1C4-02

論文データベースからの研究トピック抽出

Topic Extraction from Scientific Paper Database

榊　剛史[*1]　松尾　豊[*2]　市瀬　龍太郎[*3*4]　武田　英明[*3*4]　石塚　満[*1]
Takeshi Sakaki　Yutaka Matsuo　Ryutaro Ichise　Hideaki Takeda　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

[*3]国立情報学研究所　National Institute of Informatics
[*4]総合研究大学院大学　The Graduate University for Advanced Studies

In a current flood of computerized academic informatin , the need for search engine for scientific papers has been enhancing considerably. However, now there has been few systems, which can deal with topics and contents of papers. Now, we aim to construct a new search engine for scientific paper, which can get a overview of the past and current states of one field of research. In this paper, we have proposed new methods to cluster papers by topic and evaluated them in the preparatory stages. Our experiments showed that "structually equivalent" is more effective than other methods in this topic-clustering task.

1.　はじめに

我々研究者は研究の過程において、ある分野の論文を集め、読まねばならない場面にしばしば遭遇する。だが一つのトピックについて関連する論文は非常に多く、また以前は様々な場所にに所蔵されていたため、目的とする論文を検索し、手に入れる作業には膨大な時間がかかってしまうことが多かった。

しかし近年、学術情報の爆発的な増加と共に論文が電子化されたため、論文自体を手に入れることは容易になった [6, 2]。そこで、目的とする論文を容易に検索できるように、それらをトピックごとに整理・組織化しデータベース化する＝電子図書館を構築する研究の必要性が近年益々高まってきており、そのような研究が多く為されている。

例えば、クラスタリングやカテゴライゼーションの研究分野では、トピックが類似している論文を組織化するための様々な手法が提案されている [5, 7]。また、一方では論文間の参照・引用関係を手がかりに論文を組織化するような研究も行われている。[10]。

しかし、これらのような２種類のアプローチは必ずしも論文間の類似度を適切に表現できるわけではない。例えば、従来手法では高精度で内容の一致性かはかれるわけではないし、また引用関係があるからといって必ずしもトピックが類似しているとは限らない。

そこで、我々はこれら内容的な類似度と構造的な類似度両方を用いることで、より確実性の高い論文の組織化手法を提案したい。さらに、その論文の組織化を時系列に沿って行うことで、図１のようなある研究分野でのトピックの変遷やアイディアの相関性などを俯瞰的に捉えられるような、論文の検索システムを構築したい。本論文では、そのような検索システム構築の予備段階として、従来の２種類のアプローチ＝内容的類似性と構造的類似性に多少改良を加えた手法を提案する。そしてそれらにより、トピックごとに論文をクラスタリングする際の精度をどのように向上させることができるかを検討する。

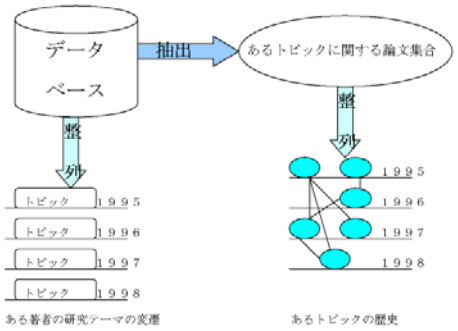

図 1: 提案論文検索システム

2.　関連研究

これまで、論文をトピックに基づいて組織化する手法はいくつか提案されてきたが、それらは、第一節で述べたように、

1. 引用関係などを手がかりとした手法
2. 内容を手がかりとする手法

大きく２つに分類することが出来る。本節では、それらについて紹介しつつ、それぞれの問題点について述べていきたい。

引用関係を基にした手法

代表的な論文の組織化手法として、引用関係を用いた手法があげられる。最もシンプルな引用関係の用い方としては、S.Hitchcock らが引用関係にある論文同士はトピックが類似している可能性が高いものとして、論文の検索に用いることを試みている [3]。また、「同じ文献を引用している論文」や「同じ文献から引用されている論文」はトピックが類似している、という仮説を用いた共引用分析 [7] や書誌結合 [5] などという手

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

法がある。これらは、共通して引用（被引用）している文献の数を、２つの論文の類似度の尺度として用いている。しかし、実際には引用関係がある２論文のトピックが必ずしも一致しているとは限らない。例えば、全く別な分野で使われている手法を引用している場合などは、手法が共通しているだけで全体のトピックはほとんど一致していない。このような問題に対する一つの解決方法として、難波らは参照の理由による参照のタイプ分けと共引用分析の手法の解決を試みているが、クラスタリングやその精度にまでは言及していない [8]。

内容を手がかりとした手法

もう一つのアプローチとして、キーワードなど論文の内容によりフォーカスした研究もある。例えば単語ベクトルを用いたベクトル空間法などがあり [9]、また Cutting らの Semantic なクラスタリングを行っている研究などもある [1]。しかし実際には、内容を手がかりとした手法の精度は高いとは言えず、内容のみを手がかりとした場合にはクラスタリングの精度に限界があると考えられる。

以上のように、２つのアプローチはいずれも確度が高いとは言えない。そこで、今回我々は、内容とネットワークの両方の手法をどのように組み合わせたら精度が上がるか、を検討するために、それぞれの手法がどれだけ論文のトピック分類に効果が現れるかを検証した。

３章ではそれぞれの分類方法について説明し、４章でその実験結果を示す。さらに５章では、それらについて議論し、今後どのように検索システムに活かしていくかについて述べていきたい。

3. 類似度の尺度の提案

前節で説明したとおり、本節では著者ネットワークにおける、ネットワーク分析と内容分析の２つのアプローチによる類似度の尺度を提案をする。

そこで、まず本節では、どのようなネットワークを対象にクラスタリングを行うかについて触れ、さらにそのネットワーク上での類似度の尺度を提案したい。

3.1 対象とするネットワーク

従来研究においては、論文をノードとしてネットワークを構築し、それを対象にクラスタリングを行っている。しかし、本研究では単純なクラスタリングではなく、時系列に沿った解析を行うことを目的としているので、論文のネットワークをそのまま解析対象とすることは妥当とは言えない。なぜなら、年によって大きくクラスタリングが変化してしまう可能性があるからである。

例えば、図２のように 2001 年と 2004 年においてそれぞれクラスタリングを行うと、参照関係の増加や、類似論文の増加から、図２のようにまったく異なるクラスタリングが出来てしまう可能性は少なくない。このように年毎にクラスタリングが異なってしまうと、全体として一貫した組織化が困難になり、時間的なトピックの変遷などをたどることが出来なくなってしまう。そこで、クラスタリングにおいて時間的な変化に左右されない軸となるものが必要である。

図 2: 年代ごとのクラスタリングの変化

そこで今回は著者のネットワークを解析対象にクラスタリングすることとする。これは、ある著者が扱うトピックが時系列によって急激に変化しにくいため、著者を軸として用いることで、全体として一貫した組織化が可能になると考えられるためである。

つまり、本研究では著者ネットワークを構築し、それを解析するものとする。

3.2 ネットワーク構造を基にした類似度

ネットワークを基にした類似度を定める前に、まず何らかの手がかりを用いて著者のネットワークを構築する必要がある。今回は、トピックが共通している確実性の高い**共著関係**を用いてネットワークを構築することとする。つまり、二人の著者が共著したことがある＝共著関係にある場合は、必ず一度以上同じトピックを扱っているので、トピックを共有している可能性は高いといえるのである。実際には、共著したことがある著者同士にエッジを張り、その重みとして共著している論文の数を用いることで、著者ネットワークを構築する。

このようなネットワークにおいて、図３のような状況を考える。図３では、著者 A,B が共著関係にあり、共著 B,C が共著関係にある。このような場合に「著者１人＝１トピック」という仮定においては、著者 A,C の研究トピックが類似していると考えられる。しかし、実際には「著者１人＝複数トピック」であることも多々あり、先ほどの例では著者 A,C のトピックが類似しているとは言えない。

そこで**構造同値**という概念を用いることを考える。構造同値とは、図３のようなネットワーク上において A,B のような関係を指す。つまり、A,B を入れ替えてもネットワークの構造が変化しない、言い換えればネットワーク的に同じ役割を果たしている２つのノードの関係が構造同値である。　先ほどの共著ネットワーク上で、図３のような構造同値の意味を考える。すると、著者 A,B は複数の共通した著者＝著者 C,D と関係を持っているため、「１著者＝複数トピック」だとしても、著者 A,B は同じトピックを扱っていると考えられる。

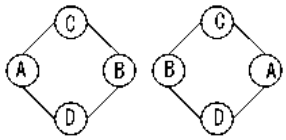

図 3: 構造同値

つまり、構造同値にある２人の著者は、類似したトピックを研究している可能性が高いと考えられる。このように共著ネットワーク上における構造同値を用いて、著者間のトピックの類似度を測る尺度を提案する。２人の著者が図３のような関係にあるとき、構造の類似度を以下の式１で表すこととする。

$$S(A,B) = \frac{v_A \cdot v_B}{\text{全ノード数} - 2} \quad (1)$$

v_A：著者 A のネットワークベクトル

例えば、図４のようなネットワークがあるとき、A,B,C の結合関係を表すベクトルは以下のようになるので、

- $A = (0,0,1,1,1,1)$

- $B = (0, 0, 1, 1, 0, 0)$
- $C = (0, 0, 0, 1, 1, 1)$

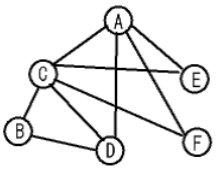

図 4: 構造の類似度

構造の類似度はそれぞれ

$S(A,B) = \frac{2}{4}$ $S(B,C) = \frac{2}{4}$ $S(A,C) = \frac{3}{4}$

となる。

3.3 内容的な類似度

ここでは、それぞれの論文の内容を用いて著者のトピックの類似度を測るための尺度を提案する。

個々の論文を基にした類似度

2 人の著者を比較する最もシンプルな方法は、それぞれが書いた論文を全て比較し、類似度を合計する方法である。ここでは、2 論文の類似度の尺度として tfidf を用いたベクトル空間法を用いるものとする。二人の著者 A,B の類似度 S(A,B) は式 2 で表される（v_i は論文 i の特徴ベクトル）

$$S(A,B) = \sum_{i \in A} \sum_{i \in B} v_i \cdot v_j \tag{2}$$

著者に特徴的な語を基にした類似度

2 人の著者を比較するもう一つの方法として、著者に特徴的な語を用いて著者ベクトルを定義し、それらの内積によって類似度を計算する方法がある。これは、著者に特徴的な語を用いて各著者のトピックを表し、それらの類似性によって著者間の類似度を定義するものである。類似度は式 3 で表される。

$$S(A,B) = v_A \cdot v_B \tag{3}$$

実際には、ある著者の全論文に出現する語によって著者ベクトルを表現した。単語の重みは tfidf を用いた。df は全文書における語の出現頻度、tf はその著者の全論文における語の出現頻度の合計である。

4. 評価実験

3 章で提案した類似度の尺度を用いて、JSAI4 年分の書誌情報とアブストラクトから著者ネットワーク上の類似度を計算し、クラスタリングを行った。さらに、その結果を JSAI のセッション分類と比較することで提案手法の評価を行った。

4.1 クラスタリング手法

今回は、最短距離法による階層的クラスタリングを行った。これは、もっとも距離の近い二つのクラスタを逐次的に併合する、という動作を、すべての対象がひとつのクラスタに併合されるまで繰り返す手法である。ただし今回は逆に、 最初に全てのノードがひとつのクラスタに含まれているとみなし、その状態から類似度の小さい順にクラスタを分割していくことで、クラスタリングを行った。実際には、エッジで結ばれているノードを同じクラスタに含まれているものとみなした。

今回のクラスタリング手順は以下のとおり。

1. 最終的な目標クラスタ数を定める。
2. 類似度の小さい順にエッジを切る。
3. エッジを切るごとに、クラスタの数をカウントする。
4. クラスタ数が目標クラスタ数を超えたら、そこで終了。

ただし、ネットワーク上に孤立したノードが多く、ほとんどの場合、初期状態で目標クラスタ数を越えてしまっていた。そこで、今回は初期状態から独立しているサイズ 1 のクラスタはクラスタリングの対象からはずすこととした。

4.2 評価手法

本研究では実験に JSAI4 年分のデータを用いたので、JSAI のセッション分類を正解としてクラスタリングを評価することとした。

実際には図 5 のように、作られたクラスタに対し、対応セッションを判別し、その中に含まれる正解、不正解の数で精度・再現率を算出した。クラスタにどのセッションが対応するか、

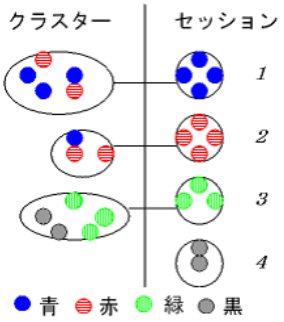

図 5: 評価方法

の判別はクラスタに含まれる著者のセッションの多数決で決定した。手順をまとめると以下のとおり。

1. 各クラスタに含まれる著者のセッションを調べ、そのクラスタが対応するセッションを決定する。
2. 対応セッションとクラスタ内の各著者のセッションを比べ、一致不一致を判定し、グループ A,B,C の数をカウントする。

   グループ A あるクラスタの中で、分類が不一致の著者

   グループ B あるクラスタの中で、分類が一致している著者

   グループ C あるセッションの中で、異なるクラスタに分類された著者

   たとえば、図 5 の 1 ならば、本来の青いノードがクラスタ 1 に分類されるはずのものであるので、

   - A_1=2（クラスタ 1 の中の赤）
   - B_1=3（クラスタ 1 の中の青）
   - C_1=1（クラスタ 1 以外に含まれる青）

   となる。

表 1: 評価実験結果

	共著		構造		論文		特徴	
	Prec	Rec	Prec	Rec	Prec	Rec	Prec	Rec
01	41.6	10.1	37.6	17.0	43.0	16.8	41.1	15.5
02	55.2	6.87	37.1	10.2	32.0	9.94	34.9	10.5
03	62.1	5.20	40.9	7.67	37.2	7.23	35.0	6.87
04	33.8	10.7	7.20	11.3	79.6	19.4	72.1	17.0

3. $A = \sum A_n, B = \sum B_n, C = \sum C_n$ とし、式 4 のように精度・再現率を計算する。

$$Precision = \frac{B}{A+B} \qquad Recall = \frac{B}{B+C} \qquad (4)$$

このような手法を用いると、たとえば図 5 の 4 のように対応するクラスタリングがないセッションがあれば、A=0,B=0,C=2 となるので、再現率のみ下がる。逆に対応するセッションのないクラスタがあれば B=,C=0 となり、精度のみが小さくなる。このように、本評価方法では適切にクラスタの精度・再現率を評価することができる。

4.3 評価実験

前述までに提案してきた 4 つの手法（共著、構造同値、著者論文、著者特徴語）を用いて、JSAI4 年分（2001〜2004 年）の論文によるネットワークを年毎に構築し、クラスタリングを行い、その評価を行った。ただし、4 年分では共著ネットワークの情報が不十分であるため、共著情報として JSAI11 年分（1990〜2000 年）のデータを用いた（CiNII[4]）。

また、今回はクラスタ数を共通にするために、4 つの手法においてクラスタ数は各年毎のセッション数に一致するようにした。（2001 年から順にクラスタ数は 46 個、62 個、74 個、57 個）ただし、共著と構造同値の 2002、2003 年においては、初期状態でセッション数以上のクラスタに分かれてしまっているため、クラスタ数は他の 2 つの手法より大きくなっている。

評価結果は表 1 のとおり。これを見ると、まずいずれの年でも著者論文や著者特徴語を用いた手法の方が値が大きくなっている。2002、2003 年の共著関係の値が大きくなっているが、これは初期状態ですでに多数のクラスタに分けられていたことに起因している。このことから、構造同値や共著といったネットワークを解析する手法は、内容を解析する手法に比べ精度が劣ると考えられる。

しかし、図 6 のグラフを見ていただきたい。この図は 2001 年の各手法でのクラスタ数と精度の関係を表したグラフである。（図 6 内の 46 は表 1 の際のクラスタ数）この図 6 においては、圧倒的に「構造同値」の精度が圧倒的に大きくなっている。ただし、クラスタ数が少ない場合は「構造同値」の値が存在していない。これはクラスタ数を初期状態より小さくできないためである。ここでは、図 6 より「構造同値」を用いれば高精度でクラスタリングが出来ることが分かる。

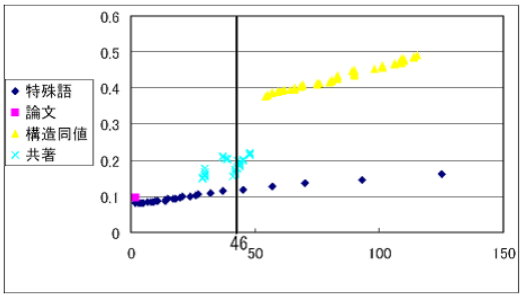

図 6: 各手法におけるクラスタ数ー精度の相関性

結局、図 6 と表 1 より、精度では 4 つの手法のうち「構造同値」を用いるものが有意に優れていること、しかし「構造同値」のネットワーク密度は「著者論文」「著者特徴語」に比べて低いために、クラスタ数を少なく出来ないこと、が分かる。

5. 結論と考察

本論文では、より便利な論文検索システム構築のための予備段階として、論文をトピックごとにクラスタリングする手法を提案し、評価を行った。実際には、複数の従来手法を拡張し、それらを用いて著者ネットワーク上でのクラスタリングを行い、その分類の妥当性によって複数の手法を比較・評価した。

その結果、精度では「構造同値」を用いる手法が優れていることが分かった。ただし、「構造同値」を用いた場合、ネットワークがスパースになるため、少ない数へおクラスタリングには不向きである。そのような場合には、「著者論文」や「著者特徴語」を用いれば良いことが分かった。

つまり、トピックごとのクラスタリングの精度を高めるためには、まず「構造同値」など引用・共著関係など、トピックが一致している確度の高い情報をメインに用い、それらを内容的な手法を用いて補強すればよい、と結論づけることが出来る。今後はこの結果を踏まえ、新たな論文検索システムを実装していくことを考える。

参考文献

[1] Pederson. CuttingD., KargerR. and Turkey. Scatter/gather: a cluster based approach to browsing large document collections. *Proc. of 15th annual International ACM SIGIR Conf. on Rand D in Information Retrieval*, 1997.

[2] ELSEVIER. Science direct.

[3] S.Hitchcock et al. Citation linking: Improving access to online journals. *Proc. of 2nd ACM International Conference on Digital Libraries*, pp. 115–122, 1997.

[4] National Information Institute. Cinii.

[5] M.M Kessler. Bibliographic coupling between scentific papers. *American Documentation*, 1963.

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

[7] H Small. Co-citation in the scientific literacture: A new measure of the relationship between two documents. *American Society for InfomationScience*, 1973.

[8] 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.

[9] 長尾真（編）. 自然言語処理. 岩波書店, 1996.

[10] 三平善郎. 論文間の引用関係を用いた主題抽出とその検索システム. 日本ソフトウェア科学 全国大会, 1995.