1B2-07

文献情報に基づく学際的分野間ネットワーク分析

Analysis on Network Structure and Dynamism in Research Fields

片上 大輔[*1] Daisuke KATAGAMI　清水英明[*1] Hideaki SHIMIZU　田中貴紘[*1] Takahiro TANAKA　新田克己[*1] Katsumi NITTA　山田隆志[*1] Takashi YAMADA

[*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

Recently, interdisciplinary research activities relating to various fields have been increasing. Therefore, analyzing research trends in journals or conferences are useful for many researchers. We developed a graph analyzing tool which visualize research networks, and analyzed constructed networks in the annual conferences of JSAI.

1. はじめに

社会がますます高度化，複雑化，多様化するなかで，各分野の研究も細分化された知を有機的に統合するグローバルな，横断型科学技術がその重要性を高めている．一昨年度は文理にまたがる 43 の学会が，自然科学とならぶ技術の基礎である「基幹科学」の発展と振興をめざして横幹連合が設立され，知の横断型会議である「第 1 回横幹コンファレンス」が盛況のうちに開催された．しかし，このような試みは近年始まったばかりであり，特に各分野の知見をどのように活用するかは未だ手探り状態であるといえる．現状では，各分野における情報交換にとどまっているといっても過言ではない．そこには，各分野に共通するような横断的かつ包含的な事象に基づいた解析とその情報提供が不可欠である．一方，これまで各研究分野における文献探索を行なう場合，検索エンジンや文献の孫引きなどを駆使して行なってきた．CiteSeer 等，幾つかの特殊なサイトには詳細な文献情報があり，参照情報や文献間の類似情報により必要な文献を見つけることが可能である．しかし，各研究分野において実際にどのような研究がどのように行われているか，ある会議の今年と昨年の違いは？この 10 年でどのように研究テーマは変遷してきたのか？など，全体像の把握を行なうことは難しかった．また，これまで研究者や文献という観点でネットワークを構築することは多かったが，研究分野という観点でネットワークを構築する研究は少なく，学会やジャーナルごとの分野の分布や偏り，時系列における研究分野の変遷や鳥瞰図的な視点などを知ることはできなかった．

本研究では，会議や論文誌単位などのテーマに沿って集められた文献情報に基づき，文献と研究内容に関する関係を研究分野間の関係性としてネットワーク的に捉えることで，様々な分野にまたがり関連する研究分野間の関係を明示的かつ多角的に整理・考察し，これまで曖昧であった全体像を明確にすることで、研究領域の垣根を越えた学際的な研究分野間の関係を理解することを目的とする．

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なってきており，松尾ら [松尾 05] は Web からの研究者関係を抽出しそのネットワークを構築，中心的な研究者を分析する研究を行っている．また，難波らは，文献間の参照関係から関連論文の組織化を行っている [難波 06]．また，Anegón ら [Anegón 05] は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．分野間の視覚化という点で本研究に類似しているがこの論文ではジャーナルごとのデータを使っており，ジャーナルまたは学会同士の比較をすることができない．本研究ではこれらの問題点を改善し，学会やジャーナルごとの分野分布の比較やより細かい分野間の関係を探るための視覚化手法を提案する．

視覚化手法に関しては，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作り，ある分野 (この論文では超ひも理論について) の転換期となった論文について分析を行っている [Chen 04]．さらに，論文をノードとし引用関係でエッジを張り，参照年代が視覚的にわかるように色分けを行っている．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ができない．

また，分析手法に関して，安田ら [安田 06] は，Web 上の情報から抽出した研究者をノードとした研究者関係のネットワークを視覚化し，社会ネットワーク分析の特に中心性を中心に分析を行っている．何らかの組織やグループがあるとき，中心的な存在，他者に影響を与える存在を，ネットワーク的な視点から捉えようというのがネットワークの中心性の分析である．これらの分析手法は我々の構築した研究分野間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．

3. 研究分野とネットワーク構造

3.1 研究分野ネットワーク

本研究では，ジャーナル・学会ごとの分野分布の比較および分野動向を知るために分野をノードとしたネットワークを作成，視覚化することを目的としている．しかし，多くの研究分野においては，分野が体系化されておらずその論文がどのような分野に関わっているかを知ることが難しい場合が多い．例えば ACM[ACM] でも分野分類はあるが，それが他の学会などで共通して用いられてはいない．そこで，各論文が分類されたカテゴリを分野として捉え，他の分野に属する論文との類似度を計算する．類似していると判断された論文セットがそれぞれ属する分野間にどのような関係があるのかを調べるネットワークが文献情報に基づく研究分野ネットワークである．このネットワークでは一枚のネットワーク上に異なる年代や学会のノードを表現し，それらのリンクを見ることで時系列分析や学会同士の比較を行う．ここで，主な解析の流れを図 1 に示す．まず，入力された文献情報に基づき，それぞれ分野をノードと

連絡先: 片上大輔，東京工業大学，〒 226-8503 神奈川県横浜市緑区長津田町 4259 J2-53，TEL&FAX:045-924-5218，katagami@ntt.dis.titech.ac.jp

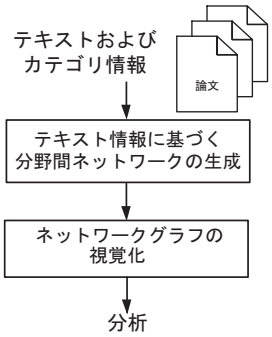


図 1: 解析の流れ

したネットワークを生成する．そしてこれらの生成したネットワークを，視覚化ツールを通して視覚化し分析を行う．

3.2 文献情報に基づく研究分野間ネットワーク構築

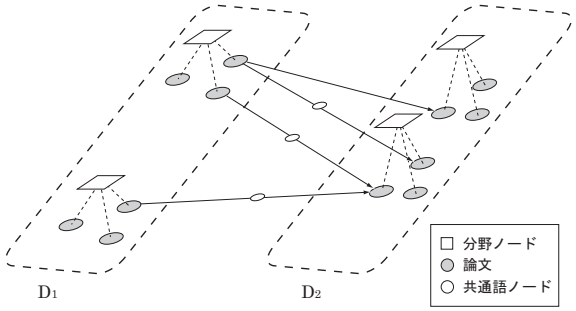
一つの論文であっても複数の技術，分野に関わっていることが多い点に着目し，その論文内に共起している分野情報を入力する．分野情報とは，論文に付加されているその論文が関わっている分野を示した分野コードやキーワードのことである．ジャーナル同士や時系列による比較をするためには，ノードに対応する分野名の統一化が必要であり，分野およびその階層構造を定義する必要もある．分野情報は，ここで定義された分野に一致させる必要があるが，完全に一致させられない場合には，定義された分野への割り当てを行うモジュールが必要となる．

本稿では，論文のテキスト情報とカテゴリ名 (たとえば全国大会のセッション名) を用いて類似度を計算し，類似文書が分類された分野間にどのような関係があるかを知るためのネットワークについて述べる．分野として認識されているカテゴリは，時間と共に流行によって細分化されたり，もしくは廃ってしまったようなカテゴリもある．そこでそのような分野の流行，廃り，枝分かれといった情報を得ることができるネットワーク表現を提案する．

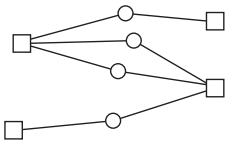
[**入力情報**] 論文のテキスト情報およびその論文が関連する分野のカテゴリ名を入力する．ここでの分野ノードは，カテゴリ名を分野として仮定し，その論文が属するカテゴリをそのまま用いる．たとえば学会などが開催する全国大会のプログラムにおけるセッション名などがそれに対応する．これにより，主にその論文の研究内容を示すカテゴリ名以外の分野との関連性を文書類似度によって発見し，属している分野およびそれ以外の分野との関わり合いを網羅的に見ることができる．

[**ネットワーク表現**] 分野が体系化されていない場合にはノードとなる分野および階層構造の定義を自分らで行う必要がある．この場合は文書データベースなどのカテゴライズ情報や大会プログラムのセッション情報などを元に作る．論文に付加されているキーワードなどを元に定義した分野への割り当てといった作業も生じる．ここで提案しているネットワークは論文を完全に一つの分野に分類することが難しい場合が多い，という点に注目している．

ここである 2 つの文書集合を D_1,D_2 とする．それぞれに属する分野，そしてそれら分野に属する論文についてもう一方の文書集合に属する論文との文書類似度を計算し，ネットワークを生成する．現在提案しているネットワークにはリンクを張る方向性がある．ここでは D_1 から D_2 へのリンクを生成する



(a) 類似度を用いた文書間リンク



(b) 視覚化ツール上でのネットワーク表示

図 2: 類似度を用いた文書間リンクの生成

場合を考える．図 2 は文書間の類似度を用いて二つの文書集合 D_1，D_2 の間で分野間のネットワークを生成する場合の図である．

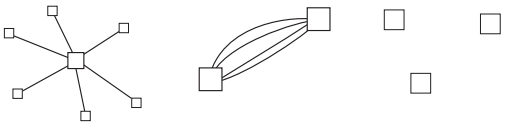
D_1 から D_2 への方向でリンクを作る場合，D_1 中の論文それぞれについて D_2 中の文書すべてとの類似度を計算し，あるしきい値以上の類似度を持つ文書のセット間でのみリンクを生成する．図 2(a) では D_1，D_2 中に計 4 つの分野があり，その中にいくつかの論文が属している．D_1 に属するすべての論文 (6 本) と D_2 に属するすべての論文 (6 本) との類似度を計算する．あるしきい値以上の類似度を持つ文書のセットを作り，それらの文書が属する分野 (これがそのままノードとなる) 同士間でリンクを生成する．このように二つの文書集合間でリンクを生成し，視覚化ツールで表示すると図 2(b) となる．なお類似度に関するしきい値はユーザーが指定できる．

分野ノードの間にはそのリンク間における共通語集合ノードを挿入する．そのリンクがどのような単語で類似しているか，といった情報を示す．リンクが結ぶ両端の分野のうちどちらの分野に関して類似しているかを知ることができ，分野の変遷を追うときに役立つ．

[**本研究における論文類似度計算と共通語の抽出**] 本研究では，論文の類似度計算に最も一般的に使われている文書ベクトルを使ったベクトル空間法を用いた．2 つの文書集合を D_1，D_2 とし，D_1 から D_2 への方向でリンクを生成するとき，D_2 をコーパスとして D_1，D_2 中のすべての論文について文書ベクトルを作り，類似度を計算する．論文ファイル (PDF) からテキストを抽出し，基本的に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プログラム「茶筅*1」を通して形態素に分け，tf*idf 値を計算して文書ベクトルの要素としている．このときストップワードリストを用いてノイズになりそうな一般的に出現する形態素は除いている．文書間の類似度は式 (1) の cosine 相関値を用いた．

$$\text{類似度} = \frac{\vec{v_A} \cdot \vec{v_B}}{|\vec{v_A}||\vec{v_B}|} \quad (1)$$

*1 http://chasen.naist.jp/hiki/ChaSen/



(a) ハブ型　(b) 多リンク型　(c) 孤立ノード型

図 3: テキスト情報に基づく分野間ネットワークにおいて観察できるネットワークパターン

分析者が指定した値以上の類似度を持つ文書間においてリンクが作られ，最終的にそれぞれが属する分野同士でリンクを張る．

共通語はそれぞれの文書ベクトルの要素である tf*idf 値の積を計算し，その値が上から高い順に数語を選び，共通語集合として抽出した．

[**観察できるネットワークパターン**] テキスト情報に基づく分野間ネットワークにおいて観察できるパターンとその解釈について述べる．観察できる観察パターンを図 3 に示す．

(a) **ハブ型**: ある分野ノードについて次数が高く，多くの分野ノードと接続しているパターン．時系列分析では分野の細分化，他分野との関連増加．学会間分析では分野の細かさの違い．

(b) **多リンク型**: ある 2 つの分野ノード間で多くのリンクを持っているパターン．時系列分析では関係が増加した分野関係．学会間分析では関連の強い分野．

(c) **孤立ノード型**: 他の分野ノードへのリンクを一つも持たないパターン．時系列分析では新しく生じた分野，廃れて消えた分野．学会間分析では一方で扱われていない分野．

また，分析方法の一つとして社会ネットワーク分析の次数（Degree）に基づく中心性をみる．特に研究分野ネットワークにおいては，人間関係のネットワークとは異なり，出次数は主に、その研究分野をメインとする研究の流行を，また，入次数はその研究分野が他分野に貢献している度合いとして考えることができる．

ネットワーク生成後，そのネットワークを視覚化ツールでネットワークグラフとして視覚化する．ツールはユーザが見たい年代のみに絞り込むフィルタリング機能，グラフ同士を比較するための差分グラフ生成，中心性（次数）表示などの機能を持っている．視覚化ツールは JAVA で実装されており，グラフ描画には JUNG[*2]を用いた．図 4 に視覚化ツールのユーザインタフェースを示す．図中には 4 つのウィンドウが表示されており，このように Multi Document Interface 形式で複数のグラフを同時に開くことができる．

4. 文献情報に基づくネットワークの分析

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ついて適用する．まず，分野の変遷を調べるために，年代毎を文書集合として年代間の類似度を計算，ネットワークグラフを生成した．ここでは人工知能学会 (JSAI) 全国大会の 2001 年，2003 年，2005 年の 3 年分の PDF ファイルを用意した．分野となるノードはそれぞれの大会プログラムのセッション名を用いた．基本的にはセッション名をそのまま分野としたが，一部，明らかに分野とすることができないものは対象外としてデータベースから除いた．また年度が違ってもセッション名が継続して使われているもの，たとえば強化学習，遺伝的アルゴリズムといったものもある．このように，全く分野名が継続されているものを年度別に区別するために，このネットワークではそれぞれの年の分野ノード (セッション) を色分けによって区別できるようにしている．以降では，得られた特徴的ないくつかの知見について考察する．

*2 http://jung.sourceforge.net/

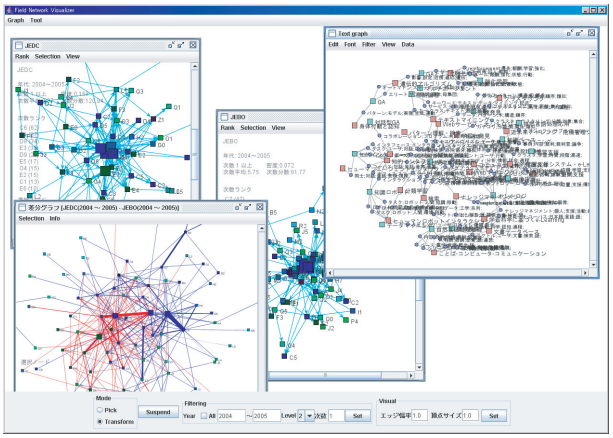


図 4: 視覚化・分析ツール

表 1: JSAI 2001-2003 間 中心性（次数）

入次数		出次数	
Web マイニング	16	ことば-コンピュータ-コミュニケーション	8
オントロジ	8	コミュニティー・知識共有	6
知識発見	7	強化学習	6
自然言語処理基礎	7	知的インタフェース・対話モデル	6
知識ロボット	6	検索	5
情報検索	6	ヒューマンロボットインタラクション	5
強化学習	5	Web サービス・オントロジー	4
知的インタフェース	5	ナレッジマネージメント	4
画像認識	4	テキストマイニング	4
GA	3	知的学習支援システム・e-Learning	3

(a) 2001-2003 間の分野間ネットワーク

2003 年から 2001 年の方向でリンクを生成し，分野間ネットワークを生成した．生成したネットワークにおける中心性（入次数，出次数）を表 1 に示す．表 1 を見ると 2001 年の Web マイニングについて特に中心性（入次数）が高い．また，Web マイニング (2001) まわりのネットワークを示した図 5 を見ると視覚的にも多くのリンクが接続されており，Web マイニングを中心に典型的なハブ型の形状をしている．Web マイニングは，テキストマイニング，コミュニティー・知識共有，パターン理解・検索，ナレッジマネージメントといった分野とリンクされており，これらの分野との密接な関係をみることができる．

また “ことば-コンピュータ-コミュニケーション (2003)” と “自然言語処理基礎 (2001)”，“強化学習 (2003)” と “強化学習 (2001)” の間でリンク数が多く (多リンク型)，これらの分野間リンクは自己分野発達型の分野として理解することができる．

表 2: 中心性（上位 10 件）

JSAI2003		JSAI2006	
タイトル	中心性（出次数）	タイトル	中心性（出次数）
ヒューマンエージェントインタラクション	10	知識獲得支援	11
Web マイニング	9	AI 応用	10
ヒューマンロボットインタラクション	9	知識発見・データマイニング	10
AI 応用	8	Web コミュニティ	9
強化学習	8	ブログ	9
知識発見・データマイニング	7	コミュニケーション支援	9
分類学習	7	人工知能の産業応用	8
知識獲得支援	6	コミュニケーション支援	7
エージェント	6	データマイニングの実践	7
情報検索・抽出・分類	6	コンテンツ作成支援	7
次数平均	10	次数平均	10.39
密度	0.435	密度	0.346
次数分散	31.58	次数分散	28.69

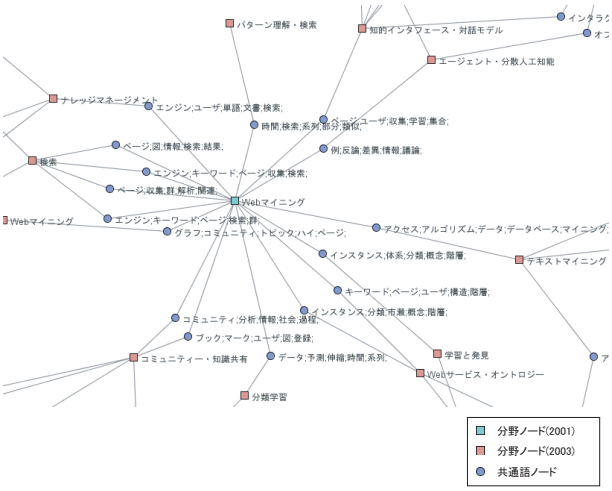


図 5: JSAI 2001-2003 間における Web マイニング (2001) のエゴセントリックネットワーク

さらに “ゲノム (2001)” と “ニュートラルネット (2001)” は孤立ノードであった．これらの分野は，2003 年には他分野との関連がなくなっており，2001 年-2003 年間においては時間発達的に孤立した分野になっていると解釈することができる．

(b) JSAI 2003-2006 間の中心性

2003 年と 2006 年の分野間ネットワークを生成した．中心性（出次数）（表 2）をみると，2003 年では，ヒューマンエージェントインタラクション，ヒューマンロボットインタラクションを主研究キーワードとする研究が合計 19 件あり、その盛況振りが伺える．一方 2006 年は，Web コミュニティ，ブログ，コミュニケーション支援など，近年のブログ，SNS などの流行が研究の盛り上がりに大きく貢献していることがよくわかる．

5. おわりに

文献情報に基づき，分野をノードとしたジャーナル・学会の分野分布の比較および分野の変遷を見るのためのネットワーク表現を提案し，分析ツールの開発と分野間ネットワークに関する分析を行った．テキスト情報に基づくネットワークは分野の変遷および学会同士の分野の対応関係を示し，今後分野の分類議論に役立つだろう．

特に，毎年行われる全国大会などのイベント空間においては，開催側における年度別の研究分野の変遷の把握や，ユーザ側においては，論文の検索，論文投稿時の判断などに利用でき，また学会間の違いの明確化や，他分野の研究理解など，研究者にとってより有益な情報を与える一つの有効なアプローチであると考えている．

参考文献

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

[難波 06] 難波英嗣，谷口裕子: 学術論文データベースからの研究動向情報の抽出と可視化, 言語処理学会 第 12 回年次大会 併設ワークショップ「言語処理と情報可視化の接点」(2006).

[Anegón 05] Anegón, M., Rodriguez, Z. C., Quesada, B. V., Álvarez, E. C.,Solana, V. H., Francisco Fernández, F. J. : *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).

[ACM] ACM `http://www.acm.org/`.

[Kamada 89] Kamada, T. and Kawai, S. : *An algorithm for drawinggeneral indirect graphs*, Information Processing Let-ters, No.31, pp.7-15 (1989).

[松尾 05] 松尾豊，友部博教，橋田浩一，中島秀之，石塚満: Web 上の情報からの人間関係ネットワークの抽出, 人工知能学会論文誌，Vol.20，No.1 (2005).

[安田 06] 安田雪，松尾豊，武田英明: 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).