1G1-2

投稿論文に基づいた研究動向の変遷の解析と検索支援

Analysis on Change of Research Trend and Support a Search for Papers

奥岡晋大[*1] Shinta OKUOKA　片上大輔[*1] Daisuke KATAGAMI　新田克己[*1] Katsumi NITTA

[*1] 東京工業大学 大学院 総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering, Tokyo Institute of Technology

Analyzing research trend of research societies are useful for many researchers. To recognize the research trend, we focus on similarities among research domains of the research society. We extract such similarities from the proceedings of conferences held by the research society and express them in the form of a research network. By comparing research networks of different year, we show transition of research trends is visualized.

1. はじめに

近年，文書間や研究者間の関係をネットワークによって視覚化し，ネットワークの構造から新たな知見を得る研究が多く行われるようになった．例えば，論文の引用関係に基づいて論文をノードとしたネットワークを生成し，ある分野の転換期となった重要な論文を発見するといった分析手法を提案する研究などがある[Chen 04]．

これまで研究者や文献そのものをノードとしたネットワークを構築することは多かったが，研究分野に着目したネットワークを構築する研究は少なく，学会内の分野間の関係性や年代ごとの分野の関係の変化が見えてこなかった．

本研究では，学会に投稿された論文に基づいて分野をノードとして，各分野に属する論文集合をテキスト解析して分野ごとにベクトルを生成し，各ベクトルの類似性が高いものをリンクで結ぶ分野間ネットワークを生成することにより，学会の特性と研究分野の変遷を解析する．また，これまであらかじめ用意された論文集合のみでネットワーク化してきたが，ユーザが提示した論文集合と用意してある論文集合との関係性を可視化することを可能にした．さらに，可視化の観点についても，今まで変遷情報を見ることはとても難しかったが，時間軸に対する分野間関係の変化の見せ方に工夫を行った．これらを実現するためのツールを紹介し，それを利用して学会の研究動向の変遷の解析方法を提案する．

以降第2章では関連研究について述べ，第3章では分野間ネットワークの生成方法，第4章では開発した視覚化ツールの機能について説明する．第5章で年代ごとの分野間ネットワークから現在の研究領域における分野の変遷に関する知見やユーザの提示論文と学会との関連性の視覚化について述べる．第6章でまとめとする．

2. 関連研究

最近ではネットワーク分析に関する研究が盛んに行われている．安田ら[安田 06]は Web マイニングの手法により収集・構築された研究者ネットワークを年代ごとに生成して比較することにより，研究者のコミュニティの様子を分析する研究を行っている．この分析手法は我々の構築した研究分野ネットワークに関しても有用な分析手法であると考える．研究分野のネットワーク構造も静的なものではなく，時間とともに構造が変化する性質を考慮するべきである．

また，論文を用いた研究では，Chen は論文の引用関係に基づいて論文をノードとしたネットワークを生成し，ある分野の転換期となった重要な論文を発見するといった分析手法を提案している．しかしこの研究では一つの研究分野の転換期となった年がわかるが，複数の分野の関係を知ることはできない[Chen 04]．

他にも，Mane らは論文をテキスト解析してバーストして出現する単語を検出し，バーストした単語をノードとして共起性が高いものをリンクで結んだネットワークによる視覚情報を提供している．論文の内容に基づいている点，バーストした時期によるノードの色分けの点は参考になるが，バーストした単語のみを扱うことからつねに高頻度に出現する単語がネットワークに出現しないため，学会における重要な情報を取り逃がしている可能性がある[Mane 04]．

論文や研究者ではなく，分野に注目した研究として Anegon らが引用データベースの分野カテゴリを用いて分野間の関係をエゴセントリック（あるノードを中心とした）ネットワークとして視覚化している．分野間という点で本研究と類似しているが，エゴセントリックネットワークは視覚化としては優れているがその犠牲として情報損失もあるため，ネットワーク分析といった解析対象としては不向きである．つまり，視覚情報以上の知見を得ることは難しい[Anegon 05]．先行研究[片上 07]は年代の異なる分野が混在した分野間ネットワークによって年代の比較を行っているため，年代ごとのネットワークの構造変化による比較ではない．

よって，本研究はある学会の分野関係を分野間ネットワークとして視覚化するツールの提案と，それを用いて年代ごとのネットワーク構造の変化の様子，そこから研究分野の特性について論ずる．

3. 分野間ネットワークの生成手法

分野間ネットワークの生成手順を説明する．まず，予めカテゴリ情報と用語辞書を用意して，学会に投稿された論文集合を入力して，論文を内容に基づいて解析し，分野間の関係性を生成する．そして，この関係性を基にネットワークとして視覚化を行う．

3.1 分野間ネットワークとは

分野間ネットワークの定義を述べる．ノードは分野名とし，リンクの生成方法は分野ごとに属する論文の用語頻度を計測して重み付けしたベクトル（分野ベクトル）を作成し，そのベクトル同

連絡先：奥岡晋大，東京工業大学，〒226-8502 神奈川県横浜市緑区長津田町 4259 J2-53，TEL&FAX:045-924-5218．okuoka@ntt.dis.titech.ac.jp

士の類似度がある閾値を超えたときに該当分野同士の関係性は高いと判断し，リンクを結ぶ．この情報をもとにネットワークとして視覚化したものを分野間ネットワークと呼ぶ．

3.2 データセット

分野間ネットワークを生成する上で必要な情報は，「カテゴリ情報」と「学会の論文」と「用語辞書」である．それぞれについて詳しく述べる．

本研究では，分野をノードとしたネットワークを生成するが，ここでの「分野」について説明する．分野といっても広義，狭義のものがある．ここでは，ある学会に注目してその学会内における分野の関係性を調べるため，その分野の名前と範囲は学会の論文募集一覧を参考にして分野名セットを用意した．しかし，オーガナイズセッションや近未来セッション，研究会セッションは特殊なものと判断し，セッション名をそのまま用いた．ここで特殊なものを除いてセッション名を用いなかった理由を述べる．今回は年度ごとのネットワークの構造変化を見ることが目的だが，セッション名は年度によって異なる名称がつけられることが多いので，その際にノードが変化すると年度ごとの比較が難しい．それを吸収するために，学会の論文募集一覧を用いた．ここで決めた分野名セットをもとに，論文がどの分野に属しているかをカテゴリ情報としてまとめる．論文の分類は，手作業により大会のセッション名を参考にして各分野に振り分けを行った．このとき，論文は必ず1つの分野にのみ属すとして，2 つ以上の分野にまたがることはないようにした．この手作業による分類が正しいと仮定した上で，分野間ネットワークを作成する．

用意する論文は学会に投稿された論文集合である．今回は2001～2007 年度の人工知能学会の全国大会の論文ファイル（PDF）を用意した．ただし，PDF からテキストを抽出する際にうまく抽出できない場合にはその論文ファイルを除外した．

次に，用語辞書について説明する．各分野をベクトル化する際に，用語の頻度を計測するが，その用語は予め登録した用語のみとした．今回は，人工知能学会に合わせて，「情報処理ハンドブック」，「人工知能ハンドブック」，「AI 事典」の索引部分の用語約1万語を手作業により登録することにより，用語辞書を作成した．

3.3 論文の内容に基づいた分野間の関係性の生成

本研究では，文書の類似度計算によく使われる文書ベクトルを使ったベクトル空間法を用いた．論文ファイル(PDF)からテキストを抽出し，全文を用いて計算を行う．抽出したテキストを形態素解析器「MeCab」を通して形態素に分け，用語辞書に登録されている用語の出現頻度を計測して，TF*IDF 値を計算して文書ベクトルの要素とした．用語辞書を用いることによって，ノイズになる一般語を用いないようにした．分野間の類似度は分野の文書ベクトル同士がなす角の余弦値とした．求めた余弦値が指定した閾値を越えた場合に，その分野間は関係性が高いと判断した．

3.4 ネットワークグラフの視覚化

分野間の関係性を基に視覚化ツールでネットワークグラフとして視覚化する．ネットワークのレイアウトアルゴリズムは文献[Fructerman 91]を用いた．人工知能学会の大会の 2006 年度の分野間ネットワークを図 1に示す．

可視化の際に，ノードやリンクの属性を表すための工夫を行った．まずノードの色と大きさであるが，こちらは分野に属する論文数を表している．（大きい，赤色）は論文数が多く，（小さい，青色）は論文数が少ないことを表す．ただし，緑色はその年度

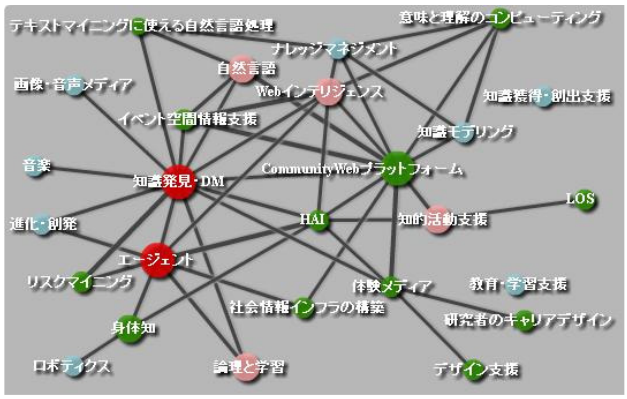


図1：2006年度の分野間ネットワーク

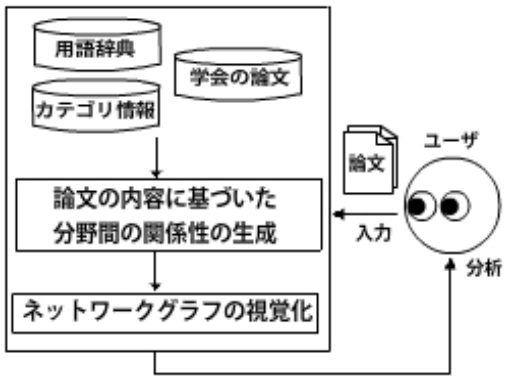


図2：分野関係解析システムの概要

における特殊なセッションを表し，特別扱いとして色を区別した．これにより，その年度における投稿論文数の多い分野が一目で分かる．次にリンクの太さであるが，これは分野間の類似性の高さを表している．太いものほど類似性が高いことを表しているので，自分の関係している分野を見ることで，どの分野と関係が深いかを判断できる．

4. 分野関係解析システム

4.1 分野関係解析システムとは

分野関係解析システムは，分野間の関係性を解析するシステム（図3）であり，分野間ネットワークを様々な視点で見ることを可能にする（図3～図5）．このシステムの概要を図2に示す．

また，論文検索支援として各分野に属する論文のタイトルがわかるようにシステムの右側の「論文タイトル」ボタンをクリックすることで，論文タイトル一覧が表示される（図3）．他，「重要用語」ボタンをクリックすると，その分野の用語で TF*IDF 値が高いベスト30が表示され，研究者へ分野の重要キーワードを提示する．この重要キーワードを眺めることによって，知らないキーワードがあれば調べるなどのきっかけ作りにも役立つ可能性がある．

以降，ユーザに有意義な情報を提供するための機能を紹介する．

4.2 2パスネットワークの生成

分野ノードとエッジの数が多い場合，無条件に全体の分野間ネットワークを表示すると分野把握が困難になることがある（図1）．Narin らは関係の特徴を容易に表示するために単純な2ステップ・マップを開発した[Narin 97]．我々はその手法を取り入れて，注目した分野を中心に表示し，その注目分野から2ステッ

プ目までのネットワークを構築した．本研究では2パスネットワークと呼ぶことにする．

4.3 時間軸に対する分野の変化を視覚化

これまでネットワークによる可視化手法は，ある時点におけるネットワークを表示するもので，他の年代とのネットワーク構造の変化を見ることが難しかった．我々は分野間の関係性の変化を見やすくするために，ユーザが注目している分野を中心に配置したネットワークを作成して，図4のように右下のバーを移動させることによりその年度ごとの分野関係のつながりの変化が見えるようにした（ツールの「変遷ボタン」の機能）．薄く表示されているノードは前年度に関係性があった分野であることを表す．ノードが位置を固定された状態で変化するため，どこがどう変化しているのかをわかりやすくした．

これによって，

- 年代が変化してもずっと固定の分野と関係性を持つ
- 色々な分野とつながりが変化したり，リンク数が増減する
- 年代ごとの論文数の増減

といった現象を容易に見ることが可能になった．

4.4 任意に提示した論文を含めたネットワークの生成

ユーザが任意に提示した論文集合に対して，分野間ネットワークを生成することを可能にした（図3）．これまではネットワークによる視覚化は予め用意された論文集合を基に生成を行うものが主流であったが，自分がどの分野と関係性を持っているかを判断することが難しかった．我々はユーザの提示論文を含めた分野間ネットワークを生成することで，ユーザと学会との関係性を視覚化できるようにした．現在，分野の複雑化が進んでいるため，自分の分野と他の分野との関係性が分かりづらくなってきている．例えば，大会のセッションに応募する際もどのセッションに応募すればよいか決定することにも使うこともできる．このことから，学会におけるユーザが行ってきた研究の位置づけを知ることにこの機能は有効であると考えられる． また，関連の深い分野の論文タイトルを見ることによって，検索支援にもなると考えている．

5. 投稿論文に基づいたネットワーク分析

ここではテキスト情報を用いた分野間ネットワーク分析例として人工知能学会の全国大会に適用する．ここでは 2001 年から 2007 年の7年分の PDF ファイルを用意した．

以降，得られた知見について考察する．

5.1 2001-2007 間の分野間ネットワーク

図5は 2001 年から 2007 年の分野間ネットワークを 2 年おきに並べたものであるが，年々ネットワークの複雑化が進んでいることがわかる．これは，現在の研究状況の複雑さを表している．あらゆる分野の情報を取り込み，自分の研究に活かそうとする動きがあるのであろう．さらに，そういった複雑化している分野とそうでない分野との二極化が進んでいる．例えば，「Web インテリジェンス」は徐々に複雑化が増してきているのに対して，「基礎・理論」はずっと次数が 0 か1である．ただし，これによって「Web インテリジェンス」がこの学会において流行していて，「基礎・理論」は廃れているという結論にはならないことに注意してほしい．しかし，今の人工知能学会では「Web インテリジェンス」の技術は多くの分野と関連を持つようになってきているということはわかる．実際，表1に分野数・論文数・次数をまとめても次数の平均は増加傾向にあり，次数の分散は 2002 年を除いて大きくなっているので，複雑化と二極化が進んでいると考えられる．

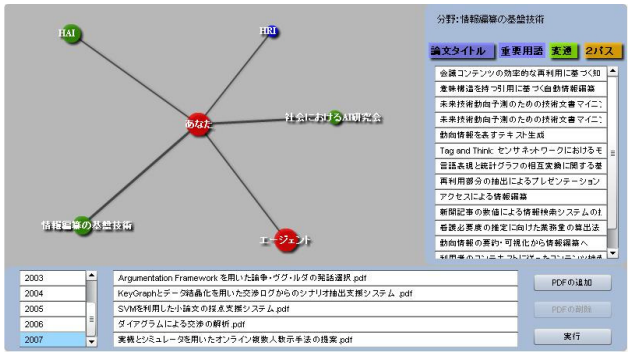


図3：分野関係解析システムのインターフェース

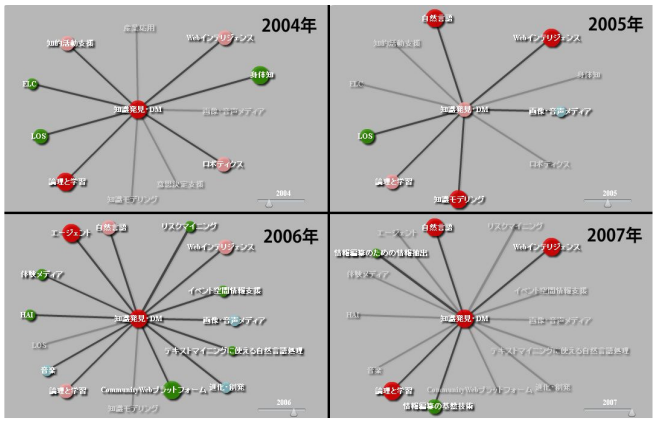


図4：時間軸に対する「知識発見・DM」の関係変化

ただ，こうした複雑度や二極化を示す指標として次数や次数の分散を使うのが必ずしも正しいとはいえないが，視覚的には捉えることができる上，どの分野が複雑化しているのかをすぐ特定することができる．可視化による利点は大きい．

他にも，年度が増すごとに特別なセッション（緑色のノード）が多くの次数を獲得する傾向が見られる．次数を獲得できない特別セッションは減り，多くの次数を獲得できるものは毎年行われる傾向がある．特別セッションはその時代において必要と感じられるものが企画されやすいので，多くの分野の技術を取り込むのではないかと考えられる．逆に，次数の低い特別セッションは実際に調べた場合にほとんどが一年で終わってしまう傾向があるので，多くの分野と関連性を持たない場合はその学会に対する適合度が低く，長く続かないのではないかと仮定する．

5.2 時間軸に対する分野の変化を視覚化

図4は2003年から2007年までの「知識発見・DM」を中心に見たときの関係変化を年代バーによって見えるようにしたものである．これによりわかることは，「Web インテリジェンス」と「論理と学習」とはずっとリンクされているので関係性の深さがわかる．しかし，それ以外では分野のつながりの変化が激しく，特別セッションとのつながりも多いことからどの分野でも利用できる技術は多いが，技術の内容の変化が激しくまだ安定していないのではないかと推測できる．

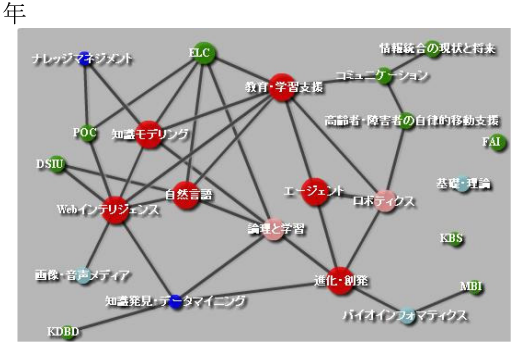
5.3 提示論文と学会との関係性

例えば，我々が所属する東京工業大学新田研究室の 2007 年に研究会で発表された5本の論文(PDF) を提示したときに 2007 年度人工知能学会の分野間との関係性をネットワーク化したものが図3である．提示した論文集合は「あなた」というノードで表現される．関連する分野とリンクで結ばれているが，リンクが

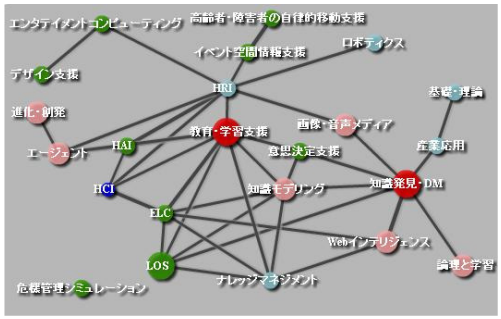
表1. 年度ごとの分野数・リンク数・論文数・次数の比較

	2001	2002	2003	2004	2005	2006	2007
分野数	22	20	22	28	20	24	31
リンク数	34	30	39	45	43	47	70
論文数	169	165	208	247	260	258	314
次数の平均	3.09	3.00	3.39	2.9	3.31	3.48	4.24
次数の分散	4.17	8.00	6.06	4.28	6.83	10.4	11.6

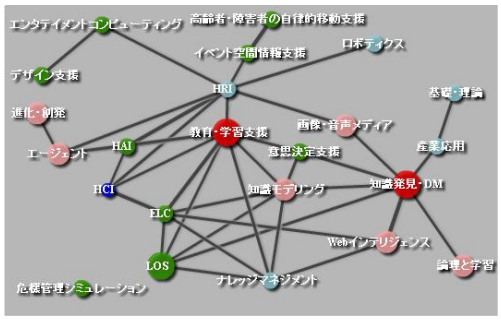
2001 年



2003 年



2005 年



2007 年

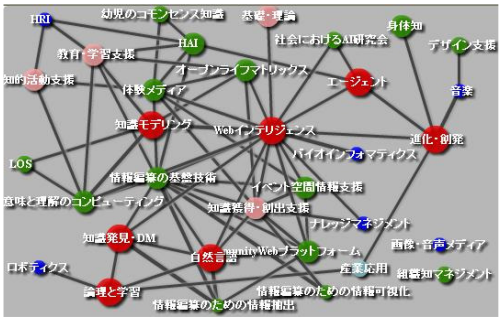


図5：2001～2007年の分野間ネットワークの変遷

太い分野「情報編纂の基盤技術」が今回提示した論文集合のもっとも関係性が高い分野であると判断できる．その分野の論文タイトルを表示することにより，検索支援にもなると考えている．

6. おわりに

本研究では，学会に投稿された論文に基づいて，分野をノードとした分野間ネットワークを生成することにより学会の分野間の関係性や変遷を可視化及び解析する手法を提案した．

その結果，人工知能学会について以下のことがわかった．

- 分野関係の複雑化及び二極化が進んでいる
- 特別なセッションは多くの分野と関連を持たないと長く続かない傾向がある
- 時間軸に対する分野のつながりの深さがわかる．また，分野関係の変化が激しいものもあり，技術の安定度合いがわかる

時間軸に対する変遷の可視化はまだ課題がたくさんあるが，本研究では有効なアプローチの第一歩を紹介できたと思う．さらに，これまでは予め用意されたものによるネットワーク化が主流であったが，ユーザを含めたネットワーク化を考えることも今後面白い課題であると考えている．

参考文献

[Chen 04]C. Chen: Searching for intellectual turning points: Progressive knowledge domain visualization, Proceedings of the National Academy of Sciences, vol.101, 5303-5310, 2004

[片上 07]片上,清水,田中,新田,山田: 文献情報に基づく学際的分野間ネットワーク分析，人工知能学会全国大会論文集, 1B2-07, 2007

[Fructerman 91]T. M. J Fructerman and E. M. Reingold: Graph Drawing by Force-directed Placement , Software-Practice and Experience vol.21, pp.1129-1164, 1991

[Anegon 05]Moya-Aneg´on, et._al: Domain analysis and information retrieval through the construction of heliocentric maps based on ISI-JCR category cocitation, Information Processing and Management, Vol.41, pp.1520-1533, 2005

[安田 06]安田,松尾,武田: 人工知能学会におけるネットワーク構造と変化,人工知能学会全国大会論文集,1F2-1,2006

[Mane 04]K.Mane and K.Borner: Mapping Topics and Topic Bursts in PNAS, Proceedings of the National Academy of Sciences of the United States of America, Vol.101, pp.5287-5290, 2004

[Narin 76]F.Narin and G.Pinski, HH.Gee: Structure of the biomedical literature, Journal of the American Society for Information Science, Vol.27, No.1, pp.25-45, 1976