4L-2

時系列変化を考慮した論文ネットワークの解析
Analysis of network of scientific network using time series modeling

榊　剛史 †　松尾　豊 ††　石塚　満 †
東京大学大学院情報理工学系研究科 †　産業技術総合研究所 情報技術研究部門††

1. はじめに

ある分野や学会全体のがどのような傾向を持って変化してきたか、といった時系列変化を俯瞰的に捉えることは難しい。そのようなサーベイを行うことは研究上必要な作業であるが、非常にコストがかかってしまうのが現状である。しかし、近年、電子化された論文の増加に伴い、それを計算機によって自動的に行う、もしくはサポートするような研究が行われている[3),4)]。本研究では、ある論文集合の持つ分野・カテゴリと行ったものがどのように変化してきたかを自動的に抽出する手法について提案を行う。

各カテゴリというのは時間を追って変化していくものである。1つのカテゴリは時間の経過共に成長したり、または縮小したりする。さらに大きくなったカテゴリが2つのカテゴリに分裂することもある。また新たなカテゴリが発生することもある。本論文では、このような論文カテゴリの時間的な変化を捉えることを目的とする。

そのような論文のカテゴリ捉えるための既存手法としては、文書分類、文書クラスタリングが考えられる。文書分類は、学習器などを用いて既存のカテゴリに文書を分類していく手法である[1)]。しかし、このような文書分類ではカテゴリの成長は捉えることはできても分裂や発生といった、カテゴリ集合自体の変化を捉えることはできない。

一方、文書クラスタリングは、文書間の関係を用いて文書をクラスタリングする手法である[2)]。この手法では、年代ごとにクラスタが変化していくため、カテゴリとクラスタを対応させれば、クラスタの変化を捉えることで、カテゴリ集合の変化を捉えることできる。しかし、クラスタリングの結果というのは図のように年代ごとに大きく変わってしまう可能性があるため、時間軸に沿って一貫性のある結果を得ることが難しくなってしまう。

このように従来手法では、既存のカテゴリを考慮しつつ、それらが成長し、分裂していく時間的な変化を捉えることができない。そこで、本論文ではそのようなカテゴリの変化を捉えることができるような手法を提案する。まず2章で手法の概要を説明し、3章で実装方法について説明する。4章では実験を行い、5章ではその結果について考察し、今後の手法の発展方法について述べる。

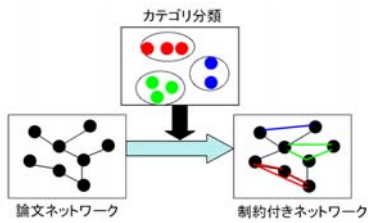

図 1　制約付きクラスタリング

2. 提案手法の概要

提案手法は、文書クラスタリングを基本とし、そのクラスタリングに文書分類に用いるようなカテゴリによる制約を加える。制約という形でカテゴリを考慮したクラスタリングを行うことで、既存のカテゴリの成長や分裂、さらに新しいカテゴリの発生などを捉えることを目指す。

本手法では、図1ある過去のある1時点のカテゴリ状態を考慮し、その上で新しく書かれた論文をクラスタリングする。例えば、1990年に3つの論文カテゴリがあり、1991年に8本の新規の論文 $(P_1, P_2, , , , P_8)$ が加わる場合、既存の3カテゴリによる制約をつけて、8本の論文のクラスタリングを行う。

まず、論文のネットワークを構築する。論文ネットワークとは、論文をノードとし、類似度や引用関係等でエッジをひいたネットワーク構造である。従来手法では、この論文ネットワーク上でクラスタリングを行うことで、文書をクラスタリングする。　次に、この論文ネットワークに既存カテゴリによる制約を加えた上でクラスタリングを行う。ここでは、同じ既存カテゴリに属する、ということを制約とする。論文を既存のカテゴリに分類する。このとき、同じカテゴリに分類されるものは関連があると仮定する。この仮定に基づき、同じカテゴリに分類された論文間のエッジに重みを付加する。前述の例ならば、P_1, P_3, P_5 が同じカテゴリに分類された場合、エッジ $P_1 - P_3, P_3 - P_5, P_1 - P_5$

† Takeshi Sakaki and Mitsuru Ishizuka
Graduate School of Information Science and Technology, University of Tokyo
†† Yutaka Matsuo
National Institute of. Advanced Industrial Science and. Technology

にそれぞれ重みを付加する。

こうして、カテゴリ分類による付加を加えた論文ネットワークを制約付きネットワークと呼ぶ。この制約付き論文ネットワークのクラスタリング結果を論文分類の結果とする。実際には、既存のカテゴリとクラスタリング結果を対応させることで、カテゴリの成長や分裂を捉える。カテゴリ C の成長、分裂、発生は以下のように定義する。

成長　制約付きクラスタリングの結果、カテゴリ C に対応するクラスタが 1 つできた場合

分裂　制約付きクラスタリングの結果、カテゴリ C に対応するクラスタが 2 つ以上できた場合

発生　クラスタに対応する既存カテゴリがない場合

3. 実装方法

提案手法で行われている制約付きクラスタリングは、論文ネットワークの構築、カテゴリ制約の付加、カテゴリの同定の 3 段階に分けられる。本章では、これらの実装方法について説明する。

3.1 論文ネットワークの構築

本論文では、論文をノード、論文のアブストラクト類似度をエッジの重みとして論文ネットワークを構築した。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した。そして、次式のような文書ベクトルの cosine 値を各論文間の類似度とした。

$$cosine(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|}$$

3.2 カテゴリ制約の付加

構築した論文ネットワークに対し、次式を用いて制約を付加した。

$$ResSim(P_i,P_j) = (1-r)\times Sim(P_i,P_j) + r\times Cat(P_i,P_j)$$

式において、$Sim()$ は文書間類似度、$ResSim$ は 2 文書の制約付き類似度、Cat は同じカテゴリに属するかどうかを表す関数である。同じカテゴリに属する場合 1、そうでない場合 0 となる。また、r はどの程度制約をつけるかをあらわす係数であり、制約係数と呼ぶ。またクラスタリング手法としては、パラメータを与える必要がない Newman 法を用いた。

3.3 カテゴリの同定

カテゴリの同定は、多数決法により行う。本手法では、制約クラスタリング結果から見て共通する論文数がもっとも多い既存カテゴリを対応カテゴリとした。この場合、クラスタリング結果と既存カテゴリが多対 1 対応となる。

4. 評価実験

本論文では、arXiv.org☆の過去の論文データとカテゴリ分類を用いて提案手法の評価を行った。本評価実験では、カテゴリの分裂を再現することで、提案手法の有効性を示す。

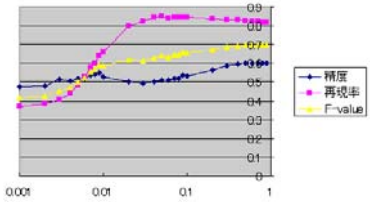

図 2　実験結果

評価では、もともと異なるいくつかのカテゴリを 1 つのカテゴリとみなす。今回は、本来 36 個あるカテゴリを 8 つのカテゴリであるとみなした。そして、この 8 つのカテゴリに分類した論文集合に対して、提案手法を実行し、どの程度本来の 36 個のカテゴリを再現できるかによって評価を行った。評価結果を図 2 に示す。

図 2 のグラフは、横軸は制約係数 $r(0.001 \leq r \leq 1)$ を、縦軸は精度・再現率・F 値の値を表している。グラフには示されていないが r=0 の結果は精度 0.46、再現率 0.36、F 値 0.40 を示しており、他の r の結果よりも低くなっている。これより制約をつけることでカテゴリ分類がよく再現されていることがわかり、提案手法の有効性が示された。　また $r \geq 0.02$ では制約が強すぎるために、全論文が既存の 8 カテゴリに分類されてしまっているので、あまり意味がない。そのため $r = 0.009$ のときに、もっともよくカテゴリが再現されていることがわかる。

5. おわりに

本論文では、論文集合のクラスタリングについて新たな手法を提案し、その有効性を示した。しかし、類似度の算出の精緻化やクラスタリング手法の最適化などにより改良の余地が考えられる。また、制約係数の r をパラメータとして与えなければならない点や、r がネットワーク全体に対して一様である点などは、今後解決していかなければならないだろう。

参考文献

1) C.Jordan. A survey of som based approaches to document classification. Dalhousie Computer Science Technical Reports, 2003, 2003.
2) J.Kogan, C.Nicholas, and M.Tboulle. Clustering large and high dimensional data. CIKM and Workshop2005, 2005.
3) S.Lawrence. Digital libraries and autonomous citation indexing. Vol. 32, pp. 66–71. IEEE Computer, 1999.
4) 難波 英嗣.奥村学. 論文の参照情報を考慮したサーベイ論文作成支援システムの開発. 自然言語処理, Vol.6, No.5, pp. 43–62, 1999.

☆ http://arxiv.org