2C1-04

論文情報を利用した研究コミュニティの発見

Discovery of Research Communities from Information on Bibliographies

市瀬 龍太郎[*1] 武田 英明[*1] 植山 浩介[*2]
Ryutaro Ichise Hideaki Takeda Kosuke Ueyama

[*1]国立情報学研究所 [*2]トライアックス
National Institute of Informatics TRIAX Inc.

The research community is very important for researchers in order to undertake new research topics. In the present paper, we propose an discovery method of research communities from information on bibliographies. We also discuss about research communities obtained by our method.

1. はじめに

近年の情報技術の発展に伴い，研究に関する情報がインターネットで多く公開され，最新の研究成果を素早く知ることができるようになってきた．それに伴い，さまざまな分野において，次々と新しい研究が起きるようになっている．そのため，全ての研究者は，研究の新しい動向を把握しておくことはもちろんのこと，新たな研究を継続的に開拓していかなければならない．新たな研究を始める時には，その研究テーマに関連する論文や研究者などについて，調べることが第一歩となる．本研究では，ある研究に関連する論文や研究者の集合を研究コミュニティと定義し，それらを発見するための手法を提案する．

研究に関する情報は，主に論文として他の研究者に伝わる．そのため，研究に関するコミュニティを発見するには，論文情報が最も重要となる．本研究では，論文情報からコミュニティの発見を試みる．論文情報から，コミュニティを発見する試みは，さまざまな方法で取り組まれている．共引用分析 [Small 73] を用いる方法では，共引用されている論文同士である研究テーマに対して関連があるとし，それらがある研究トピックに関してコミュニティを形成しているとみなす．しかし，この方法で形成されるコミュニティは，全体を俯瞰するためには適しているが，得られるコミュニティが巨大になるため，細かい研究テーマを見ることには適さない．一方，CiteSeer [Sci 05] のように，各研究から，周囲の研究を徐々に調べていき，コミュニティを探していくアプローチもある．しかし，このような方法は，全体からの見通しが悪いという問題点がある．そこで，本研究では，ユーザとのインタラクションを通して，徐々にコミュニティを洗練していくことで，ユーザが求める適切な規模のコミュニティを発見できるような手法を提案する．

2. コミュニティの要素

本章では，論文情報から得られるコミュニティの要素を検討する．共引用分析では，引用されている論文の間にある研究テーマに関して関連があるとみなし，それらがある研究トピックに関してコミュニティを形成しているとみなしている．同様に，書誌結合 [Kessler 63]，引用，共著 [Newman 04] では，特定の関係がある論文，著者の間でコミュニティを形成しているとみなしている．本研究では，論文情報から得られるコミュニティの要素として，共著関係，論文引用関係，著者間引用関係の３つを取り上げる．それぞれの関係は，図 1 のようになっている．

連絡先: 市瀬 龍太郎，国立情報学研究所 知能システム研究系，
〒 101-8430 東京都千代田区一ツ橋 2-1-2，ichise@nii.ac.jp

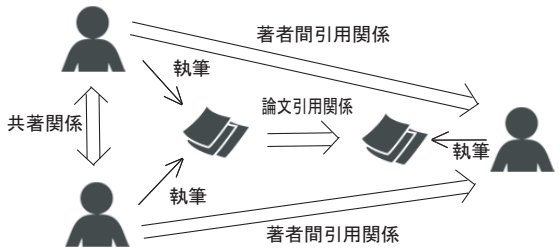

図 1: コミュニティ発見に使用する関係

共著関係では，ある論文を共同で執筆した研究者の間にコミュニティを形成する要素，すなわち，ある特定の研究テーマに関して，共に研究していたり，共に興味を持っていたりすると仮定する．研究者をノード，共著関係をアークとし，共著した論文数をアークの重みとすると，重み付きグラフを生成することができる．この重み付きグラフから，コミュニティを発見する．同様に，論文引用関係では，論文を引用している場合に，引用元と引用先の論文の間に，コミュニティを形成する要素があると仮定する．この場合には，論文をノード，引用関係を引用元から引用先に引かれるアークとし，重みが 1 となる重み付き有向グラフが得られる．このグラフからコミュニティを発見する．最後の著者間引用関係は，共著関係と論文引用関係を拡張した考え方である．本論文では，引用関係にある論文と共著者の関係にある著者の両方について，コミュニティの要素，すなわち，共通の研究テーマや共通の興味があると仮定している．この仮定を拡張すると，引用関係にある論文を書いた著者同士は，共通のテーマや共通の興味を持っていると仮定することができる．その仮定に基づくと，研究者をノード，論文を執筆した著者とその論文が引用した論文の著者の関係をアーク，その回数を重みとした重み付き有向グラフを生成することができる．このグラフからコミュニティの発見をする．

3. コミュニティ発見の指標

前章で述べた関係を使うと，グラフを構成することが可能となる．このグラフは，それだけでコミュニティを示していると言うこともできるが，一般的にこのようなグラフは巨大になることが知られており [Barabási 02]，この状態で，コミュニティの意味を理解することは困難である．そこで，適切な規模のコミュニティに洗練をしてやる必要性がある．

本研究では，ユーザが望む適切なコミュニティに洗練するために，２つのステップを用いる．１つ目のステップでは，計算機がある指標と閾値にしたがって，自動的にコミュニティを生成する．そして，２つ目のステップで，計算機の出すコミュニティをユーザが理解し，計算機に新たな指標や閾値を指示する．この２つのステップをインタラクティブに繰り返すことにより，目的となるコミュニティの発見を行う．そのためには，コミュニティとして与えられたグラフを何らかの指標で洗練する必要性がある．そのような指標として，いくつか考えられている [Freeman 79]．本研究では，コミュニティ発見のための指標として，単純重み，最大流量，Closeness の３つを用いて，グラフの洗練を行う．

単純重みとは，グラフのアークを切断することで，コミュニティの分割をおこなう方法である．ある一定の閾値を設定し，その閾値に満たない部分を消去することで，関係の弱い部分を切断する．その結果，関係が密なコミュニティのみが残っていくため，コミュニティ発見の重要な指標の一つになると考えられる．

最大流量では，アークの重みをパイプの太さとみなして，あるノードから水を流した時に，別のノードでどれだけの水が得られるかを計算して関係性の指標とする．単純重みと比べると，この手法では，直接つながっていないノード間に対しても，間が太い関係でつながれている場合には，関係性が大きくなるという特徴を持つ．この指標では，あるノードとそれ以外の全てのノードに対して，どれだけ水が流れるか計算を行い，その総和が少ないノード，すなわち，他の全てのノードと関係が弱いノードを削除することで，コミュニティの洗練をする．

最後の Closeness は，異なるノードの近さを考慮する指標である．異なるノード間をつなぐ経路が複数ある場合に，お互いをつなぐ最良の経路を通ると別の経路よりも近い場合がある．それを見るために，あるノードと他のノードの間の距離を指標とする．この時，他の全てのノードとの距離の総和を計算すると，距離が小さいもの程，コミュニティの核になっている可能性が高いと考えられる．逆に，距離が大きいもの程，コミュニティの外縁にいると考えられる．そのようなノードを消去することで，コミュニティの洗練を行う．

4. コミュニティの分析

前章で述べた手法に基づいて，コミュニティを発見するシステムを試作した．このシステムは，CiNii [Nat 04] で使われている論文の情報を加工し，約１２万件の論文情報，９万件の著者情報から，コミュニティの発見を行うことができる．詳しいデータの加工方法やシステムの動作などは，[市瀬 05] を参照されたい[*1]．

紙面の都合上，ここでは共著関係を単純重みで取り扱った場合のみを取り上げることとする．全ての論文の著者情報に対して，さまざまな閾値の単純重みを使って，コミュニティを生成させてみた．その結果が図 2 である．図の横軸はコミュニティの大きさ，縦軸はコミュニティの数を示す．この図は，対数グラフになっている点に注意をしてもらいたい．この設定では，例えば，閾値５の場合には，共著を５回以上した場合のみ，著者を示すノード間にアークが引かれ，それ以下のアークは，削除されたコミュニティが得られる．このグラフから，どの閾値においても，コミュニティの大きさが小さいもの程数が多く，コミュニティの大きさが大きいものほど数が少なくなることが

*1 一部の手法の組合せに関しては，未実装であるが，http://irweb.ex.nii.ac.jp/ でシステムを試用できる．

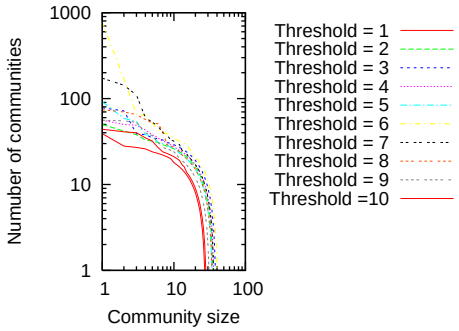

図 2: 得られるコミュニティの大きさと数の関係

読み取れる．また，大きさが小さいものの数が非常に多いのに対して，大きさが大きいものが少数個しか存在しないことが分かる．しかし，閾値を上げることによって，より多くの小さいコミュニティが得られていくことが分かる．

5. おわりに

本論文では，コミュニティ発見のための３つの関係性と３つの指標を提案し，それによって，コミュニティを洗練化し，必要なコミュニティを発見するための方法を示した．本研究では，指標を使って，ユーザがインタラクティブに操作をすることでコミュニティの発見を行う方針を用いたが，今後は指標や閾値の自動選択を行う機構について研究を進めて行きたい．

参考文献

[Barabási 02] Barabási, A.-L.: *LINKED: The New Science of Networks*, Perseus Books Group (2002), 邦訳：青木 薫 [訳], 「新ネットワーク思考」

[Freeman 79] Freeman, L. C.: Centrality in social networks: Conceptual clarification, *Social Networks*, Vol. 1, pp. 215–239 (1979)

[Kessler 63] Kessler, M. M.: Bibliographic Coupling between Scientific Papers, *American Documentation*, Vol. 14, No. 1, pp. 10–25 (1963)

[Nat 04] CiNii (Citation Information by NII), National Institute of Informatics (2004), http://ci.nii.ac.jp/

[Newman 04] Newman, M. E. J.: Coauthorship networks and patterns of scientific collaboration, *Proceedings of the National Academy of Sciences of the USA*, Vol. 101, No. suppl. 1, pp. 5200–5205 (2004)

[Sci 05] Scientific Literature Digital Library (2005), http://citeseer.ist.psu.edu/

[Small 73] Small, H.: Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents, *Journal of the American Society of Information Science*, Vol. 24, pp. 265–269 (1973)

[市瀬 05] 市瀬 龍太郎, 武田 英明, 植山 浩介：コミュニティマイニングのための研究者情報の視覚化, 信学技報, Vol. 104, No. 587, pp. 1–6 (2005)