの距離学習手法を比較実験し、最大マージン化最近傍法が本タスクにおいて最も効果があることを確認した。またこの距離学習の手法の有効性を語義識別問題により確認した。最終的には、大規模データを k-means により小規模クラスタに分割し、各クラスタの重心と最も近いデータから Committee を作成し、いくつかの Committee 間に同じカテゴリかどうかのラベルを与え、そこから Committee 間の距離を最大マージン化最近傍法により学習し、それを基に縮約類似度行列を作成した。その行列を利用してスペクトラルクラスタリングを行い、当初の大規模データのクラスタリングが行えた。結果は、直接 k-means でクラスタリングを行うよりも精度が向上した。

今後の課題としては縮約の度合いの調整がある。縮約の度合いと精度との関係を調べると、縮約の度合いが大きいほど精度が悪くなることを確認した。ただし縮約の度合いが小さいと、計算コストが高くなるために、現実的な妥当な縮約の度合いを設定することが課題である。

また本手法はクラスタリング結果の改善手法という位置づけで本手法を捉えることもできる。まず最初のクラスタリングで正解と思われるものを固定して、分類が曖昧になるデータに対してだけ、高精度のクラスタリング手法を行っている形になる。そのためここでのアプローチは CBC (Clustering by Committee) の一種とも考えられる。CBC では Committee と呼ばれる各クラスタの核となるデータセットを作り、各データは Committee との距離によってどの Committee に属するかを判定することでクラスタリングを行う。各データの Committee への割り当ては、単連結法（最短距離法）によるクラスタリングと見なせる。一方、本手法においては k-means で得られたクラスタ中の信頼性のある集合を Committee とし、各データの振り分け部分にスペクトラルクラスタリングを利用している。また混合分布モデルを用いたクラスタリングにおいても、分散共分散行列のモデルを推定するために、最初にクラスタリングを行う場合があり、これもクラスタリング結果の改善手法と見なせる。本手法をクラスタリング結果の改善手法と見なした場合の、様々な改善手法と併用して利用することも今後の課題である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計 15 件）

① 佐々木稔, 新納浩幸,``距離学習に基づく語義識別の性能分析″, 言語処理学会第 17 回年次大会, E2-7 ,2011 年 3 月 11 日, 豊橋.

② Minoru Sasaki and Hiroyuki Shinnou, Document Clustering Using Semantic Relationship Between Target Documents And Related Documents″, The Fourth International Conference on Advances in Semantic Processing, pp.91-95, 2010 年 10 月 25 日, フィレンツェ(イタリア)

③ Hiroyuki Shinnou and Minoru Sasaki, ``Detection of Peculiar Examples using LOF and One Class SVM″, LREC-2010, 2010 年 5 月 20 日, バレッタ（マルタ共和国)

④ 茂木哲矢, 新納浩幸, 佐々木稔, ``文書クラスタリングを対象とした Weighted Kernel K-means の初期値設定法″, 言語処理学会第 15 回年次大会, D4-5, pp.693-696, 2009 年 3 月 5 日, 鳥取

⑤ Hiroyuki Shinnou and Minoru Sasaki, ``Spectral Clustering for a Large Data Set by Reducing the Similarity Matrix Size'', LREC-2008, 2008年5月28日, マラケッシュ（モロッコ）

⑥ Hiroyuki Shinnou and Minoru Sasaki, ``Ping-pong Document Clustering using NMF and Linkage-Based Refinement'', LREC-2008 , 2008 年 5 月 28 日, マラケッシュ（モロッコ）

６．研究組織

(1)研究代表者

新納 浩幸（SHINNOU HIROYUKI）

茨城大学・工学部・准教授

研究者番号：10250987