3B8-5

著者話題モデルを用いた研究話題の関係の発見

Relationship Discovery of Research Topics using Author-topic Model

市瀬 龍太郎[*1] 藤田 摂[*2] 村木 太一[*2] 武田 英明[*1]
Ryutaro Ichise Setsu Fujita Taichi Muraki Hideaki Takeda

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

In research and development in evolving fields in science and technology, it is essential to keep abreast with current trends and to predict future trends. In this study, we propose a relationship discovery system of research topics for finding them. In order to evaluate the proposed method, we conducted experiments. The experimental results indicate that this system can induce the appropriate relationships for finding research trends.

1. はじめに

社会の発展に伴い，最先端の研究の論文数は年々増加傾向にある．論文のデータベースを提供している Thomson 社によると，世界中で 5 年間に発表された論文の数は，1981 年から 1985 年まで約 230 万であったのに対して，20 年後の 2001 年から 2005 年には 400 万近くと増加の一途をたどっている．このような膨大な研究の中には，常に新しい研究分野が発生している．そのような新しい研究分野は，既存の分野との関係を保ちつつも，さまざまな分野と関係することにより，新たな研究分野として確立していくことが多い．例えば，最近，注目を集めているネットワーク科学の分野は，物理学，社会学，情報工学など，さまざまな研究分野と関係しつつ，どの分野にも含まれない学際的な話題として研究分野が確立しつつある．したがって，このような研究の話題間の関係性を発見することにより，さまざまな研究分野の中で新たに発生しつつある研究話題を見付けることができるようになると考えられる．本論文では，著者話題モデルと自己組織化マップを組み合わせて用いることにより，研究話題の関係を視覚的に発見する手法について述べる．

2. 提案手法

本研究では，文献のデータを利用して関係の発見を試みる．研究論文には，タイトル，著者，抄録の 3 点が含まれるのが一般的である．これらの情報から，まず，研究者，話題，論文の関係を抽出する．そして，それらの関係を視覚化することにより，研究の動向を掴むことを試みる．

本研究では，関係の抽出に，著者話題モデル [Steyvers 04] を用いた．著者話題モデルは，文書の生成にある確率モデルを想定し，そのモデルに沿って確率を推定することにより，著者と話題の関係，話題とキーワードの関係の 2 つを確率的な表現で得る手法である．ここで確率的な表現とは，お互いにどの程度の関係があるのかを確率で示したものである．例えば，ある著者が話題 A に関係している確率が 0.3，話題 B に関係している確率が 0.1 であるというような形式で，著者と話題の関係が得られる．この時，話題は予め与えられたものではなく，モデルに基づき自動的に話題が得られる．同様に，キーワードに対しても話題とキーワードの関係が確率的な表現で得られる．この確率の推定には，ギブスサンプリングを用いる．本研究では，キーワードとして論文の抄録の中から名詞を抽出して利用した．その結果，キーワードの頻度と論文の関係が分かるため，キーワードと話題の確率的な関係を用いると論文と話題の確率的な関係についても推定することが可能となる．本研究では，この著者と話題の関係，話題と論文の関係を用いた．

上で述べたように，著者と話題の関係，話題と論文の関係は，確率のベクトルによって表される．しかし，関係の近さ，遠さなどはベクトル形式のものを見ただけでは分かりづらい．そこで，関係性を発見するために，これらの確率のベクトルを図示することを試みる．本研究では，この図示のために自己組織化マップ [Kohonen 01] を利用した．自己組織化マップは，格子状のニューラルネットワークを使って，高次元のデータを低次元のデータに落す手法であり，一般的には 2 次元で視覚化するのに使われる．2 次元の格子の上で，近いベクトルデータ同士が集められるため，視覚的にデータ間の関係を発見することができる．この手法を著者と話題の関係を示した確率のベクトルに適用することにより，同じ研究話題に関わる研究者のコミュニティやそれぞれのコミュニティ同士の関係，話題同士の関係を把握することが可能となる．さらに，話題と論文の関係を利用して，論文も同じ 2 次元空間上にマッピングすることができるため，著者，話題，論文の 3 者の関係を視覚的に発見することが可能となる．

3. 論文マッピングシステム

前章で提案した手法に基づいて，論文マッピングシステムと呼ばれるシステムの作成を行った．図 1 は，論文マッピングシステムの画面である．このシステムは，対話的に操作することが可能であり，このシステムを通して研究者が研究している話題や研究論文の位置付けが分かるようになっている．

図 1 の 1 の部分は，本システムの核となる部分で，計算された関係を 2 次元で視覚化して表示する．ここでは，著者話題モデルによって導出された話題を別々の色で表示する．表示の際には，著者表示モードと論文表示モードの 2 つがあり，初期状態では著者モードとなる．著者モードでは，著者名を画面上に表示する．その際，全ての著者を表示すると画面が読めなくなるため，画面の拡大度に応じて著者名が表示されるようになっている．前章で述べたように，著者の話題に対する確率ベクトルを使ってこの画面が構成されるため，研究話題が近い著者同士が近傍に配置されることになる．著者モードの時に，画面上の著者を指定すると論文モードに切り替わり，その著者

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，Fax:03-3556-1916，E-mail:ichise@nii.ac.jp

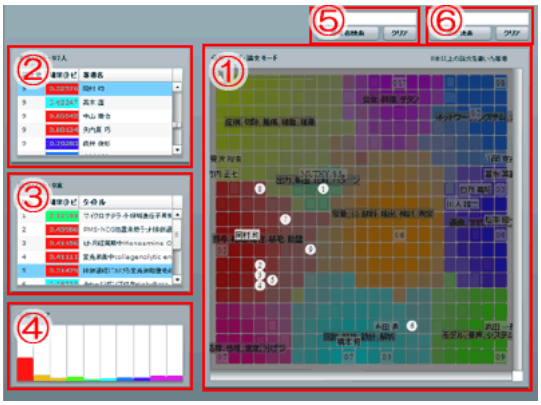

図 1: 論文マッピングシステムの画面

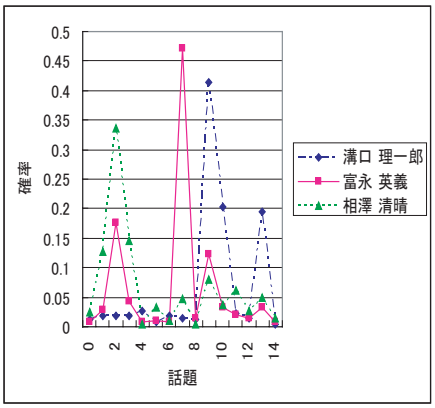

図 2: 著者に対する話題と確率の関係

が書いた論文がマップ上で表示される．論文も著者と同様に，話題に応じて配置される．

図 1 に表示されている 2 から 4 の部分は，情報の詳しい表示に使われる．2 では，研究者のリストが表示される．各々の研究者は，著者話題モデルにより計算された最も関連の深い話題の色と同じ色が付けられてリストで表示される．同様に，3 では，論文に対して最も関連の深い話題の色が付けられ，リストで表示される．しかし，このような方法だけでは，各々の研究者や論文に対して，その他の話題に対する関連度を知ることができない．そこで，4 の部分に棒グラフを用意し，指定された研究者や論文に対して，話題との関係がどのようになっているのかを棒グラフで表示するようになっている．5 と 6 は，検索のためのものであり，それぞれ，著者，論文が検索できるようになっている．

4. 実験

本システムにおいて，どのような関係が得られるかを調べるために実験を行った．実験には，CiNii の文献データベースを用いた．まず，データベースに５０本以上の論文データがある研究者をランダムで１０００人抽出し，各々の著者に対して１０本の論文をランダムに抽出した．その結果得られた論文のデータを用いて実験を行った．関係を見るために，著名な研究者である溝口理一郎氏，富永英義氏，相澤清晴氏の三人について分析を行った．著者と話題の関係を表したのが図 2 である．図の横軸が話題を示し，縦軸が確率を示している．溝口氏は，話題 9 に大きな関係が見られる．話題 9 には，システム，開発，情報，利用などが確率の高いキーワードとして与えられており，情報系の研究分野に割り当てられていることが分かる．富永氏は，話題 7 に対して大きな関連が見られる．話題 7 は，ネットワーク，方式，提案，通信などが確率の高いキーワードとして与えられており，通信ネットワーク系の研究分野が割り当てられていることが分かる．相澤氏は，話題 2 に大きな関係が見られる．話題 2 には，画像，手法，次元，領域，抽出などが確率の高いキーワードとして与えられており，画像処理系の研究分野が割り当てられていることが分かる．一方，3 者の関連が低いとされる話題１４では，特性，発光，製作，基板，素子などが確率の高いキーワードとして与えられており，電気電子工学系の研究分野が割り当てられていることが分かる．富永氏，相澤氏は，それぞれ，「高知能映像情報ネットワーク」，「映像・メディア信号処理」を自己の研究課題として取り上げており，類似した分野を研究していると言える．図 2 を見ると，話題 2，7，9，１３が他の話題よりも共に高くなっており，研究分野の類似性が正しく抽出できていることが分かる．一方，富永氏は，ネットワークという語を研究課題の中に挙げており，相澤氏と比較した時の大きな差異の一つとなっている．このことは，話題 7 の大きさの差として見られる．この結果より，提案手法を使うとそれぞれの研究者や研究話題の関係性を導出できることが分かる．したがって，この関係を表示することにより，研究話題の関係を発見することができると言える．なお，その他の実験を含めた詳しい実験結果については，[市瀬 07] で報告されている．

5. おわりに

本論文では，研究話題の関係を発見するための新たな方法として，著者話題モデルを用いて研究者，研究話題，論文の間の関係を導出し，それを視覚化して対話的に研究話題の関係を発見する手法を提案した．そして，その手法に基づくシステムを構築し，実験を行った．実験の結果，提案した手法では研究者と話題の関係，話題とキーワードの関係が適切に導出できることが示された．したがって，本システムを使うと研究者が集中している分野などが分かるため，新たな動向の分析などに応用できると考えられる．

今後は，さまざまな論文のデータセットに対して，このシステムを適用し，どのような時に研究の動向が変わるのかを同定する手法を研究していく予定である．また，現状のシステムは，確率の計算に大きな計算コストがかかるため，その計算量を減らすことも大きな課題であると考えられる．

参考文献

[Steyvers 04] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T.: Probabilistic Author-Topic Models for Information Discovery, In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 306-315, (2004).

[Kohonen 01] Kohonen, T.: Self-Organizing Maps, Springer (2001), (邦訳：自己組織化マップ, 徳高平蔵ほか監修，シュプリンガー・フェアラーク東京 (2005)).

[市瀬 07] 市瀬龍太郎，藤田 摂，村木太一，武田英明：著者話題モデルによる話題予測の評価，信学技報, Vol. 106, No. 473, pp. 13-18, (2007).