

ベイズ的アプローチによるクラスタ及び 共引用の特徴量を用いた学術動向予測

大槻 明[†] 川村 雅義^{†‡}日本大学[†] MK Future Software[‡]

1. はじめに

計量書誌学を用いた学術論文ネットワーク分析によって、これまでに学術動向の進展を分析したり、トレンドな学術研究領域を特定したり、さらには、技術と科学との成長関係を明らかにするような研究が盛んに行われてきた[1-9]。しかし、共引用論文の特徴量を用いて既存の研究領域同士の結合を予測するようなモデルは少ない。ゆえに、本研究では、他クラスタとの結合度及び共引用論文の特徴量をベイズ的アプローチによるモデルに適用することによって、既存の研究領域同士の結合により新たな研究領域が生まれる可能性について予測するモデルを提案する。

2. 提案コンセプト

本研究では、直接引用分析を用いて学術論文ネットワークを構築し、このネットワークに対し、Newman 法[10-11]を用いてクラスタリングを行うことで研究領域コミュニティ（クラスタ）を形成する。そして、各クラスタについて、自他クラスタ結合度及び共引用の特徴量を算出し、これらの値を、ベイズ的アプローチによるモデルに適用することにより、既存の研究領域同士の結合によって新たな研究領域が生まれる可能性を予測する手法を提案する。

2.1 新たなクラスタが生まれる期待値の条件付き確率モデル

ナイーブベイズ[12-13]は確率的分類器であり、ナイーブベイズを参考にして新たなクラスタ C が発生する期待値 $E(\text{Expected Value})$ の条件付き確率モデルとして表す。これは、次式(1)のとおり表される。

$$E(C|DC_1, DC_2, \dots, DC_n) \quad (1)$$

$DC_1, \dots, DC_n$ は、新たに発生するクラスタ C の依存クラスタ群を表す。つまり、式(1)は、 $DC_1, \dots, DC_n$ の当該クラスタに依存する形でクラスタ C が新たに生成される確率を表すモデルである。

3.2 自クラスタ内結合度の算出

自クラスタ内結合度については、全次数に占める自クラスタ内リンク数の割合を求める式(2)を定義した。これを I_i とする。

$$I_i = \frac{l_{ii}}{d_i} \quad (2)$$

d_i は i クラスタの全リンク数を表し、 l_{ii} は i クラスタが所属するクラスタ内へのリンク数を表す。

3.3 他クラスタとの結合度の算出

他クラスタとの結合度は、式(3)によって求める。これを O_{is} とする。

$$O_{is} = \frac{l_{is}}{d_i} \quad (3)$$

d_i は i クラスタの全リンク数を表し、 l_{is} は i クラスタから S クラスタへのリンク数を表す。本計算は、対象クラスタに対して外部クラスタ毎に行うため O_{is} という表記となる。

3.4 共引用の特徴量を算出

クラスタごとに共引用の特徴量を算出する式(4)を次のように定義する。

$$COC_{is} = \sum_{j=1}^{N_i} \text{count}_n(\text{cited}_j(P_1) \cap \text{cited}_j(P_2), \dots, \text{cited}_j(P_n))$$

ここで $\{P_1, P_2, \dots, P_n\} \in S$, $n \geq 2$ (4)

- N_i : クラスタ i の論文数
- $\text{cited}_j(P)$: j 論文から引用されている論文 P
- $\cap$: 共引用を示す
- $\text{count}_n()$: $()$ 内要素の数を示す。 n 個ある場合は n となる。
- $\{P_1, P_2, \dots, P_n\} \in S$: クラスタ S に属している論文群

count_n は、 i クラスタにおいて、 S クラスタに属する共引用論文の引用数をカウントしている。つまり、 COC_{is} は、 i クラスタの論文から共引用している他の S クラスタに所属する論文引用数（リンク数）をカウントする。式(4)のイメージを図1に示す。

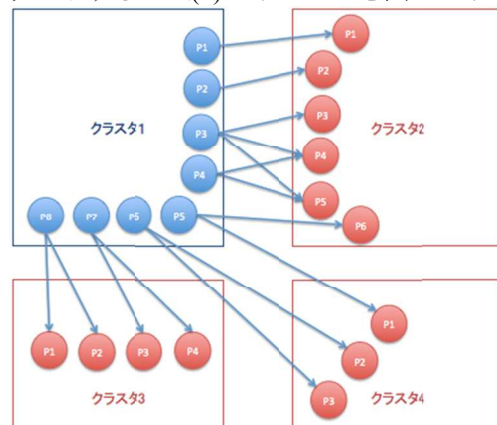


図1 共引用の特徴量をカウントするイメージ

図1では、クラスタ1を基準としてクラスタ1の各論文から共引用関係になっている対クラスタ内だけの論文数をカウントする（重複を除く）。具体的には、リンク数がクラスタ生成に影響するのでエッジ数をカウントする。この結果、クラスタ1から各クラスタへの COC_{is} は次の通りとなる。

- クラスタ2への $COC_{12} = 5$

- クラスタ3への $COC_{13} = 4$
- クラスタ4への $COC_{14} = 2$

これらの値を式(1)に重み付けとして適用する。

3.5 新しいクラスタ（研究領域）が新たに生まれる期待値の計算

まず、 O_{is} （他クラスタとの結合度）を式(1)に重み付けとして適用する。例えば、1 番目のクラスタを基準に考える場合は、 I_i （自クラスタ内結合度）と O_s を、 $\{I_i, O_2, O_3, \dots, O_{ns}\}$ という形で合計が 1 になるように正規化する。 O_{is} の i は、1 番目のクラスタの場合は 1 となるため省略される。そして、このうち $\{O_2, O_3, \dots, O_{ns}\}$ のみを式(1)に適用する。これは式(5)のように表される。

$$E(C|DC_1^{O_1}, DC_2^{O_2}, \dots, DC_n^{O_n}) \quad (5)$$

2 番目のクラスタを基準に考えた場合は、 $\{O_1, O_3, \dots, O_{ns}\}$ となるため、結果的には式(5)の添え字のようになる。さらに、式(5)の依存クラスタ群に式(4)の共引用論文が存在する場合は、その特徴量も重み付けとして式(5)に適用する。図 1 に示した通り、共引用関係が多いクラスタ群は、学術研究成果という観点ではより密接な関係性を持っていると考えられるため、将来の融合を示唆する特徴の一つであると考えられる。ゆえに、式(5)に COC_{is} も重みづけとして適用することによって、他クラスタとの結合度の特徴量に加えて、共引用の特徴量も考慮する形で新しいクラスタ C が発生する期待値を計算することができるように考えられる。具体的な適用例を述べると、例えば、1 番目のクラスタを基準に期待値を考える場合には、 $COCs$ の箇所を $\{COC_2, COC_3, \dots, COC_{ns}\}$ として合計で 1 になるように正規化したうえで該当する依存クラスタ $DC_1, DC_2, \dots, DC_n$ に重み付けする。この重み付けを式(5)に適用したものを式(6)に示す。

$$E(C|DC_1^{O_1COC_1}, DC_2^{O_2COC_2}, \dots, DC_n^{O_nCOC_n}) \quad (6)$$

最後に、式(6)をベイズの定理にあてはめて表す。この式は(7)のように表される。

$$\begin{aligned} & E(C|DC_1^{O_1COC_1}, DC_2^{O_2COC_2}, \dots, DC_n^{O_nCOC_n}) \\ &= \frac{E(C) E(DC_1^{O_1COC_1}, DC_2^{O_2COC_2}, \dots, DC_n^{O_nCOC_n} | C)}{E(DC_1^{O_1COC_1}, DC_2^{O_2COC_2}, \dots, DC_n^{O_nCOC_n})} \end{aligned} \quad (7)$$

式(7)の「 $E(C|DC_1^{O_1COC_1}, DC_2^{O_2COC_2}, \dots, DC_n^{O_nCOC_n})$ 」は Posterior（事後分布）を、「 $E(C)$ 」は Prior（事前分布）を、「 $E(DC_1^{O_1COC_1}, DC_2^{O_2COC_2}, \dots, DC_n^{O_nCOC_n} | C)$ 」は Likelihood（尤度）を、「 $E(DC_1^{O_1COC_1}, DC_2^{O_2COC_2}, \dots, DC_n^{O_nCOC_n})$ 」は Evidence（証拠）をそれぞれ表す。

3. 新たな研究領域を予測するイメージ

前章のコンセプトにて新たな研究領域の予測をした結果（イメージ）を図 2 に示す。図 2 にはクラスタ A～E が存在し、各クラスタの下には、内（In）外部（OutA～E）結合度が表示されている。そして、図 2 では、クラスタ A, B, D が結合して「New

Cluster1」が新たに生まれるイメージを表している。

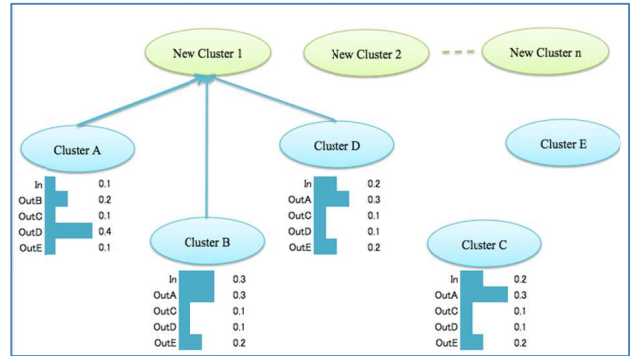


図 2 新たなクラスタが生まれる可能性を予測するイメージ

4. おわりに

本研究では、他クラスタとの結合度及び共引用論文の特徴量をベイズ的アプローチによるモデルに適用することによって、既存の研究領域同士の結合によって新たな研究領域が生まれる可能性を予測するモデルを提案した。しかし、新しい研究領域が生まれるプロセスは、必ずしも既存の研究領域の結合のみに留まらないため、今後は、本研究で明らかにした以外のプロセス（新たな研究領域が生まれる過程）についても研究していきたいと考える。

5. 参考文献

- [1] H. Small: "Co-citation in the scientific literature: a new measure of the relationship between two documents". Journal of the American Society for Information Science, 24:265-269.
- [2] H. Small: Update on science mapping: Creating large document spaces, Scientometrics, Volume 38, Issue 2, pp 275-293, 1977.
- [3] H. small: Tracking and predicting growth areas in science, Scientometrics, Vol.68, No.3, pp.595-610, 2006.
- [4] Igami, Masatsura: Exploration of the evolution of nanotechnology via mapping of patent applications, SCIENTOMETRICS, Vol.77, No.2, pp.289-308, 2008.
- [5] Persson, Olle: Identifying research themes with weighted direct citation links, JOURNAL OF INFORMETRICS, Vol.4, No.3, pp.415-422, 2010.
- [6] Narin, F., Hamilton, K.S., Olivastro, D.: Linkage between agency supported research and patented industrial technology, Research Evaluation 5 (3), 183-187, 1995.
- [7] Meyer, M: Does science push technology? Patents citing scientific literature, RESEARCH POLICY, Vol.29, No.3, pp.409-434, 2000.
- [8] Shibata, Naoki; Kajikawa, Yuya; Takeda, Yoshiyuki; et al. JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, Vol.60, No.3, pp.571-580, 2009.
- [9] http://shibataism.com/pdfs/090119_Dr_thesis_shibata_v1-2_clean.pdf
- [10] M.E.J.Newman and M.Girvan.: Finding and evaluating community structure in networks. Phys. Rev. E 69, 026113, 2004.
- [11] M.E.J.Newman.: Fast algorithm for detecting community structure in networks. Phys. Rev. E 69, 066133, 2004.
- [12] S.L. Ting, W.H. Ip, Albert H.C. Tsang: Is Naive Bayes a Good Classifier for Document Classification?, International Journal of Software Engineering and Its Applications, Vol.5, No.3, 2011.
- [13] Ioan Pop: An approach of the Naive Bayes classifier for the document classification, General Mathematics Vol. 14, No. 4, 135-138, 2006.

Academic Trend Prediction by Bayesian approach Using a Feature Amount of Cluster and Co-citation

† Akira Otsuki

†† Masayoshi Kawamura