1A1-1

制約付きクラスタリングを用いた論文分類

Topic Extraction from Scientific Paper Database

榊　剛史[*1]　松尾　豊[*2]　石塚　満[*1]
Takeshi Sakaki　Yutaka Matsuo　Mitsuru Ishizuka

[*1]東京大学　情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo

[*2]産業総合研究所　サイバーアシスト研究センター
Cyber Assist Research Center, National Institute of Advanced Industrial Science and Technology

In a current flood of computerized academic information, the categorization them has been more and more important. However, categories in academic fields has been changing and it is difficult to apply current methods to scientific article categorization.

In this paper, we have proposed new methods to categorize papers, considering time series variation of academic categories. We have demonstrated that our method is more effective to categorize scientific papers than current methods.

1.　はじめに

論文集合を整理し、構造化する技術は多くの研究で行われている [4, 2])。本研究では、論文集合を各論文分野のカテゴリに分類することに注目し、論文分類の新しい手法を提案する。

文書集合の分類については、今まで様々な文書分類や文書クラスタリングを初めとする様々な分類手法が提案されてきている。しかし、これらのような従来の分類手法が対象としてきた文書集合と比べ、論文集合には大きく異なる点がある。それは、分類すべきカテゴリが時間と共に変化していくという点である。例えば、近年の「人工知能」に関する分野であれば、以前は存在しなかった「Web マイニング」や「複雑ネットワーク」といった分野が出現している。しかし、既存手法を用いてこのようなカテゴリの時系列変化を俯瞰的に捉えることは難しい。

そこで本研究では、従来のアプローチを組み合わせ、お互いの欠点を補完することで、論文分野の時間的な変化を捉えることのできる手法を提案する。特に、制約付きクラスタリングという新たなクラスタリング手法が大きな特徴である。

2.　時系列的な観点から見た従来手法

ある論文集合を分類することを考えるとき、我々は２つの情報を考慮していると考えられる。一つは現在あるカテゴリの特徴、もう一つは分類対象となる文書集合の関係性である。分類対象をなるべく現在あるカテゴリに分類することを考えるが、その論文集合に偏りがある場合、新しいカテゴリを作ることもあり得る。

このように人間が捉えるカテゴリというものは時系列的に変化していくものである。しかし、時系列的な分析という観点から考えた場合、従来の文書分類手法には大きな問題点がある。

2.1　時系列的な観点から見た文書分類

文書分類は、あらかじめ与えられたカテゴリに文書を分類する手法である [5],。この手法では、図１のようにあらかじめ分野カテゴリごとに分類された論文集合を用いて分類器を学習し、それを用いて新規論文を各カテゴリに分類すると考えられる。このプロセスを時系列的な観点から捉えれば、過去の時点のデータを用いて現在のデータを解析していると言える。過去の時点のデータとは、あらかじめ分類された論文集合であり、現在のデータとは新規論文である。

このように論文のカテゴリ分類に文書分類の手法を用いる場合、過去の時点から見た現在の状況を反映できる反面、現在の論文集合のみで考えた状態を反映する事ができないのである。

2.2　時系列的な観点から見た文書クラスタリング

文書クラスタリングは、類似性や関連性といったある文書集合全体の関係性を手がかりとして、関係の強い論文が１つのグループ（クラスタ）にまとまるように、全体を分割する手法である [6, 1]。これを論文集合の分類にあてはめると、新規論文集合の中で関係の強い論文同士を一つのクラスタにまとめているといえる。時系列的な観点から捉えれば、現在の状況のみで論文を整理していると考えられる。そのため、過去の時点でのカテゴリを考慮することができないため、図２のように各年によって全く異なる結果が得られてしまう可能性が高い。

このように論文のカテゴリ分類に文書クラスタリングの手法を用いる場合、現在の論文集合のみで考えた状態を反映できる反面、過去の時点から見た現在の状況を反映する事ができないのである。

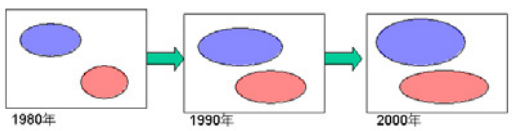

図 1: 文書分類の問題点

3.　提案手法の概要

3.1　基本的な考え方

本論文では、時系列的な変化を考慮した論文のカテゴリへの分類手法を提案する。

連絡先: 榊　剛史，東京大学大学院　情報理工学系研究科　石塚研究室，〒 113-8656 東京都 文京区 本郷 7-3-1，03-5841-6775，sakaki@miv.t.u-tokyo.ac.jp

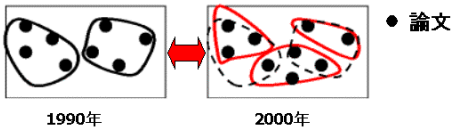

図 2: 文書クラスタリングの問題点

論文集合のカテゴリ分類を考えると、文書分類は、過去のデータを生かしているもの現在の論文集合の状態を反映していないことが問題である。また、文書クラスタリングでは、逆に現在のデータを生かしているものの、過去のデータを無視していることが問題点となっている。そこで、本研究では、2 つのデータ系列を足し合わせ、そのデータ上でクラスタリングを行うことで、両方の影響を加味したクラスタリングを行うことを目指す。

まず、ある論文集合 D があったとき、それらを既存のカテゴリに分類する。このカテゴリ分類を同カテゴリ行列 C によって表す。同カテゴリ行列 C とは、各論文を要素とする n^2 行列であり、2 つの論文が同じカテゴリに所属しているとき、その 2 つの論文にあたる要素が 1 となった行列である。次に、論文間の類似度に基づき類似度行列 S を構成する。S は各 2 論文の類似度を表す隣接行列である。

以上 2 つの行列を C, S とすると、図 3 のように C, S を合わせることで制約付きネットワークを構成する。制約付きネットワークとは、図 3 のようにこれは類似度の逆数を距離とする論文ネットワークに対し、さらに各 2 論文が同じカテゴリに含まれるか否かで、制約を付加したネットワークである。ここでは、類似度に対し、さらに同じカテゴリに含まれることを数値として付加する。

このとき制約付きネットワークの隣接行列 R は次式で表されるものと定義する。

$$R = (E - r)S + rC \tag{1}$$

E は単位行列を表す。r は S による制約の強さを表すパラメータであり、制約行列と呼ぶ。制約付きネットワークを構築することで、2 つの系統のネットワークの影響を考慮することができる。したがって、このような制約付きネットワーク上でクラスタリングを行うと、2 系統のデータの影響を受けたクラスタ結果が得られると考えられる。

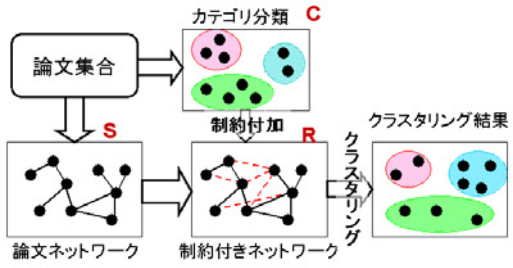

図 3: 制約付きクラスタリング

3.2 提案手法の流れ

本研究では、類似度や引用関係によるネットワークに対して、カテゴリ分類によるネットワークによる制約をつけることで、カテゴリ分類による影響を加味した文書クラスタリングを行う。

手法の流れは以下に示すとおり。

1. 類似度や引用関係などのデータから、論文ネットワークを構築する。
2. 次に何らかの手法で論文のカテゴリ分類を行う。
3. 同じカテゴリに分類された論文同士を結んだ、カテゴリネットワークを構築する。
4. 2 つのネットワークから、制約付きネットワークを構築する。
5. 制約付きネットワーク上でクラスタリングを行う。
6. 既存カテゴリとクラスタ結果の対応づけを行う。

3.3 提案手法の実装

論文ネットワークの構築

本論文では類似度を手がかりとしてネットワークを構築する。類似度の算出方法としては、最も一般的な手法の 1 つである、文書ベクトルを用いたベクトル空間法を用いる。ただし、論文の本文全てを用いる場合、文書ベクトルのノイズが多くなってしまうので、今回はアブストラクトを用いる。各論文のごとに、アブストラクトに出現する語の tfidf 値を要素とする文書ベクトルを定義する。式 (2) のような文書ベクトルの cosine 値を各論文間の類似度とする。

$$cosine = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} \tag{2}$$

そして、論文をノード、論文間の類似度をエッジの重みとして論文ネットワークを構築する。

既存カテゴリへの分類

論文を自動的にカテゴリ分類する手法は、今まで数多く研究されている。しかし、ほとんどの手法が人手で分類したデータを元に分類器を作成する教師付き機械学習を用いた手法であり、あくまで人手による分類が 100%の正解としている。そこで本研究では、人手によるカテゴリ分類をそのまま用いる。

制約付きクラスタリング

類似度による論文ネットワークに対して、カテゴリ分類によるネットワークを用いて制約を付加することで、制約付きネットワークを構築する。類似度ネットワークの隣接行列を S、カテゴリネットワークの隣接行列を C とすると、制約付きネットワークの隣接行列を R は、式 1 より次式のようになる。

$$R = (1 - r)S + rC \tag{3}$$

今回は、モデルの簡単化のために r を行列ではなく単なる定数とする。このように構築した制約付きネットワーク上でクラスタリングを行う。クラスタリング手法としては、Newman 法を用いる [3]。

Newman 法は、階層的クラスタリング手法の一つであるが、クラスタリングを式 4 のような評価関数 Q の最大値導出問題に置き換えた手法である [3]。

$$Q = \frac{1}{2m}\left[\left(\sum_{v,w} A_{vw}\delta(c_v, c_w)\right) - \left(\sum_{x,w} \frac{k_v k_w}{2m}\delta(c_v, c_w)\right)\right] \tag{4}$$

ここでは、Q を次式 5 のように書き換えた重み付き Newman 法を提案する。

$$\begin{aligned} \Delta Q_{ij} &= 2(e_{ij} - a_i a_j) \\ e_{ij} &= \frac{\text{クラスタ } i, j \text{ 間のエッジの重みの和}}{\text{全エッジの重みの和}} \\ a_i &= \frac{\text{クラスタ } i \text{ 内の頂点と結び付いたエッジの重みの和}}{\text{全エッジの重みの和}} \end{aligned} \tag{5}$$

これにより、制約付きネットワークのエッジの重みの違いを考慮したクラスタリング結果が得られる。

4. 評価実験

本節では提案手法の評価を行う。

ある論文集合のカテゴリ分類の評価を行う場合、人間によるカテゴリ分類を正解として適合率、再現率を測る評価手法が一般的である。しかし、カテゴリの分裂や発生といった現象を考慮したデータがあるわけではないので、直接それらを評価することは難しい。

そこで本論文では、ある学会や論文誌の論文集合の過去のカテゴリ分類の変化を正解データとし、それらをどの程度再現できるか、ということで提案手法の評価を行う。

4.1 評価実験の概要

本論文では、ある過去の 1 時点のカテゴリ分類をもとに制約付きクラスタリングを行い、それによりその時点からみて未来のカテゴリ分類が再現できるかどうか、ということで評価を行う。

具体例をもって説明しよう。ある学会において、1980 年では、その年の論文が、図 4 のように「言語処理」「人工知能」の 2 つのカテゴリに分類されているとする。また、1990 年では、その年の論文が、「言語処理」が分裂してできた「形態素解析」「機械翻訳」「意味解析」の 3 つのカテゴリと、「人工知能」が分裂してできた「学習」「エージェント」2 つのカテゴリにそれぞれ分裂しているものとする。このようなデータがあるとき、1990 年の 6 つのカテゴリ分類を正解データとする。次に 1990 年の論文を、「言語処理」と「人工知能」に再分類する。そして、この 1980 年のカテゴリ分類によって制約を付加した制約クラスタリングを行う。その結果、図 4 のように 1990 年のカテゴリ分類と一致したクラスタ結果となれば、提案手法は、有効に働いていると言える。

このように過去の 1 時点のカテゴリ分類用いて、それより進んだ過去の時点のカテゴリ分類をどの程度再現できるかどうかで、提案手法の評価を行う。そして、提案手法の効果を確認するために、文書クラスタリングのみの場合 (r=0) と文書分類のみの場合 (r=1) と提案手法を比較する。

4.2 カテゴリの再分類

今回用いる arXiv.org には、元々36 個のカテゴリが存在している。本実験では、それらを 8 つのカテゴリに再分類した。この 8 つのカテゴリは、1993 年から 1996 年の論文で出現回数の多いカテゴリから上位 8 つを取り出したものである。

各論文には主たる分野の他に、関連する分野がいくつか割り当てられる。この時、一つの論文に割り当てられている分野同士を分野共起していると呼び、またこの分野共起している頻度を分野共起頻度と呼ぶ。

本来ある 36 個のカテゴリの 8 つのカテゴリへの再分類は、この分野共起頻度を用いた。つまり各カテゴリは、8 つのカテゴリのうち、最も分野共起頻度の高いものに再分類を行う。

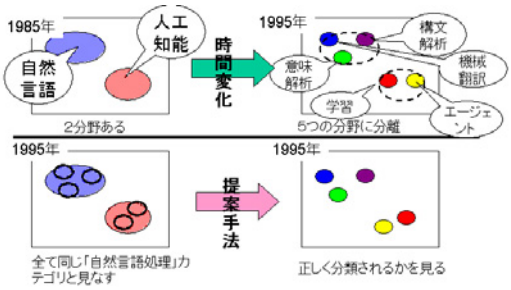

図 4: 制約付きクラスタリングの評価実験手法

4.3 評価実験とその結果

本評価実験では、あらかじめカテゴリ分類された論文データを使用する。今回は、Cornell 大学の arXiv.org[*1] から Computer Science の分野の論文のうち、1993 年から 2005 年までの 13 年分収集し、それを実験データとした。

また、arXiv.org では元々36 個のカテゴリーがある。それを本実験のためにそれらのカテゴリーを表のように 5 つのカテゴリーに再分類した。本実験の手順を以下に示す。

1. 論文集合を発行年ごとに分類し、それぞれ y 年の論文集合を P_y とする。。
2. 論文集合 P_y の各論文の類似度を計算し、類似度行列 S_{P_y} を算出する。
3. 表の 5 つのカテゴリ分類を基に P_y の各論文をカテゴリ分類し、カテゴリ行列 C_{P_y} を算出する。
4. S_{P_y}, C_{P_y} を用いて、式 6 から制約付きネットワーク行列 R_{P_y} を算出する。
5. R_{P_y} をもとに制約付きクラスタリングを行う。
6. 図**??**のカテゴリ分類に基づいて、多数決法によりクラスタ結果と 36 個のカテゴリを対応させる。
7. 適合率・再現率を計算する。
8. 1.〜7. を $P_y(y = 1993 \cdots\cdots 2005)$ について行う。

$$R_{P_y} = (1 - r)S_{P_y} + rC_{P_y} \tag{6}$$

まず、予備実験として r を色々と変化させ、最も値の良い r を選択する。r を 0.001〜1 まで変化させた場合の各 P_y の適合率・再現率をあらわしたグラフを図 5 に示す。図 5 において、縦軸は割合を横軸は r の値を表す。

図 5 において、精度は $r = 0.1$ 前後までは単調増加し、$r \geq 0.1$ ではあまり変化していない。また、再現率は $r = 0.03$ 前後までは単調増加し、その後 $0.03 \leq r \leq 0.3$ では減少し、再び $r \geq 0.3$ において増加している。

数値的には、$r = 0.9$ 付近が精度・再現率共に高くなっている。しかし、$0.3 \leq r \leq 1$ において再現率が高いのは、大きなクラスタができているためである。そこで、再現率が高く、また精度も $r = 0.3$ 付近とほとんど変わらない $r = 0.03$ を今回で最も良い値と考えられる。

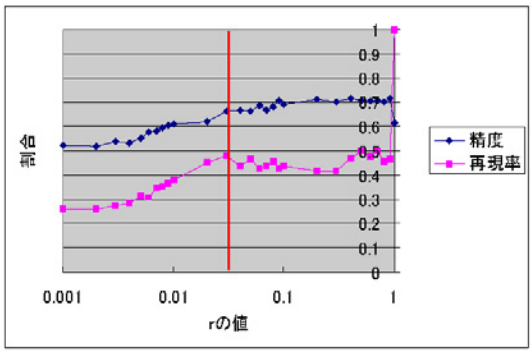

図 5: 論文カテゴリ分類の予備実験結果

表 1: 従来手法との比較

	文書クラスタリング	提案手法	文書分類
精度	0.520	0.662	0.616
再現率	0.263	0.480	1.00

次に r=0.03 と、r=0 つまり文書クラスタリングのみの場合と、r=1 つまり文書分類のみの場合とをそれぞれ比較する。比較結果を表 1 に示す。

まず、提案手法 (r=0) と文書クラスタリング (r=0.03) を比較すると、精度、再現率共に提案手法の方が高い。また、提案手法と文書分類 (r=1.0) を比較すると、精度は提案手法が上回っているもののの、再現率は文書分類の手法が高い。これは、文書分類のデータ用いて論文の分類を用いているため、当然である。しかし、実際には再現率が高い文書クラスタリングにおいては各クラスタのサイズが大きいために、本手法で求める結果としてはあまり良い結果ではない。

これより、提案手法を用いることで文書クラスタリング、文書分類よりも高い精度で論文を分類することができる。以上より、提案手法の有効性を確かめることができた。

5. 議論

本章では、制約付きクラスタリングにより、時系列的なカテゴリの変化を考慮した論文のカテゴリ分類を行った。評価実験にでは、単なる文書クラスタリングの手法と文書分類の手法と提案手法を比較することで、提案手法が確かに論文の分類において有効であることを示した。

今回用いている過去のデータというのは 1993 年〜1996 年で最も多く出現した 8 つのカテゴリである。しかし、本来ならば複数の時点での過去のデータを用いるべきである。過去のデータをきちんと連続的に用いることで、より精度・再現率が向上すると考えられる。

また、カテゴリの実験の結果はカテゴリの再分類にも影響されるため、より正確なカテゴリの再分類を行う必要がある。これについては、論文検索サイトにおいて論文にインデックス付けするためのより詳細なカテゴリの分類が、いくつものサイトに存在しているため、これらを適用することで、さらに適切にデータを取り扱うことを考えたい。

本実験では、予備実験により最もよい値を返す r を決定している。しかし、制約係数 r、つまり過去のデータをどの程度分類結果に反映させるかは、本来自動で決定されるべきである。今後は、そのように r を自動決定する手法についても検討していきたい。

6. 関連研究

本研究に類似した手法として教師ありクラスタリングがあげられる。

教師ありクラスタリングとは、神嶌らが提案している手法であり、事例データの集合から目標とする分割に望ましい規準を獲得し、それをもとに未知のデータ集合のクラスタリングを行う手法である [7]。この手法はクラス分類とクラスタリング手法を合わせているという点で、提案手法と非常に似通っている。

教室ありクラスタリングが事例データからの学習という観点からクラス分類をクラスタリングに組み合わせている。一方、提案手法では時系列的な影響力という観点からクラス分類をクラスタリングと組み合わせている。つまり、教師有りクラスタリングとは観点が異なっていると考えられる。

7. まとめと future work

本論文では、文書分類を文書クラスタリングの手法を組み合わせた制約付きクラスタリングによって、カテゴリが時間変化を考慮に入れた論文のカテゴリ分類手法を提案した。実験結果では、文書分類および文書クラスタリングよりもよいカテゴリ分類の結果が得られており、本手法の有効性が示せていると言える。このように分類すべきカテゴリ自体が変化するという条件において、既存手法よりも高い精度分類結果が得られるのが大きな特徴である。

実際に人間が文書のカテゴリ分類を行う際に、判断基準に用いられるのは、論文の類似度だけではなく、引用文献・非引用文献の一致や、著者が共通している点、などが用いられる。そこで、引用文献・被引用文献情報や著者情報などのデータに本手法を適用することが考えられる。

また、本論文では論文のみを対象としているが、その他の時系列を持った様々なデータに応用可能である。今後は、他の時系列的なデータにも本手法を適用していきたい。

参考文献

[1] *A comparison of document clustering techniques*, 2000.

[2] S. Lawrence. Digital libraries and autonomous citation indexing. Vol. 32, pp. 66–71. IEEE Computer, 1999.

[3] M.E.J. Newman. Fast algorithm for detecting community structure in networks. In *Phys. Rev. E 69,2004*, 2004.

[4] ACM Portal. Acm. http://portal.acm.org/.

[5] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Survey*, Vol. 34, No. 1, pp. 1–47, 2002.

[6] P. Willett. Recent trends in hierarchic document clustering: a critical review. *Information. Processing Manage.*, Vol. 24, No. 5, pp. 577–597, 1988.

[7] 神嶌敏弘, 元吉文男. クラスタ例からの学習 : 分類対象集合全体の属性の利用. 情報処理学会, Vol. 40, No. 9, 1999.

∗1 http://arxiv.org