DISCUSSION PAPER No. 110

国際学会に注目した萌芽的研究の発展過程分析
－ World-Wide Web Conference の事例分析 －

2014 年 11 月

文部科学省　科学技術・学術政策研究所

科学技術動向研究センター

古川 貴雄　森 薫　有野 和真

林 和弘　白川 展之　野村 稔

本 DISCUSSION PAPER は、所内での討論に用いるとともに、関係の方々からのご意見をいただくことを目的に作成したものである。

また、本 DISCUSSION PAPER の内容は、執筆者の見解に基づいてまとめられたものであり、機関の公式の見解を示すものではないことに留意されたい。

DISCUSSION PAPER No. 110

An Analysis of Evolutionary Process on Emerging Research Focusing on International Confereces
— A Case Study of the World-Wide Web Conferences —

Takao FURUKAWA, Kaoru MORI, Kazuma ARINO,
Kazuhiro HAYASHI, Nobuyuki SHIRAKAWA, Minoru NOMURA

November 2014

Science and Technology Foresight Center
National Institute of Science and Technology Policy (NISTEP)
Ministry of Education, Culture, Sports, Science and Technology (MEXT)
Japan

国際学会に注目した萌芽的研究の発展過程分析
－ World-Wide Web Conference の事例分析 －

文部科学省 科学技術・学術政策研究所 科学技術動向研究センター
古川 貴雄　森 薫　有野 和真　林 和弘　白川 展之　野村 稔

要旨

本調査研究では、計算機科学の中でも応用研究の傾向が顕著なウェブ関連研究を例に、当該領域における萌芽的研究の発展過程を分析する手法を提案し、その有用性について検討する。2002 年から 2011 年に開催された World-Wide Web カンファレンスのセッションを取り上げ、プロシーディングペーパーのアブストラクトを用いたテキスト分析により、セッション間を接続するネットワークを生成した。その結果、萌芽的な研究と考えられるソーシャルネットワークやマネタイゼーション研究の発展する過程が示された。さらに、カンファレンスセッションの時系列ネットワーク分析により次の知見が得られた。(1) 過去のセッションとの接続が多い収束セッションは、過去の研究トピックを統合したと考えられる。(2) その後のセッションとの接続が多い分岐セッションは、他の研究に影響を与えたセッションと考えられる。テキスト分析の安定性などの課題は残るが、提案手法は萌芽的研究の発展過程の分析に有用と考えられる。

An Analysis of Evolutionary Process on Emerging Research Focusing on International Confereces — A Case Study of the World-Wide Web Conferences —

Takao FURUKAWA, Kaoru MORI, Kazuma ARINO,
Kazuhiro HAYASHI, Nobuyuki SHIRAKAWA, Minoru NOMURA
Science and Technology Foresight Center, National Institute of Science and Technology Policy (NISTEP), MEXT

ABSTRACT

As an example of web-related research that shows remarkable tendency of applied research in computer science, this study proposes a method to analyze evolutionary process of emerging research and discusses the availability. This paper took up organized sessions at the World-Wide Web conferences held from 2002 to 2011 and created the networks connecting sessions based on textual analysis of proceedings papers' abstracts. The results unveiled the evolutionary processes of social networks and monetization studies that were considered as emerging research. Furthermore, the chronological network analysis of conference sessions made the following findings. (1) Convergent sessions nodes that have many links to the past sessions integrate the research topics in the past (2) Divergent session nodes that have many links to the succeeding sessions affect other research content. Although stability problems on textual analysis still remain, the proposed method is considered to be useful to analyze the evolutionaly process of emerging research.

目次

概　要

1. 調査研究の目的

科学技術政策のベンチマーキングに科学技術動向の定量分析は不可欠であり、これまでに基礎科学と中心とする研究領域については、共引用分析等の計量書誌学を用いた分析が行われている。しかし、工学領域のように基礎科学の研究領域と比較して学術文献の引用回数が比較的少ない領域については、共引用分析だけで研究の動向を正確に把握することは容易でない。また、共引用分析の場合、学術文献が引用されるまでに時間を要することから、その研究領域における萌芽的研究の動向を正確に把握することは困難である。本調査研究では、学術文献の引用回数が基礎科学領域に比較して少ないとされる計算機科学を取り上げ、その中でも応用研究の傾向が顕著なウェブ関連研究を例に、当該領域における萌芽的研究の発展過程を分析する手法を提案し、その有用性について検討する。

2. 萌芽的研究の発展過程を分析する手法

2.1 学術文献の分析手法

学術文献の代表的な分析手法である共引用分析とテキスト分析の特徴を概要表 1 にまとめる。学術文献のテキスト分析は、文献に記載された単語の出現頻度等から論文間の関係を生成し、これまでに把握されていなかった潜在的な知識の抽出に利用されている。ここでは、学術文献間の引用関係等の情報を必要とせず、最新の研究成果の分析に適したテキスト分析手法を用いる。

概要表 1　学術文献分析における共引用分析とテキスト分析の比較

	共引用分析	テキスト分析
(1) 基本データの構造	学術文献間の**引用関係**を示す**構造化データ**である。	**非構造**の**テキストデータ**である。
(2) 学術文献間の関係	引用関係によって**直接的**、かつ、**明示的**に示されている。	テキスト分析によって**間接的**に学術文献間の関係を生成するため、**明示的に示されていない**。
(3) 分析結果の安定性	**共引用関係**を用いるため、分析結果が**安定**している。	**テキスト分析に依存**するため、分析結果が**安定しているとは言えない**。
(4) 分析における情報探索範囲	引用・被引用文献に限定されるため、基本的に**論文著者の有する知識の範囲**に制限される。	収集したデータ全体を網羅するため、論文著者に**認識されていない潜在的な知識**も含まれる。
(5) 迅速性	ある**学術文献が公表されてから**、他の学術文献に**引用されるまでに一定の期間**を要する。	**学術文献が公表された段階で**、即時に分析に用いる**テキストデータが得られる**。

2.2 カンファレンスセッションに注目した分析

プロシーディングペーパーは、ジャーナルペーパーよりも公表されるまでの期間が短いため、速報性が高いとされている。他の研究領域と比較して計算機科学の研究領域では、プロシーディングペーパー比率の高いことが知られている。そこで、カンファレンスで発表されたプロシーディングペーパーとカンファレンスセッションに注目した分析手法を提案する。ここでは、カンファレンスセッションの名称が研究内容を表現する場合の抽象度や粒度として適切であると仮定し、カンファレンスセッションの時系列変化から萌芽的研究の発展過程を分析する。

2.3 カンファレンスセッションのネットワーク生成方法

最初に、発表された各論文のアブストラクト中の単語の出現頻度を用いて論文間の類似度を計

算する。次に、各セッションで発表された論文間の類似度からセッション間の類似度を計算する。最終的に、セッション間の類似度が設定値よりも高い場合にそれらのセッションを接続し、セッションの時系列変化を示すネットワークを生成する。

3. 分析データと分析結果

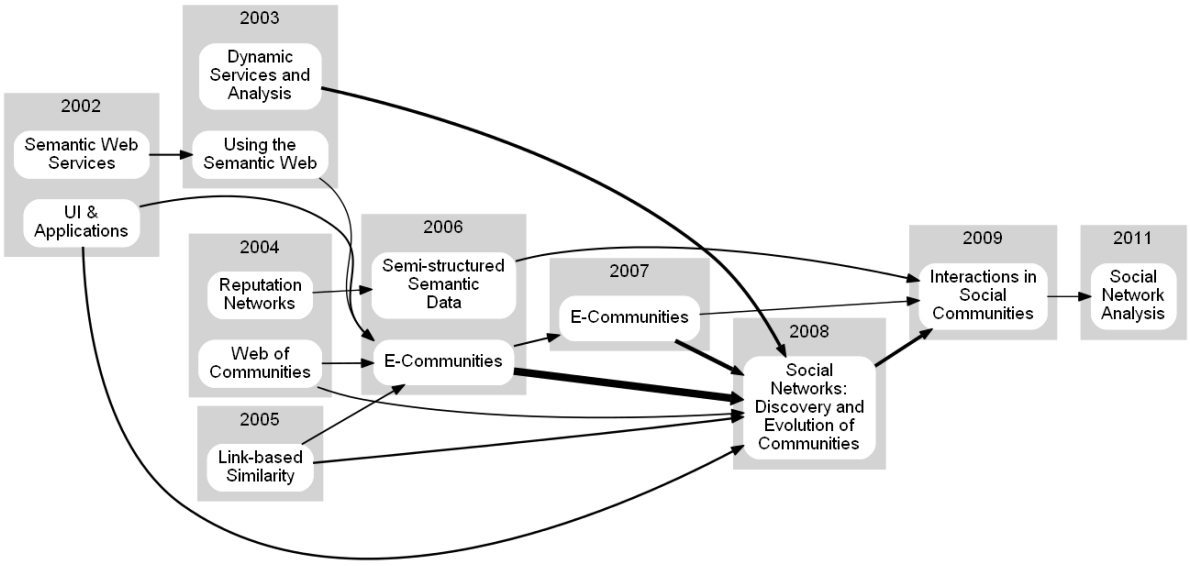
3.1 分析データ

2002 年から 2011 年に開催された World-Wide Web (WWW)カンファレンスを調査し、894 件のプロシーディングペーパーと295 件のセッションに関する情報を収集した。次に、注目すべきセッションを選択し、このセッションに関連性の高いセッションを接続するネットワークを生成して、萌芽的研究の発展過程を分析した。

3.2 分析結果

3.2.1 過去のソーシャルネットワーク研究を総括する収束セッションノード

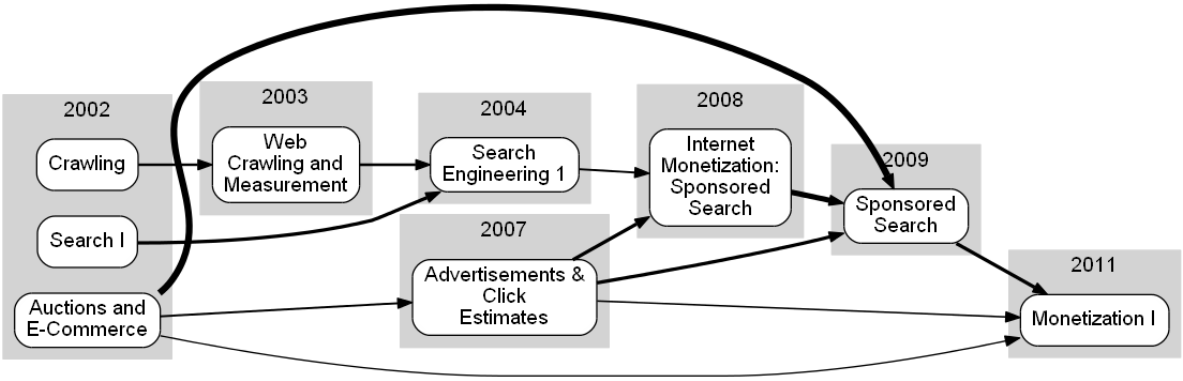
概要図 1 の 2008 年 Social Networks: Discovery and Evolution of Communities のように、過去のセッションとの接続が多いセッション(収束ノード)は、過去の研究トピックを総括したセッションと考えられる。また、2004 年の Web of Communities、2006 年、2007 年の E-communities といったセッションで発表された研究はソーシャルネットワーク研究に発展に寄与したことが推察される。



概要図 1　2011 年の Social Network Analysis セッションに至る時系列ネットワーク。セッション間類似度の大きさを矢印の太さに反映させた。

3.2.2 マネタイゼーション研究の発展に寄与する分岐セッションノード

概要図 2 に示す 2002 年の Auction and E-Commerce や 2007 年の Advertisements & Click Estimates は、その後のセッションとの接続が多いセッション(分岐ノード)であり、他の研究に影響を与えたセッションと考えられる。マネタイゼーションという名称は、計算機科学における一般的な研究トピックとは考えにくいが、これらのセッションで発表された研究がマネタイゼーション研究に発展したことが概要図 2 から示唆される。



概要図 2　2011 年の Monetization I セッションに至る時系列ネットワーク。セッション間類似度の大きさを矢印の太さに反映させた。

4. おわりに

提案手法の特徴は、研究者コミュニティにおける新たな研究領域を開拓しようとする意思や将来展望が反映されたと考えられるカンファレンスセッションに注目し、萌芽的研究の発展過程を可視化した点にある。個々の論文よりも抽象度の高いセッション名を扱うことで、最新の研究動向を容易に把握できるようになった。カンファレンスセッションの時系列ネットワーク分析により、過去の研究を総括するような収束セッションノードと、その後の研究に影響を与えたと思われる分岐セッションノードの存在が示された。

テキスト分析における安定性などの課題が残されているものの、提案手法は、萌芽的研究の発展過程を分析する手法として有用と考えられる。ただし、セッションの推移を可視化した結果の妥当性については、当該研究領域の研究者にインタビューするなど定性的な評価を行う必要があろう。調査研究の実施段階では、分析に用いたすべてのプロシーディングペーパーがデータベースに収録されておらず、計量書誌学的な手法とは結果を比較できなかったが、今後は共引用分析等の手法で得られた結果との比較も課題として挙げられる。萌芽的研究として抽出されたセッションから、そのセッションで発表された論文の著者や、セッションチェア等の研究者情報も取得できるため、これらの研究者情報を用いた共著者分析についても興味が持たれる。また、医学、物理学、化学領域の代表的な学会の年次大会についても、名称の付与されたセッションが開催されていることから、提案手法が他の研究領域にも適用できる可能性が示唆された。

1. 調査研究の背景と目的

1.1 科学技術動向の定量分析

科学技術政策のベンチマーキングに科学技術動向の定量分析は不可欠であり、これまでにサイエンスマップ等の分析手法が提案されている[1-5]。サイエンスマップの場合、論文データベースを用いて、基礎科学を中心とする研究の動向を俯瞰的に捉えるとともに、論文数の増加する注目研究領域の抽出と、それら研究領域の時系列変化を観測することを目的としている。これらの研究では、サイエンスマップを生成する過程で、論文間の引用関係や論文の共著関係等の計量書誌学で扱われるデータを分析している。学術論文や特許などの学術文献を分析する手法として、共引用分析に加えて、テキスト分析[例えば 6]も、情報処理技術の発展に伴って利用される機会が増えている。

1.2 学術文献分析

学術文献分析における共引用分析とテキスト分析の比較した例[7]を表 1 に示し、その基本的な特徴を以下にまとめる。

共引用分析の場合、学術文献間の引用関係という明示的、かつ、明確な構造をもつ基本データを用いることからあいまいな要素が含まれず、結果として分析結果は安定する傾向がある [1]。しかし、分析対象となるのは共引用関係の示された学術文献に限られるため、情報の探索範囲は基本的に論文著者の有する知識の範囲に制限されてしまう。また、学術文献が公表されてから他の学術文献に引用されるまでに一定の期間を要するため、最新の研究よりも過去に行われた研究の動向を分析することになりがちである。

一方、テキスト分析の場合、非構造のテキストデータを対象とするため、引用関係のような明示的な情報を扱うことはなく、自然言語処理や機械学習を用いて生成した間接的な学術文献間の関係を分析をすることが多い。学術文献分析の結果はテキスト分析手法に依存するため、分析の条件を統一しない限り、常に安定した結果が得られるとは限らない。しかし、テキスト分析における情報探索範囲は、収集したデータ全体を網羅するため、研究者にも認識されていない潜在的な知識を抽出できる可能性のあることが指摘されている[6]。例えば、異なった研究領域に分類され、関連性が知られていなかった知識を顕在化することが期待されている。また、学術文献が公表された段階で、基本データとなるテキストデータが存在するため、最新の研究成果を対象とした分析が可能となる。

[1] 学術文献を分類する場合、クラスタリングアルゴリズムや設定値等により結果が異なるため、クラスタリング結果の安定性について留意する必要がある。

表 1　学術文献分析における共引用分析とテキスト分析の比較

	共引用分析	テキスト分析
(1) 基本データの構造	学術文献間の**引用関係**を示す**構造化データ**である。	**非構造**の**テキストデータ**である。
(2) 学術文献間の関係	引用関係によって**直接的**、かつ、**明示的**に示されている。	テキスト分析によって**間接的**に学術文献間の関係を生成するため、**明示的に示されていない**。
(3) 分析結果の安定性	**共引用関係**を用いるため、分析結果は**安定**している。	**テキスト分析に依存**するため、分析結果が**安定しているとは言えない**。
(4) 情報探索範囲	引用・被引用文献に限定されるため、基本的に**論文著者の有する知識の範囲**に制限される。	収集したデータ全体を網羅するため、論文著者に**認識されていない潜在的な知識**も含まれる。
(5) 迅速性	ある**学術文献が公表されてから**、他の学術文献に**引用されるまでに一定の期間**を要する。	**学術文献が公表された段階で**、即時に分析に用いる**テキストデータが得られる**。

1.3　革新的技術と科学技術ロードマップ

革新的技術を形成する科学に基づくイノベーションは、新しい産業を創出し、既存産業を転換する可能性を秘めている[7]。革新的技術は、既存技術の改善の積み重ねとは根本的に異なり、技術的な不連続な変化をもたらすとともに、従来の産業、市場、企業に対して破壊的なインパクトを与えている。その点で、革新的技術への戦略的な研究開発(R&D)投資は、産業競争力を効果的に高めることに寄与することから、企業の経営層をはじめとする民間セクターだけでなく、政策立案者や行政からも注目を集めている。既存技術の積み重ねの場合、連続性、均衡性、合理性、最適性といった前提に基づいた議論も有効であるが、革新的技術に対してこれらの前提を仮定するのは適切とは言えない[8]。従って、これらの前提を仮定し、既存データの外挿による将来予測に合理性を見い出すことは困難である。急速に進展するR&Dの最先端領域から革新的技術を検出し、識別することは専門家にとっても極めて困難な課題と言える。

様々な科学技術 R&D とその応用について構造的な関係を記述する科学技術ロードマップは、公的・民間セクターの両方で戦略的 R&D の立案に用いられてきた[9]。ロードマップを作成するロードマッピングの過程では、専門家の集団における新たな集合知を創出する手法と、コンピュータを用いた学術文献の自動分析によって抽出した知識を活用する手法が組み合わされる。近年では、破壊的なインパクトをもたらす革新的技術についてもこのような手法が適用されつつある[10,11]。今後も、ロードマッピングにおいて、専門家の知識は不可欠であることに変わりはないが、学術文献の爆発的な増加、計算機科学の進展や計算機ハードウェアの性能向上を考慮すると、情報処理技術を利用した分析手法の重要性がより高まると考えられる。

1.4　本調査研究の目的

本調査研究では、科学技術動向の定量的な分析手法の確立に向けて、急速に発展する研究領域における萌芽的な研究トピックを検出、識別し、その発展を過程を分析する手法を提案するとともに、その有用性について検討する。また、情報処理技術を利用した科学技術ロードマッピングに資する新た

な学術情報の分析手法としても検討を行う。これまでに基礎科学と中心とする研究領域については、共引用分析等の計量書誌学を用いた分析が行われている。しかし、工学領域のように基礎科学の研究領域と比較して学術文献の引用回数が比較的少ない領域については、共引用分析だけで研究の動向を正確に把握することは容易でない。また、共引用分析の場合、学術文献が引用されるまでに時間を要することから、その研究領域における萌芽的研究の動向を正確に把握することは困難である。本調査研究では、学術文献の引用回数が他の領域に比較して少ないとされる計算機科学を取り上げ、その中でも応用研究の傾向が顕著なウェブ関連研究を例に、当該領域における萌芽的研究の発展過程を分析する。

革新的技術は何らかの破壊的なインパクトをもたらしていることに間違いはなく、それによって我々の生活は劇的に変化している。情報通信技術(ICT)の進歩は代表的な事例であり、その中でも特にインターネット、ウェブ、モバイル関連技術は、パーソナルコミュケーションだけでなく商業や製造業におけるビジネスコミュニケーションの形態も大きく変化させている。ここでは、影響が大きく、かつ、広範にわたるウェブ関連研究を調査対象に選択した。

計算機科学の領域では、プロシーディングペーパーの比率がジャーナルペーパーと比較して高く、さらに、プロシーディングペーパーは研究成果の速報性も高いことが指摘されている[12,13]。そこで、本調査研究では、、カンファレンスで発表されたプロシーディングペーパーとカンファレンスセッションに注目した分析手法を提案する。カンファレンスは、研究者コミュニティにおける最先端の知識を共有し、さらに新たな知識を創出する機会を提供する場であることも、分析対象として選択した要因である。また、萌芽的研究の発展過程を分析する上で、カンファレンスセッションに使用される名称が研究トピックの抽象度や粒度として適切であると仮定し、カンファレンスセッションの時系列変化についてネットワーク分析を行う。

本報告書の構成を以下に示す。第 2 章では、ロードマッピング等に利用される学術文献の定量的な分析手法について先行研究を紹介する。第 3 章では、テキスト分析手法を用いて研究トピックの時間的な変化を可視化する手法について述べる。第 4 章では、カンファレンスセッションの分析から得られた時系列ネットワークを示し、萌芽的研究の発展過程の事例について検討する。第 5 章に提案手法と特徴を示し、第 6 章で本研究の結果と意義をまとめる。

2. 学術文献の分析手法

科学技術に基づく研究成果は、経済成長を加速する技術的なイノベーションを引き起こす要因とされ[14,15]、公的・民間セクターのいずれも最新の研究動向を注視している。科学技術ロードマップは、公的・民間セクターにおける R&D 戦略を策定するための意思決定に有用な資料として活用されることが期待されてきた[9,10]。例えば、特許と技術の構造的な関係を記述した技術ロードマップは、企業間の共同研究や特許のクロスライセンス等の戦略的な意思決定に利用されている[16]。科学技術ロードマップを作成する過程で、学術論文や特許に代表される学術文献の分析から抽出される最新の研究トレンドや、専門家によるミーティング、パネル、ワークショップ等における議論から形成される集合知が活用される[9]。現在では、科学技術に基づく研究成果は学術文献として発表されるとともに、デジタルデータとしてデータベースに蓄積されることが一般化している。科学技術ロードマップの作成には、依然として専門家の議論から形成される集合知が不可欠であるが、学術文献の爆発的な増加や、計算機科学の進展を考慮すると、情報処理技術を利用した分析手法の重要性が高まると考えられる。専門家であっても、様々な分野に分かれた膨大な学術文献を調査することは容易ではないため、データベースに蓄積された膨大な学術文献の中から革新的技術を自動的に抽出、識別し、分析する手法が、科学技術ロードマップ作成の効率化に寄与することが期待される。既存の学術文献分析手法は、計量書誌学的な手法、テキスト分析、両者の混合手法に分けることができる。さらに、既存研究の分析結果は、研究トピックの抽出と研究トレンドの分析に分類できるため、以下ではこれらの手法について整理する。

2.1 計量書誌学的な分析とテキスト分析

2.1.1 計量書誌学的な分析

計量書誌学的な分析では、引用文献、共著者、所属組織等の学術文献に特有の情報が用いられる。共引用文献を用いた分析では、論文や特許を接続した階層的なネットワーク構造を生成し[17]、類似した論文や特許等を含む共引用文献のクラスタを生成し[18-28]、研究トピックの抽出や識別を行っている。新たな研究領域を開拓するような影響の大きな文献は、共引用文献を接続するネットワークにおいて、多くの被引用文献と接続されたハブを形成する傾向がある。共引用分析はネットワーク構造を基盤にした強力、かつ、効果的な分析手法であるものの、学術文献が公表されてから他の学術文献に引用されるまでに一定に期間を要するという問題、また、引用論文の名寄せ処理等の学術文献間を接続する Linked data を生成するためのコストが高いという問題もある。結果として、データベースに引用文献も含めて登録され、データとして整備されるまでに時間がかかることから、急速に発展する研究領域の調査に共引用分析が最適であるとは言い難い。

2.1.2 テキスト分析

テキスト分析は、計量書誌学的に手法に代わる学術文献の分析手法であり、文献に記載された単語の出現頻度等から論文間の関係を生成し、これまでに把握されていたなかった潜在的な知識

を抽出することを目的に利用されることが多い。テキスト分析の場合、共引用関係や共著関係などの計量書誌学的な情報からは関連性を見出すことの困難な、異分野に分類されるような論文間の関係を提示できるという特徴がある[6]。従って、共引用分析にように明示的、かつ、明確な論文間の関係に対して、テキスト分析では、明示的ではない曖昧な論文間を関係も扱うことができる。このような関係を用いて複数の学術文献に含まれる潜在的な知識を抽出する手法は、Literature-Based Discovery (LBD) [29]と呼ばれている。これまでに、医学系論文のテキスト分析により、専門家にも知られていない知識を自動的に抽出することを目指したコンピュータ支援 LBD システムが開発されている[30-32]。初期のテキスト分析では、論文に記載された単語や単語が接続されたフレーズの出現頻度の傾向から、当該領域における注目すべき研究トピックが抽出された[33]。その後、テキスト分析手法は初期の簡単な単語やフレーズ分析から、概念抽出を目指した複雑かつ高度な手法に発展してきた。例えば、Latent Semantic Analysis (LSA)は、論文等の文書データから生成した単語の出現頻度行列に特異値分解を適用して、関連した単語グループに対応する上位概念を生成する手法[34]である。これまでに医学系論文に LSA を適用した例が報告されている[35]。Latent Dirichlet Allocation (LDA)は、より統計的に洗練された手法であり、テキスト分析によって抽出された論文クラスタに対応する概念を抽出するために用いられている[36,37]。この手法は、ベイズ理論に基づいたトピックモデルという手法に発展し、現在は様々な研究や応用が進められている[38-40]。

2.1.3　混合手法

計量書誌学的な手法とテキスト分析を組み合わせた混合手法も提案されている[41]。初期の混合手法では、計算量の大きなテキスト分析の効率を改善するために、引用分析の結果を用いてテキスト分析の対象となる論文数を限定している[42]。また、テキスト分析に焦点を当てた LDB と、学術文献の著者情報を用いて抽出した専門家によるワークショップ等において創出された集合知を利用する Literature-Assisted Discovery (LAD)を統合した方法も混合手法とされている[43,44]。LBD と LAD は、それぞれ、科学技術ロードマッピングにおける情報処理技術を用いた学術文献分析による知識抽出と、専門家による集合知の創成に対応している。LRD における学術文献分析は、(i) 中核となる文献の抽出、(ii) 直接的に関係する論文の抽出と分析、(iii) 間接的に関係する論文の抽出と分析の 3 段階からなる[45]。なお、LRD は、異った研究領域に分類されている論文間に存在する潜在的な関係の抽出に利用されている[46,47]。

2.2　研究トピックと研究トレンドの分析

2.2.1　研究トピック分析

前述した複雑かつ高度な計量書誌学的な手法やテキスト分析を用いても、学術文献間の関係から生成された学術文献クラスタのもつ意味や対応する概念を把握することは依然として容易でない。医学系論文を対象とした MEDLINE データベースの場合、登録された論文に Medical Subject

Headings(MeSH)タームと重要度指数を割り当てている[48]。MeSH タームは、個々の著者が論文に記述したキーワードとは別に、専門家の間で認識が統一されたキーワードであり、論文の分類や論文クラスタに対応する概念の把握に有用とされている[49]。例えば、急速に発展している医学系研究トピックを抽出するために、MeSH タームと共著者ネットワークを組み合わせて分析した例が報告されている[50]。医学系領域における最近の LRD 研究では、MeSH タームを用いた因子分析と階層的なクラスタリングを組み合わせた研究トピックの抽出方法が検討されている[51]。しかし、MeSH タームを用いた分析でも、結果の確認作業に専門家は不可欠であり、個々の論文クラスタに対応する研究トピックに対応する概念を自動的に生成するボトムアップアプローチが難しいことに変わりはない。

特許分析の研究では、特許クラスタの内容を示すラベルを自動生成する包括的なテキスト分析手法が提案されている[52]。この手法では、特許クラスタに含まれる文書について、単語の同時生起関係を分析して特許クラスタを代表するラベルを決定する。しかし、ラベルは特許文書に含まれる単語に限定されるため、抽出された単語だけでは表現の困難な包括的な概念を扱うことはできない。また、これまでに用いられている単語との対応が明確でない未定義の概念を扱うことも困難である。LSA や LDA などの数理的に洗練された手法であれば、これらの問題を解決する可能性があるものの、学術文献クラスタのラベリングが容易ではないことに変わりはない。

2.2.2 研究トレンド分析

時間的な変化に注目した学術文献分析の研究もこれまでに行われている。以下では、これらの研究トレンド分析手法を紹介する。

計量書誌学的な手法では、共引用論文クラスタの成長曲線[22,23]や時間発展 [25-27]を調べ、特定の研究トピックに関する論文の増加傾向を詳細に分析した例がある。例えば、論文[26]や特許[27,28]の共引用文献クラスタのタイムラインチャートを描くことにより、研究トピックの分岐、統合、移行に焦点を当て、革新的技術の発展過程が分析されている。

テキスト分析では、技術経営系カンファレンスのプロシーディングペーパーのアブストラクトから単語出現頻度の時間変化を抽出し、新規研究トピックとそのの変遷を分析した例が報告されている[53,54]。また、デルファイ予測調査法のワークショップで取り上げられたトピックタームを用いて検索した論文数の時系列変化に対して、成長曲線をフィッティングすることにより、研究トレンドを定量的に示した例もある[55]。PubMed データベースから抽出した論文について、月間の論文数とトピックタームの出現頻度を分析し、新しい研究トピックを抽出した例も報告されている[56]。他にも、テキスト分析によって生成された特許クラスタのタイムラインチャートを用いて研究の発展過程を可視化した例がある[57]。

混合手法では、共引用論文の間に密接な関係があること利用し、共引用ネットワークとテキスト分析を組み合わせて研究トピックの時間的変化の検出精度を向上させた例が報告されている[58]。その他に、共引用論文ネットワークのタイムラインチャートに、テキスト分析を用いて抽

出したラベルを付加して研究トレンドを分析した例もある[59]。さらに、共引用分析に基づく手法でも、抽出された論文クラスタについてキーワードの分析を行い、その研究領域を表現するラベルの自動生成や、論文クラスタの時系列変化に関する分析が行われている[5]。

これらの研究トレンド分析手法は、注目すべき研究領域の発展過程の分析には有用と考えられるが、結果を解釈する段階で専門家の知識が不可欠であることに変わりはない。

3. 時系列ネットワークの生成方法

3.1 分析データ

ウェブ関連技術は、この 20 年間に急速した革新的技術の一つである。本調査研究では、ウェブ関連技術に焦点を当て、当該研究領域においてトップランクに評価される WWW カンファレンスを取り上げ、萌芽的研究の発展過程を分析する。WWW カンファレンスの場合、セッション名はウェブ上で公開されているプログラムに記載されている。なお、調査段階では、カンファレンスのセッション名はIEEE やACM などの情報通信や計算機科学系の学術団体による文献データベースや、Scopus や Web of Science 等の商業学術出版社のデータベースにも収録されていなかった。

2002 年から 2011 年の間に開催された WWW カンファレンスのプログラムや文献データベースを調べ、894 件のプロシーディングペーパーと 295 件のセッションに関する情報を得た。図 1 にプロシーディングペーパー(以下では単に論文と表記)とセッション数の変化を示す。論文数は 75 件から 115 件の間で多少は変動しているが、セッション数はこの期間を通して大きな変化は見られなかった。1 カンファレンス当たりのセッション数は平均で約 30 件あり、1 つのセッションで約 3 件の論文が発表されている。なお、カンファレンスには、ペーパーセッションやレギュラーセッションと呼ばれるセッションに加えて、ポスターセッション等も実施されているが、今回の分析では、セッション名との関係が明確に示されているペーパーセッションに限定して分析を行った。

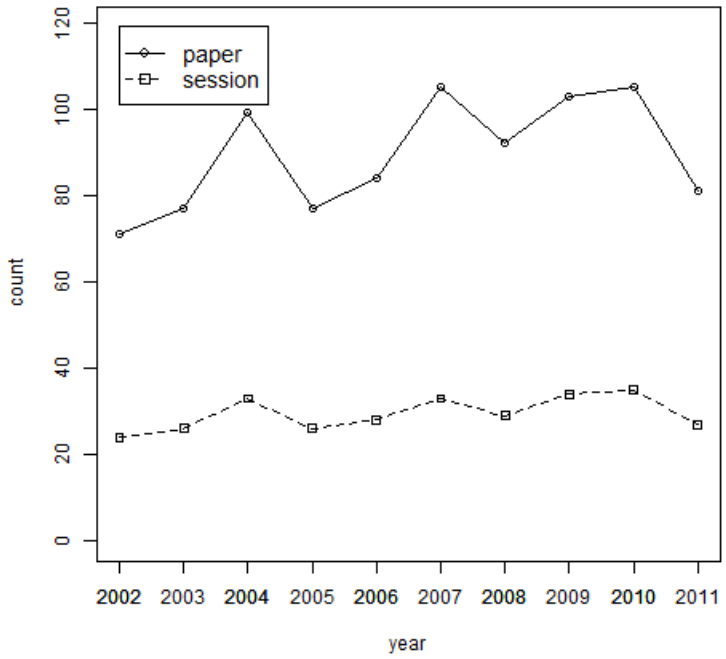


図 1　2002 年から 2011 年に開催された WWW カンファレンスにおける論文数とセッション数の関係

3.2　論文・セッションの類似度

論文は題目、著者、著者の所属、キーワード、引用文献などの属性をもつが、ここでは分析を簡略化するためアブストラクトのテキストデータのみを用いた。まず、論文の内容を要約したアブストラクトの文書データは、term frequency–inverse document frequency (*tf-idf*) [60]の値を要素とするベクトルとして記述する。*tf-idf* は、簡単な単語の出現頻度よりも、特定の文書データに含まれる単語の重要性を強調した指標である。

近年では、*tf-idf* の他にも文書データをベクトルとして記述する手法が提案されている。例えば、LSA [34]や LDA [36]といった手法では、単語の集合から潜在的な意味情報を抽出するために、高次元空間の文書ベクトルを低次元空間に投影して分析を行っている。このような手法の場合、投影される空間の次元など未知の定数を事前に決定しておく必要がある。しかし、未知の定数を決定する方法によって結果が変化するという不安定性もあるため、ここでは、不確定要因を避けるために文書ベクトルの要素として *tf-idf* を用いることにした。

以下に論文間類似度の定義を示し、論文間類似度に基づくセッション間類似度の計算方法を示す。

3.2.1　論文間類似度

tf-idf ベクトルによって記述された論文 i と論文 j をそれぞれ、ベクトル$\boldsymbol{x}_i$と $\boldsymbol{x}_j$と表記し、これらの論文間類似を次のように定義する。

$$s_{i,j} = \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\| \|\mathbf{x}_j\|} \tag{1}$$

tf-idf は非負の値をとることから論文間類似度 $s_{i,j}$ の範囲は 0 から 1 になる。論文間類似度 $s_{i,j}$ が 1 に等しい場合、論文間の *tf-idf* ベクトルの比率が一致し、論文間類似度 $s_{i,j}$ が 0 の場合、2 つの論文間には共通する単語が存在しないことになる。

3.2.2　セッション間類似度

セッション間類似度は、セッションに含まれるすべての論文ペアについて求めた論文間類似度 s_{ij}の平均値と定義する。セッション I とセッション J 間の類似度を次に示す。

$$S_{I,J} = \frac{\sum_{i=1}^{N_i} \sum_{j=1}^{N_j} s_{i,j}}{N_i N_j} \tag{2}$$

ここで、N_I と N_J はそれぞれセッション I とセッション J に含まれる論文数を示す。

3.3　時系列ネットワーク生成アルゴリズム

図 2 に示すアルゴリズムによりカンファレンスセッションの時系列ネットワークを生成する。各セッションはネットワークを構成するノードに対応するため、2 つのセッションノードを接続するエッジの挿入を繰り返すことで、カンファレンスセッションの時系列ネットワークが生成される。時系列ネットワークを生成するアルゴリズムを以下に示す。

(1) 基準年からルートノードとなるセッションを選択する。基準年以外の全セッションノードを接続されるセッションノードの候補とする (図2 (a))。
(2) 各セッション候補について、ルートセッションとの類似度を計算する(図2 (b))。
(3) セッションペアの類似度が設定値よりも大きい場合、セッションノード間を接続するエッジを挿入する。接続されたセッションは候補セッションノードから除く(図2 (c))。
(4) 新たに接続されたリーフノードを選択し、リーフノードが含まれる年のセッションを候補セッションノードから除く。
(5) 各候補セッションノードについて、リーフセッションノードとの類似度を計算する (図2 (d))。
(6) ステップ(5)で計算したセッション間類似度が設定値よりも大きな場合には、これらのセッションノードを接続するエッジを挿入する(図2 (e))。
(7) 全セッションのペアについて接続が確認されるまでステップ(4)に戻って処理を続ける。

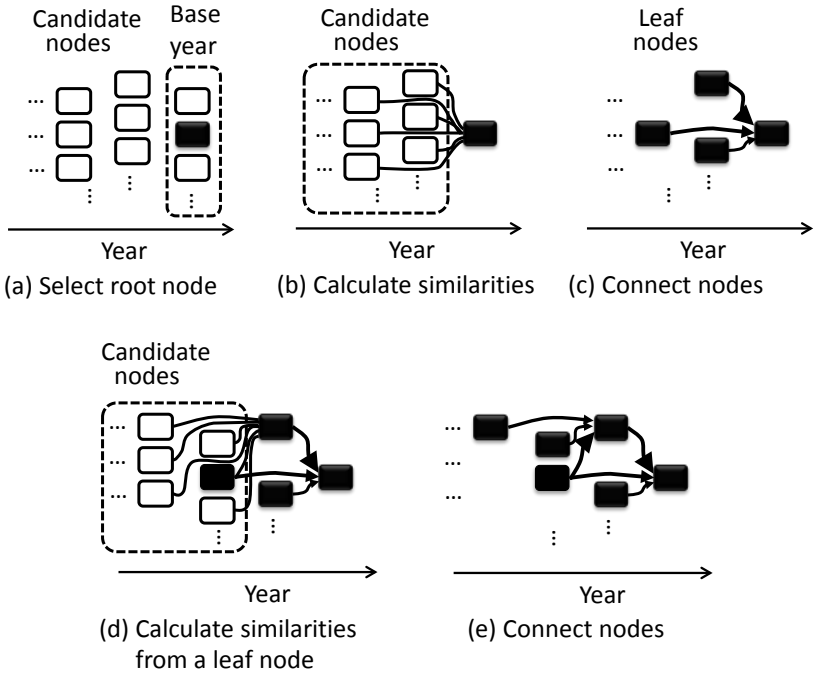


図 2　カンファレンスセッションの時系列ネットワークを生成するアルゴリズム

基準年が調査期間の途中であった場合、前述の後方処理と、時間を反転した前進処理を組み合わせて、セッション間を接続する時系列ネットワークを生成する。

4. 時系列ネットワークの分析例

カンファレンスセションの時系列ネットワークを生成するために、2002 年から 2011 年に開催された WWW カンファレンスのペーパーセッションで発表されたすべての論文からアブストラクトを抽出してテキスト分析をした。本研究では、WWW カンファレンスを代表するような研究の発展過程を明かにするために、次のセッションに注目して分析を行った。

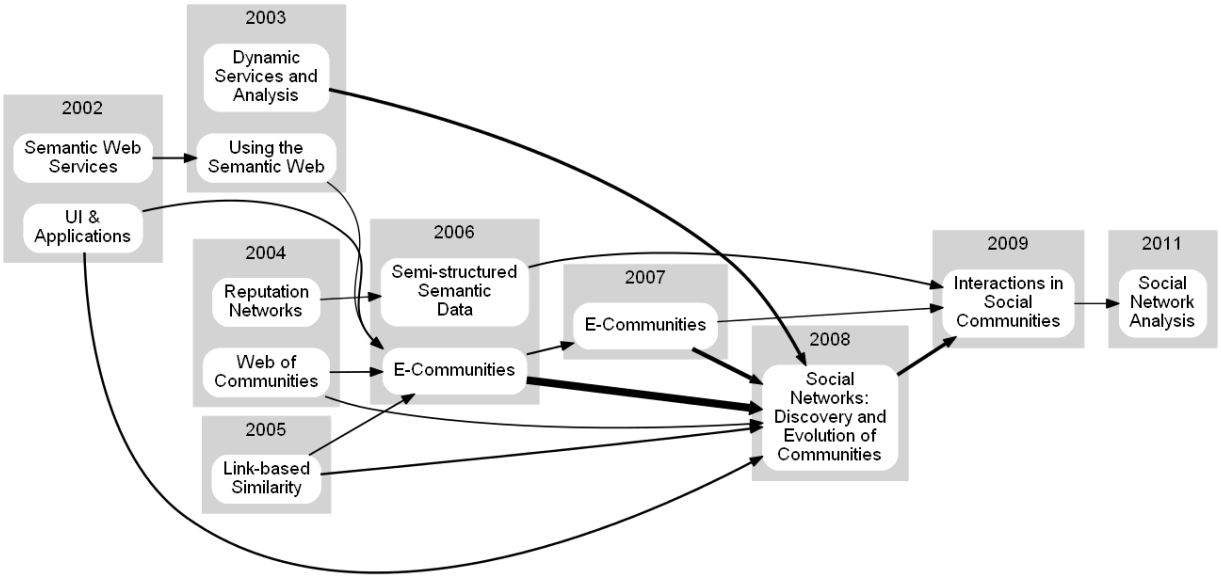
(1) Social Network Analysis (WWW 2011)
(2) Monetization I (WWW 2011)
(3) Semantic Analysis (WWW 2011)
(4) Technology for Developing Regions (WWW 2008)

ソーシャルネットワークサービス(SNS)は我々のパーソナルコミュニケーションだけでなく、ビジネスコミュニケーションにも大きな変化をもたらしている。そこで、2011 年の Social Network Analysis というセッションに注目し、WWW カンファレンスからソーシャルネットワーク研究の発展過程を分析する。2011 年の Monetization I というセッションからは、学術的なカンファレンスにおける研究トピックとは異った印象を受けるため、この研究に注目して、その発展過程を分析する。2011 年の Semantic Analysis というセッションは、WWW 関連研究の中では比較的歴史の長い研究テーマであるため、他の研究トピックと比較するために取り上げた。2008 年の Technology for Developing Regions といセッションは、発展途上国における社会的な問題の解決に寄与する先端技術を扱ったセッションであることが伺える。このセッションからも、既存の研究トピックとは大きく異なった印象を受けたため、この研究の発展過程を分析した。

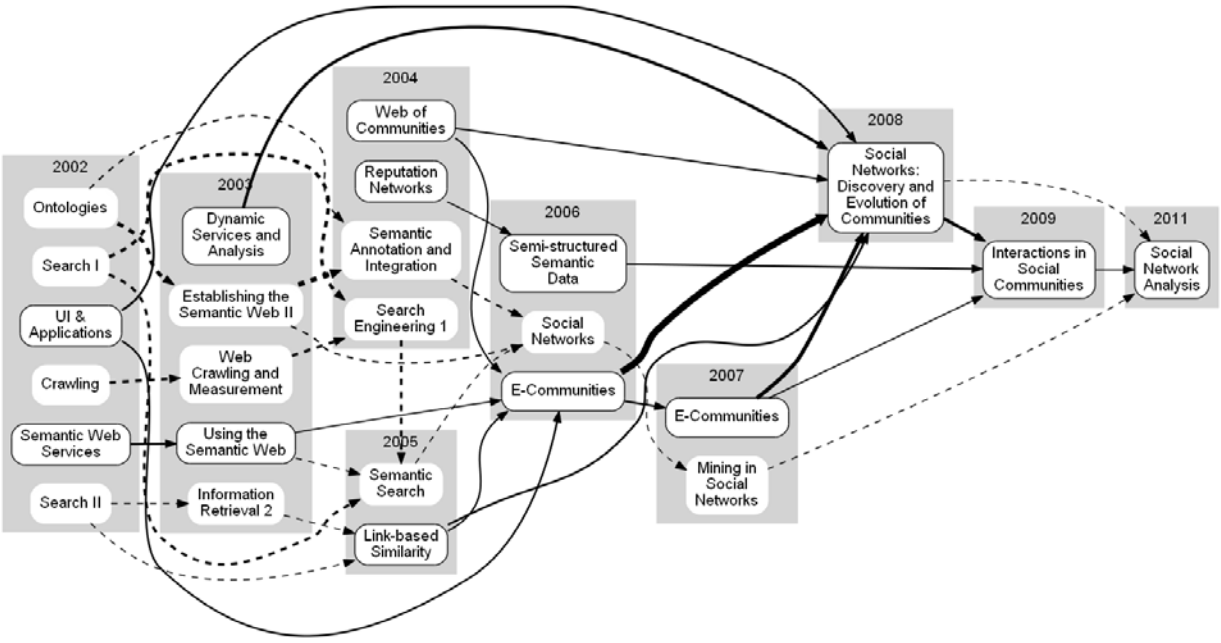
4.1 ソーシャルネットワーク研究を総括する収束セッションノード

2011 年に開催されたWWWカンファレンスのSocial Network Analysisというセッションに注目し、2002 年から 2011 年までの 10 年間に開催されたWWWカンファレンスセッションとの関係を分析して生成した時系列ネットワークを図 3 に示す。図 3 (a)と図 3 (b)に、セッション間の接続を判定するためのセッション間類似度のしきい値を高く設定した結果と低く設定した結果を示した。図 3(a)は図 3(b)の部分集合であり、図 3 (b)の実線で囲まれたセッションノードと、実線の矢印で表示された部分は、図 3 (a)の時系列ネットワークに一致する。図 3(b)の実線で囲まれていないセッションノードと、点線で示された矢印は、図 3(a)に対する付加的な部分ネットワークである。また、セッション間類似度の大きさを矢印の太さに反映したため、太い矢印で接続された 2 つのセッションは類似度が大きいことがわかる [2]。

[2] セッションノードはルートノードに対して順番に接続されるため、木構造のルートから離れた部分(図 3 左側のセッションノード)では、図 3 (a)には存在しない太い矢印で示された類似度の高いセッションが図 3 (b)に含ま



(a) High similarity threshold



(b) Low similarity threshold

図 3　2011 年の Social Network Analysis セッションに至る時系列ネットワーク。セッション間類似度の大きさを矢印の太さに反映させた。実線で囲まれたセッションノードと実線で示された矢印は、上段の時系列ネットワークに一致する。

れることもある。

図 3 の右側に表示された 2011 年の Social Network Analysis というセッションから左側に遡って関連するセッションを辿ると、2009 年の Interactions in Social Communications を経由して、2008 年の Social Networks: Discovery and Evolution of Communities (以下、DEC と省略して表記する。) に辿りつく。図 3 (a)と(b)のいずれも、2008 年の Social Networks: DEC というセッションは、過去の 6 つのセッションに接続されている。さらに、2008 年の Social Networks: DEC から遡ると、2003 年の Dynamic Services and Analysis、2006 年、2007 年の E-Communities に到達する。この時系列ネットワークの構造を見ると、2008 年の Social Networks: DEC というセッションは収束セッションノードとみなすことができる。つまり、過去には分かれていた研究トピックがこのセッションによって統合されたことを示唆している。同様に、2009 年の Interactions in Social Communications や 2006 年の E-Communities も接続されるセッションノードが多いことから、これらも収束セッションノードとみなすことができる。これらの収束セッションノードは、過去の研究トピックを統合し、その後のソーシャルネットワーク研究の発展に寄与した重要なセッションであると考えられる。

図 3 において、2006 年の E-Communities と 2008 年の Social Networks: DEC を接続する太い矢印は、これらのセッションの類似度が高いことを示している。また、2006 年の E-Communities から 2007 年の E-Communities を経由した 2008 年の Social Networks: DEC に至る間接的な経路も存在しているが、これらのセッション間を接続する矢印は太くないためセッション間類似度もそれ程高くはない。2006 年と 2007 年の E-Communities というセッションは同一の名称でありながら、セッション間を接続する矢印は太くはないため、セッション間類似度もそれほど高くはない。この結果は、2008 年の Social Networks: DEC セッションに含まれる研究は、主に 2006 年の E-Communities セッションに含まれた研究から派生した内容であり、2007 年の E-Communities というセッションに含まれた研究との関連性はあまり高くないことを示唆している。さらに、2004 年の Web of Communities を起点とする経路は、ソーシャルネットワーク研究が Web 上のコミュニティ研究から発展したこと示唆している。

表 2 に、図 3(b)の時系列ネットワークに示したセッションと*tf-idf*スコアの高い単語の関係を示す。ここで、*で示したセッションは、図 3 (a)の時系列ネットワークのセッションノードである。*tf-idf* スコアの高い単語が、各セッションで扱われた研究内容を示していることが推測されるため、表の右側に、各セッションで発表された論文のアブストラクトから抽出した*tf-idf* スコアの高い単語を示した[3]。2006 年のE-Communitiesや 2004 年のWeb of Communitiesでは、SNSやblogといったセッションを代表する研究トピックに対応する単語の*tf-idf* スコアが高いことがわかる。図 3 に示した収束セッションノードの 2008 年、Social Networks: DECの場合、community、social、network などが*tf-idf* スコアの高い単語として抽出された。これらの単語は、2003 年のDynamic Services and Analysis、2006 年と 2007 年のE-Communities、2004 年のWeb of Communitiesといったセッションに

[3] なお、*tf-idf*スコアの高い単語のみが、セッション間類似度の計算結果に反映されているわけではないため、セッション間の関係を詳細に分析する場合、*tf-idf*スコアの高くはない単語についても考慮する必要があろう。

おける*tf-idf* スコアの高い単語としても抽出されている。複数のセッションに共通する*tf-idf* スコアの高い単語は、これらのセッションで発表された研究が密接に関係していることを示唆している。特に、2006 年のE-Communitiesと 2008 年のSocial Networks: DECでは、community、social、networkが*tf-idf* スコアの高い単語として共通することから、これらのセッションで発表された研究が密接に関係していることが伺える。

表 2　2011 年の Social Network Analysis に至るカンファレンスセッションの時系列ネットワークから抽出した *tf-idf* スコアの高い単語とセッションの関係。*で示したセッションは、セッション間類似度のしきい値を高く設定した時系列ネットワークに含まれるセッションである。

Year	Session title	Terms with high *tf-idf* scores				
2011	Social Network Analysis*	social	network	list	reliable	score
2009	Interactions in Social Communities*	network	analysis	community	user	interact
2008	Social Networks: Discovery and Evolution of Communities*	community	network	social	connect	discover
2007	E-Communities*	system	web	community	forum	network
2007	Mining in Social Networks	network	identify	propagate	social	node
2006	Semi-Structured Semantic Data*	content	semantic	integrate	design	community
2006	E-Communities*	community	network	semantic	social	SNS
2006	Social Networks	social	semantic	network	discover	annotate
2005	Link-based Similarity	search	web	similar	index	graph
2005	Semantic Search	search	query	web	method	engine
2004	Reputation Networks*	evaluate	recommend	people	predict	system
2004	Web of Communities*	web	culture	content	blog	community
2004	Semantic Annotation and Integration	semantic	annotate	ontology	category	taxonomy
2004	Search Engineering 1	search	engine	user	web	page
2003	Dynamic Services and Analysis*	community	content	context	distribute	network
2003	Using the Semantic Web*	semantic	system	service	query	distribute
2003	Establishing the Semantic Web II	ontology	semantic	web	database	annotate
2003	Web Crawling and Measurement	page	crawl	perform	web	URL
2003	Information Retrieval 2	index	collect	scalable	query	service
2002	Semantic Web Services*	service	language	web	semantic	protocol
2002	UI & Applications*	web	distribute	system	user	toolkit
2002	Ontologies	ontology	process	semantic	domain	schema
2002	Search I	search	query	result	engine	meta
2002	Search II	answer	search	query	user	index
2002	Crawling	crawl	page	metric	parallel	href

4.2　マネタイゼーション研究の発展に寄与する分岐セッションノード

図 4 に、2011 年の Monetization I セッションから遡って生成したカンファレンスセッションの時系列ネットワークを示す。図 4(a)と(b)は、セッション間の接続を判定するセッション間類似度のしきい値を変化させて生成した時系列ネットワークである。マネタイゼーションという名称からは、計算機科学における一般的な研究トピックとは考えにくい。そのため、当該領域の研究者でなければ、WWW カンファレンスでマネタイゼーション研究が扱われるようになった背景を理解するのは容易ではないと思われる。

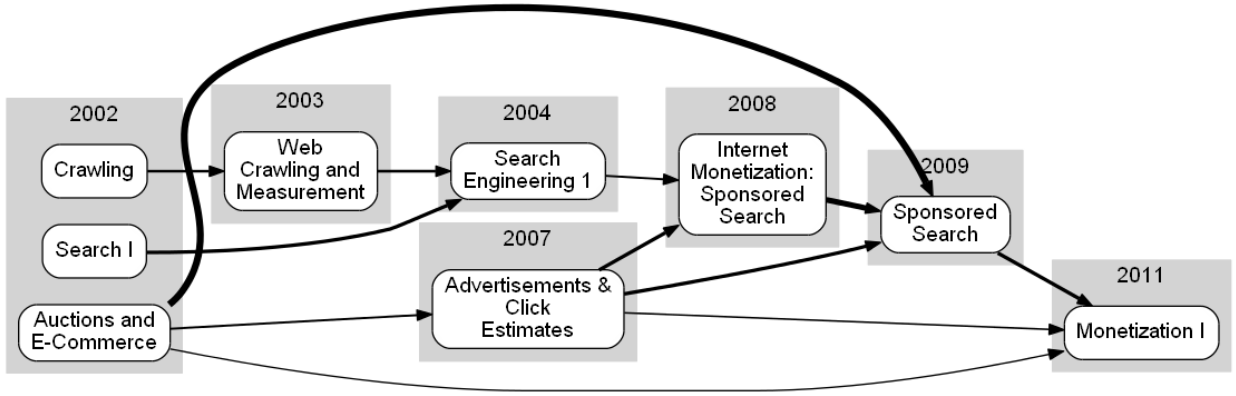
2002 年から 2011 年の全セッション名を確認すると、monetization という単語は 2008 年の Internet

Monetization: Sponsored Search で初めて登場している。このセッションノードを見ると、2002 年の Auctions and E-Commerce、2007 年の Advertisements & Click Estimates、2009 年の Sponsored Search から接続されていることがわかる。これらのセッション間の接続関係から、インターネットオークションや広告を含む電子商取引等に関する研究が、マネタイゼーション研究を形成したことが読み取れる。

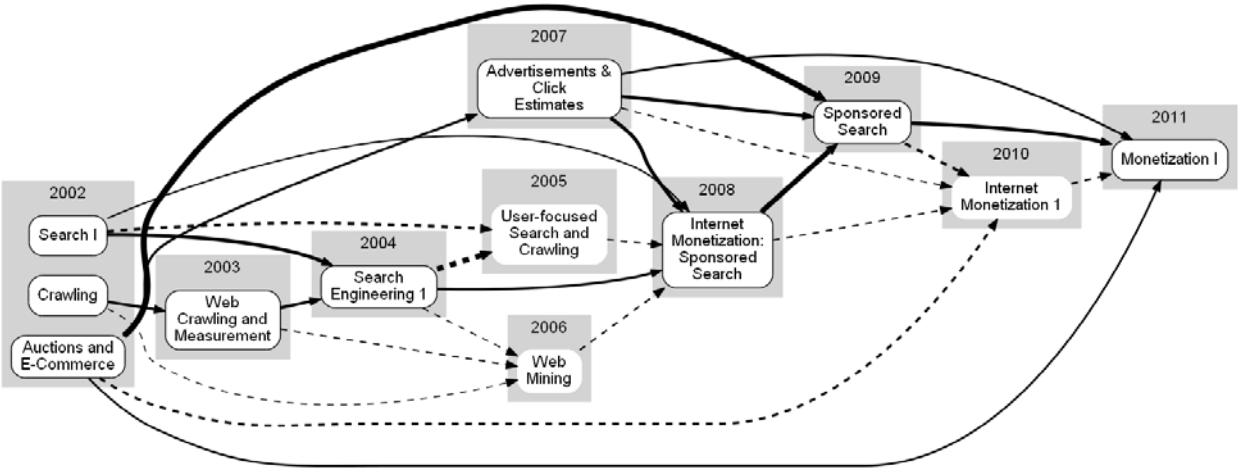
図 4(a)を見ると、2002 年の Auctions and E-Commerce と 2007 年の Advertisements & Click Estimates は、それぞれのセッションから派生した 3 つのセッションノードに接続されている。このようなセッション間の接続関係から、2002 年の Auctions and E-Commerce と 2007 年の Advertisements & Click Estimates は、その後のマネタイゼーション研究に大きな影響を与えた分岐セッションとみなすことができる。図 4(b)を見ると、2002 年の Search I や 2004 年の Search Engineering 1 といったセッションノードとの接続関係もあることから、これらのセッションも、その後のマネタイゼーション研究に影響を与えたセッションとみなすことができる。

また、図 4 (b)において、2002 年の Auctions and E-Commerce は、2007 年から 2011 年までのセッションノードに接続され、最短でも 5 年以上の間隔が空いていることがわかる。2002 年の Auctions and E-Commerce と 2009 年の Sponsored search の間には 7 年の間隔があるにもかかわらず、太い矢印で接続されていることからセッション間類似度は高い。このような結果は、2002 年の Auctions and E-Commerce に関連する研究が、数年の間隔を置いて再び盛んになったことを示唆している。

図 4(a)に示す 2009 年の Sponsored Search と 2011 年の Monetization I は、それぞれ 3 つセッションノードから派生していることから、収束セッションノードとみなすことができる。図 4(b)に示した 2008 年の Internet Monetization: Sponsored Search と 2010 年の Internet Monetization 1 も複数のセッションノードから派生していることから、過去の研究を統合した収束セッションノードと考えることができる。



(a) High similarity threshold



(b) Low similarity threshold

図 4　2011 年の Monetization I セッションに至る時系列ネットワーク。セッション間類似度の大きを矢印の太さに反映させている。実線で囲まれたセッションノードと実線で示された矢印は、上段の時系列ネットワークに一致する。

表 3 に、図 4 の各セッションで発表された論文のアブストラクトから抽出した *tf-idf* スコアの高い単語を示す。ここで、*で示したセッションは、図 4 (a)に時系列ネットワークのセッションノードに対応する。2002 年の Auction and E-Commerce セッションにおいて *tf-idf* スコアの高い代表的な単語は auction と advertise であり、これらは、2007 年の Advertisements & Click Estimates と 2009 年の Sponsored Search でも *tf-idf* スコアの高い単語として抽出されている。2011 年の Monetization I でも、auction が *tf-idf* スコアの高い共通の単語として抽出されていることから、マネタイゼーション研究の一部は Auction and E-Commerce セッションから派生したことが伺える。2007 年の Advertisements & Click Estimates、2008 年の Internet Monetization: Sponsored Search、2009 年の Sponsored Search、2011 年の Monetization I のセッションでは、advertise、click、auction、search と

いった単語の *tf-idf* スコアが高いことから、これらに関連する研究がマネタイゼーション研究の発展過程において重要な役割を果たしたことが推測される。

表 3　2011 年の Monetization I に至るカンファレンスセッションの時系列ネットワークから抽出した *tf-idf* スコアの高い単語とセッションの関係。*で示したセッションは、セッション間類似度のしきい値を高く設定した時系列ネットワークに含まれるセッションである。

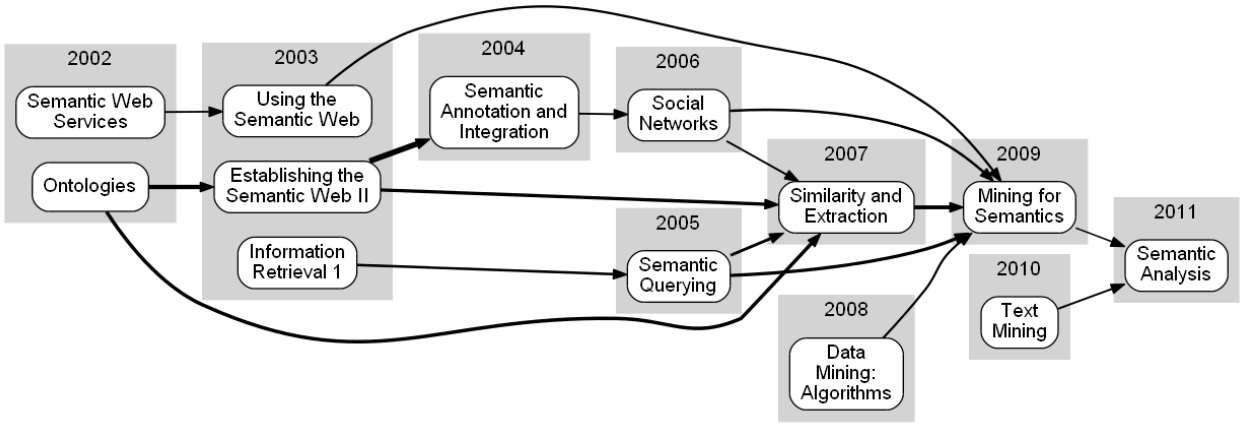
Year	Session title	Terms with high *tf-idf* scores				
2011	Monetization I*	mechanism	optimize	bidder	auction	equilibrium
2010	Internet Monetization 1	page	content	bid	sponsor	auction
2009	Sponsored Search*	auction	bid	advertise	search	price
2008	Internet Monetization: Sponsored Search*	advertise	click	sponsor	search	engine
2007	Advertisements & Click Estimates*	advertise	click	model	user	auction
2006	Web Mining	page	extract	analysis	web	search
2005	User-focused Search and Crawling	search	engine	query	user	web
2004	Search Engineering 1*	search	engine	user	web	page
2003	Web Crawling and Measurement*	page	crawl	perform	web	url
2002	Search I*	search	query	result	engine	meta
2002	Crawling*	crawl	page	metric	parallel	href
2002	Auctions and E-Commerce*	auction	agent	market	internet	advertise

4.3　セマンティックアナリシスの発展過程

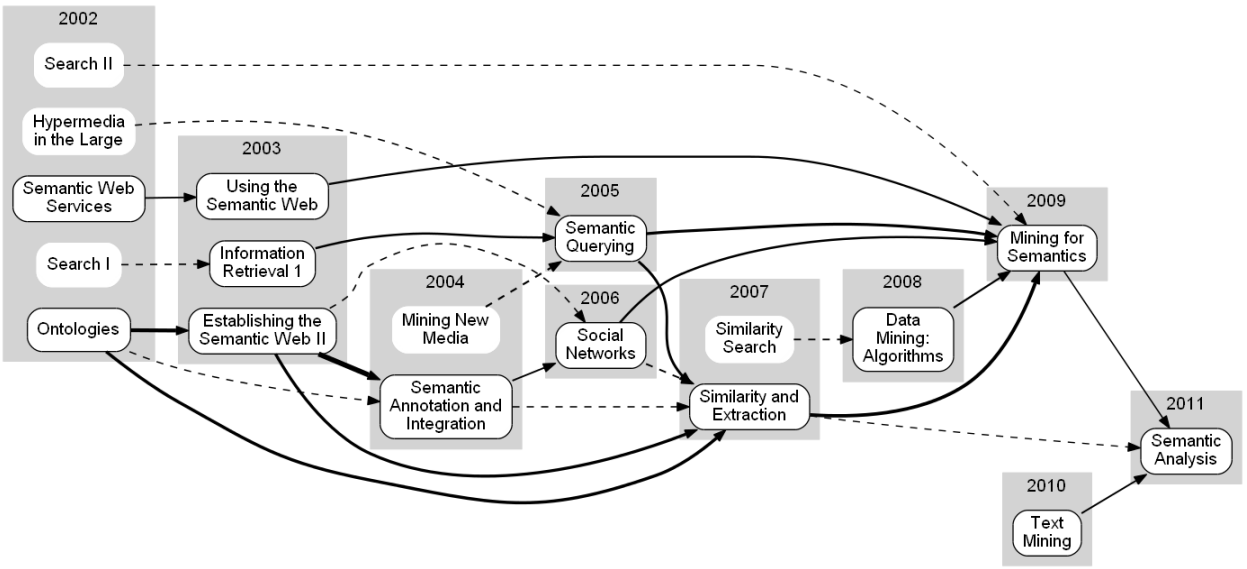
図 5 に、2011 年の Semantic Analysis セッションから遡って生成したカンファレンスセッションの時系列ネットワークを示す。図 5 (a)、(b)はセッション間の接続を判定するセッション間類似度のしきい値を変化させて生成した結果である。図 5(b)の点線で示した矢印は、図 5(a)で示されなかった潜在的なセッション間の関係を示している。セマンティックアナリシスは、前述したソーシャルネットワークやマネタイゼーション研究と比較して、この研究領域では比較的歴史の長い研究トピックと言える。図 5 (a)を見ると、2002 年に Semantic Web や Ontologies を含むセッションが登場している。2007 年の Similarity and Extraction や 2009 年の Mining for Semantics は、過去に実施された複数のセッションに接続されていることから、収束セッションノードとみなすことができる。そのため、2007 年の Similarity and Extraction において、過去のセマンティックウェブ技術が統合され、さらに、2009 年の Mining for Semantics において、セマンティックウェブ技術とデータマイニング技術が統合されたことが推察される。図 5 に示す時系列ネットワークは、セマンティックアナリシスに関する研究はデータマイニング技術を取り込みながら発展したことを示唆している。

表 4 に、図 5 に示した各セッションで発表された論文のアブストラクトから抽出した *tf-idf* スコアの高い単語を示す。ここで、*で示したセッションは、図 5(a)に示す時系列ネットワークのセッションノードに対応する。表 4 では、semantic、web、services、ontology などが *tf-idf* スコアの高い単語として抽出されている。これらの単語は関連するセッションで共有されていることから、

オントロジーを基盤とするセマンティックウェブサービスに関する研究トピックが発展したことが伺える。収束セッションノードと考えられる 2007 年の Similarity and Extraction では、*tf-idf* スコアの高い単語として semantic と ontology を含むことから、オントロジーを基盤とするセマンティックウェブに関する過去の統合されたことが示唆される。さらに、2009 年の Mining for Semantics は *tf-idf* スコアの高い単語として semantic に加えて similarity を含むことから、similarity に関する研究トピックがオントロジーを基盤とするセマンティックウェブ技術とデータマイニング技術を統合されたことが推測される。



(a) High similarity threshold



(b) Low similarity threshold

図 5　2011 年の Semantic Analysis セッションに至る時系列ネットワーク。セッション間類似度の大きを矢印の太さに反映させている。実線で囲まれたセッションノードと実線で示された矢印は、上段の時系列ネットワークに一致する。

表 4　2011 年の Semantic Analysis に至るカンファレンスセッションの時系列ネットワークから抽出した *tf-idf* スコアの高い単語とセッションの関係。*で示したセッションは、セッション間類似度のしきい値を高く設定した時系列ネットワークに含まれるセッションである。

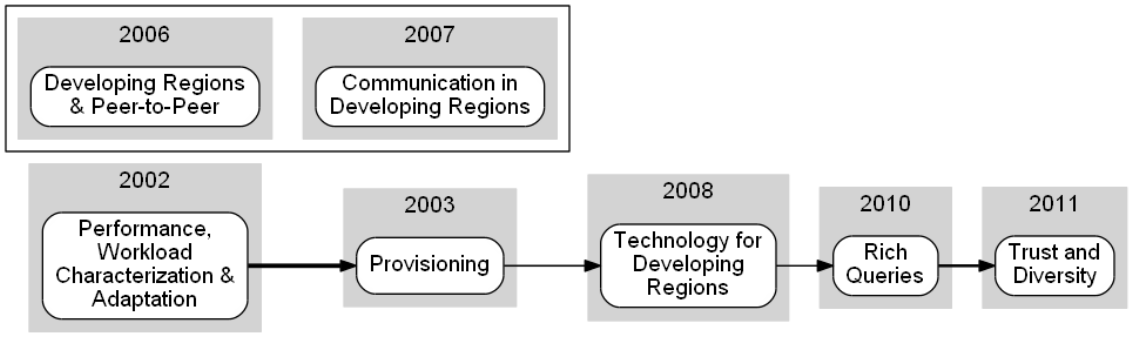
Year	Session title	Terms with high *tf-idf* scores				
2011	Semantic Analysis*	language	document	analysis	lexicon	semantic
2010	Text Mining*	algorithm	robust	text	source	web
2009	Mining for Semantics*	semantic	relation	detect	web	similarity
2008	Data Mining: Algorithms*	method	algorithm	web	similarity	classify
2007	Similarity and Extraction*	semantic	similarity	extract	measure	ontology
2007	Similarity Search	similarity	score	assess	web	search
2006	Social Networks*	social	semantic	network	discover	annotate
2005	Semantic Querying*	rank	search	measure	semantic	topic
2004	Semantic Annotation and Integration*	semantic	annotate	ontology	category	web
2004	Mining New Media	news	individual	algorithm	person	dynamic
2003	Using the Semantic Web*	semantic	service	query	distribute	similarity
2003	Information Retrieval 1*	query	web	search	topic	page
2003	Establishing the Semantic Web II*	ontology	semantic	web	annotate	layer
2002	Semantic Web Services*	service	language	web	protocol	semantic
2002	Ontologies*	ontology	process	semantic	domain	schema
2002	Hypermedia in the Large	estimate	classify	algorithm	topic	hypermedia
2002	Search I	search	query	result	engine	meta
2002	Search II	answer	search	query	user	index

4.4　開発途上地域のための技術

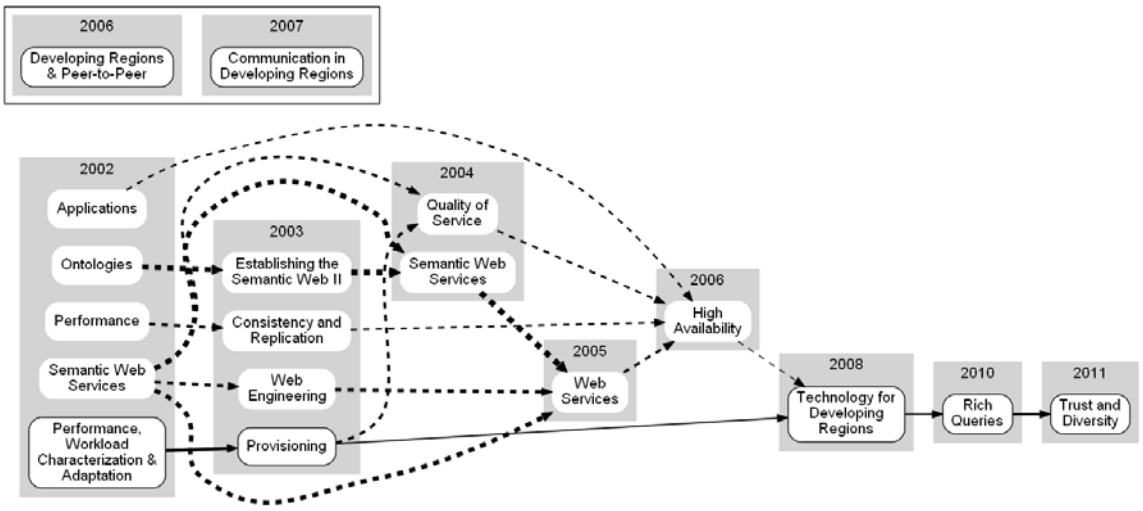
Technology for Developing Regions というセッション名には、伝統的な研究よりも、社会への貢献を意識した新たな研究領域を開拓しようという研究者コミュニティの意思が反映されているように思われる。図 6 に、2008 年の Technology for Development Regions を起点として生成したカンファレンスセッションの時系列ネットワークを示す。図 6(a)、(b)はセッション間の接続を判定するセッション間類似度のしきい値を変化させて生成した結果である。図 6(a)、(b)の左上に示した 2006 年の Developing Regions & Peer-to-Peer と 2007 年の Communication in Developing Regions は開発途上地域に関連する技術として、2008 年の Technology for Development Regions に接続されることが期待された。しかし、図 6(a)と(b)のいずれも 2008 年の Technology for Development Regions から生成した時系列ネットワークには接続されずに独立した構造となっている。図 6 (a)は、2008 年の Technology for Developing Regions がシステムパフォーマンスや信頼性と多様性といった研究トピックから派生したことを示唆している。図 6(b)の点線で示した矢印は、2000 年代初期のウェブサービスが 2006 年の High Availability、2008 年の Technology for Developing Regions に発展したことを示唆している。

表 5 に、図 6 の各セッションで発表された論文のアブストラクトから抽出した *tf-idf* スコアの高い単語を示す。2000 年代初めのセッションに共通する *tf-idf* スコアの高い単語として web と service が抽出され、2000 年代中期のセッションから service と quality が抽出されている。これらのセッションに共通する単語から、2000 年代の初期から中期にかけてウェブサービスの品質に関する研

究が進展したと推測される。2006 年から 2008 年にかけて登場した Developing Regions という名称を含むセッションは互いに接続されることが期待されたが、これらのセッションは時系列ネットワークの中で接続されていない。表 5 に示した *tf-idf* スコアの高い単語を確認すると、これらのセッション間で共通する単語が存在しないため、時系列ネットワークでセッションノードが接続されないことが説明できる。2008 年の Technology for Developing Regions の場合、microbusiness や non-profit といった特徴的な単語の *tf-idf* スコアが高い。2007 年の Communication in Developing Regions でも illiteracy という特徴的な単語の *tf-idf* スコアが高い。これらの単語は ICT が開発途上地域で貢献する方向性を示唆するものの、Developing Regions を名称に含むセッション間の関連性を示すことにはなっていない。開発途上地域における ICT 応用に関する研究は発展しつつあるが、研究トピックには多様性があり、本調査研究で用いいた簡単なテキスト分析ではこのようなセッション間の共通性を見い出せなかったと解釈できる。



(a) High threshold



(b) Low threshold

図 6　2008 年の Technology for Developing Regions セッションを含む時系列ネットワーク。セッション間類似度の大きを矢印の太さに反映させている。実線で囲まれたセッションノードと実線で示された矢印は、上段の時系列ネットワークに一致する。

表 5　2008 年の Technology for Developing Regions セッションを含むカンファレンスセッションの時系列ネットワークから抽出した *tf-idf* スコアの高い単語とセッションの関係。*で示したセッションは、セッション間類似度のしきい値を高く設定した時系列ネットワークに含まれるセッションである。

Year	Session title	Terms with high *tf-idf*				
2011	Trust and Diversity *	source	trustworthy	relevance	query	method
2010	Rich Queries *	web	user	query	source	person
2008	Technology for Developing Regions *	translate	tool	microbusiness	non-profit	action
2007	Communication in Developing Regions **	communicate	social	people	region	illiteracy
2006	High Availability	infrastructure	distribute	system	service	bottleneck
2006	Developing Regions & Peer-to-Peer **	network	distribute	system	feed	contribute
2005	Web Services	service	web	enterprise	protocol	interface
2004	Quality of Service	service	quality	framework	context	QoS
2004	Semantic Web Services	service	ontology	semantic	web	framework
2003	Establishing the Semantic Web II	ontology	semantic	web	databas	annotate
2003	Consistency and Replication	composit	framework	individual	web	service
2003	Provisioning *	request	service	control	quality	QoS
2002	Semantic Web Services	service	languag	web	protocol	semantic
2002	Ontologies	ontology	process	semantic	domain	schema
2002	Applications	standard	resource	content	system	catalog
2002	Performance	implement	web	proxy	server	overhead
2002	Performance, Workload Characterization & Adaptation *	request	web	client	site	payload

5. 時系列ネットワークを用いた分析手法に関する検討

5.1 研究者コミュニティの将来展望を反映するカンファレンスセッション

学術文献の分析結果をデルファイ予測調査[55]やシナリオプランニング[61]と組み合わせて将来予測に応用する研究も進展している。例えば、キーワード検索で抽出された文献数や単語やフレーズなどの出現頻度から求めた指標の時系列変化を成長曲線で近似し、さらに成長曲線を外挿することにより新しい研究の発展を予測するといった方法が提案されている[55]。成長曲線を外挿するという考え方は合理的ではあるものの、過去の学術文献分析に基づいた分析手法であるため、予測手法として考えると将来展望は反映されにくいと言わざるを得ない。

カンファレンスのセッション名には、研究者コミュニティにおける将来展望が反映されていると考えられるため、提案手法は将来を指向した分析手法であるとみなすことができる。一般に、カンファレンスセッションの構成はプログラムコミッティーで決定され、セッション名には、先駆的な論文によって提示された新たな研究トピックや将来への展望が反映されていると考えられる。また、社会科学的な観点で考えると、プログラムコミッティーのメンバーが今後重要になると思われる研究トピックを選択するというゲートキーパーとのしての役割を果していると言える[62]。

過去の研究トピックを統合する収束セッションノードの場合、かつては個別に検討されていた研究を融合した新しい概念がセッション名によって提示されることがある。また、その後の研究に大きな影響を与えたと先駆的な論文を含む分岐セッションノードの場合にも、新しく創出された概念がセッション名によって提示されることもある。例えば、本研究で対象とした Monetization というセッションや E-Communities から Social Networks へのセッション名の変化は、研究者コミュニティが新たな研究領域を開拓しようとする動きと捉えることができる。このように考えるとセッション名そのものが、研究者コミュニティが新たに創出した概念と解釈することもできる。

5.2 研究を推進する要因とその後の研究に影響を与えるカンファレンスセッション

研究を推進する要因は、研究者に関する内部要因と外部要因に分けることができる。研究者のもつ純粋な科学的興味は、研究を推進する代表的な内部要因であり、伝統的な基礎科学を推進する重要な要因となっている。研究者の内部要因が原動力となって推進される研究は Mode 1 と呼ばれている[63]。ICT 産業と密接に関係するウェブ関連技術の研究は、実際のアプリケーションやサービスを意識した応用研究の色彩が強く、このような研究は Mode 2、あるいは、context-intensive science と呼ばれている[63]。この場合、研究を推進する要因として、研究者の内部要因に加えて、社会的な環境を含めた外部要因が考えられている。例えば、研究成果の商業化による経済効果を期待した研究ファンディングは研究を推進する代表的な外部要因である。その他に、経済原理だけでは説明が困難な研究開発による社会福祉への貢献なども研究を推進する要因の一つとして考えられる。2008 年の WWW カンファレンスには、Technology for Developing

Regions というセッションが実施されたことから、発展途上地域における社会福祉の実現に寄与するような研究を推進しようとするコミッティーメンバーの意思があった考えられる。図 6 に示した時系列ネットワークでは、論文アブストラクトを用いたテキスト分析手法における情報量の限界から、このような研究トピックの背景を含めた発展過程を適切に示すことはできなかった。しかし、このようなセッションの存在は、研究者コミュニティにおける先駆的な研究に対する将来展望がカンファレンスセッションに反映されていることを示唆している。

5.3 応用に関する検討

5.3.1 他の研究領域への応用

計算機科学の場合、論文は採択率の低い著名なカンファレンスで発表されることも多く、ジャーナルペーパーではないプロシーディングペーパーも十分に高く評価される傾向がある。そのため、WWW カンファレンスは当該研究領域における研究トレンドの変遷の分析に適していた。以下に、計算機科学を除く他の代表的な研究領域について、提案手法の適用可能性を検討するために公開されているカンファレンスプログラムを調査した事例を示す。医学系の研究領域の場合、American Association for Cancer Research が開催している年次大会では、ガン研究における最新の学術的な成果や発展しつつある研究テーマがセッションに取り上げられている。物理学の American Physical Society や化学の American Chemical Society といった学会でも年次大会のセッションには何らかの名称が付与されている。すべての研究領域で同様のことが確認されたわけではないが、少なくとも計算機科学以外の研究領域でも、カンファレンスセッションに注目した手法で最先端の研究トレンド分析ができる可能性が示唆された。

5.3.2 類似度のしきい値

提案手法では、カンファレンスセッション間の接続関係を生成するために、セッション間の類似度にしきい値を設定した。しきい値設定の影響を検討するために、複数のしきい値を用いてカンファレンスセッションの時系列ネットワークを生成した結果を示した。図 3 から図 6 に示したように、低いしきい値を設定した場合、カンファレンスセッション間の潜在的な接続関係を示すことができた。しかし、セッション間の接続関係の評価ではテキスト分析におけるノイズの影響を考慮する必要もあり、情報の粒度と類似度のしきい値設定の関係については、今後も検討すべき課題である。

5.3.3 クラスタとしてのカンファレンスセッション

既存の計量書誌学的な手法やテキスト分析手法と比べて、カンファレンスセッションの時系列ネットワーク分析では、学術文献のクラスタリング結果の検証や、抽出されたクラスタの解釈などの問題を回避できる利点がある。その一方で、提案手法にはセッションに含まれる論文数の制約に起因する問題も存在する。例えば、カンファレンスセッションは数理的に最適なクラスタで

あるとは言えず、同一セッションに含まれる論文が必ずしも類似しているとは限らない。静的なデータを対象とした数理的なクラスタリング問題に限定すれば、このような性質は不適切であると言える。しかし、本研究では、革新的技術が発展する過程の動的な変化に注目しているため、ある瞬間において生成された学術文献のクラスタが数理的に最適でないとしても、深刻な問題として扱う必要はないと思われる。

提案手法の場合、セッション名については次の問題が残る。特定の研究トピックに関する多くの論文がカンファレンスで採択された場合、セッションが複数に分割されることがある。例えば、2012 年の WWW カンファレンスでは、Monetization I と Monetization II という 2 つのセッションに分かれている。このようなセッションの場合、以下に示す 2 通りの解釈ができる。

- Monetization IとMonetization IIは同一の研究トピックを扱ったセッションであるが、1つのセッションで発表可能な論文数の制約から2つのセッションに分割された。
- Monetization IとMonetization IIは異なった研究トピックを扱っているが、研究者間でも概念や用語が共有されておらず適切なセッション名を付与できなかった。

本調査研究では、このような差異を区別することができないため、さらなる調査研究が必要であろう。

6. おわりに

公的、あるいは、民間の科学技術投資の方向性を議論する際に、当該領域の専門家だけでなく非専門家を含めたステークホルダーの間で、先端研究の動向に関する共通認識をもつことの重要性が指摘されている。しかし、急速に発展している研究領域の動向については、非専門家が研究の動向を正確に把握することは極めて困難である。本研究では、科学技術政策のベンチマーキングと科学技術ロードマッピングへの応用を目指し、萌芽的研究の発展過程を分析するための手法を検討した。

まず、基本データの構造、学術文献間の関係、分析結果の安定性、情報探索範囲、基本データ先生に要する時間の観点から、代表的な計量書誌学的な手法である共引用分析とテキスト分析を比較した結果について論じた。このような比較から、学術文献を用いて専門家にも認識されていないような知識の抽出を目的とした場合、あいまいなデータを扱うことのできるテキスト分析が適することを示した。次に、カンファレンスのセッションと発表されたプロシーディングペーパーのアブストラクトを用いて、研究領域の発展過程を示す時系列ネットワークの生成手法を示した。

本調査研究では、2002 年から 2011 年に開催された WWW カンファレンスのセッションをノードとする時系列ネットワークを生成し､研究トピックの変遷から最先端研究の動向を可視化した。4 つの特徴的なセッション Social Network Analysis、Monetization I、Semantic Analysis (WWW 2011) と Technology for Developing Regions (WWW 2008)に注目し、これら研究トピックの発展過程について詳細な分析を行った。カンファレンスセッションの時系列ネットワーク分析によって抽出された特徴的なセッションから、異なった研究トピック間の相互作用が研究の発展過程において重要な役割を果たすことが示唆された。さらに、以下のセッションによって研究の発展過程が特徴づけられた。

- 過去のセッションとの接続が多い収束セッションノードは、過去の研究トピックを総括したと考えられる。
- その後のセッションとの接続が多い分岐セッションノードは、他の研究に大きな影響を与えたセッションと考えられる。

当初、ソーシャルネットワークやマネタイゼーションといった研究が登場した背景は不明であったが、カンファレンスセッションの時系列ネットワークを生成することによって、過去に行われた研究の関係など、これらの研究が発展する過程を可視化することができた。結果として、時系列ネットワークの分析によって、以下の知見が得られた。

- Social Networksという単語を含むセッションの登場と、E-communitiesからSocial Networksへの名称の変化は、ソーシャルネットワーク研究の急速な発展に対応している。
- マネタイゼーション研究は、Sponsored Search、Auction and E-Commerce、Advertisement & Click Estimatesなどの研究から派生し、相互に関係しながら発展してきた。

提案手法の特徴は、研究者コミュニティにおける新たな研究領域を開拓しようとする意思や将来展望が反映されたと考えられるカンファレンスセッションに注目し、萌芽的研究の発展過程を可視化した点にある。個々の論文よりも抽象度の高いセッション名を扱うことで、最新の研究動向を容易に把握できるようになった。カンファレンスセッションの時系列ネットワーク分析により、過去の研究を総括するような収束セッションノードと、その後の研究に影響を与えたと思われる分岐セッションノードの存在が示された。

テキスト分析における安定性などの課題が残されているものの、提案手法は、萌芽的研究の発展過程を分析する手法として有用であると考えられる。ただし、セッションの推移を可視化した結果の妥当性については、当該研究領域の研究者にインタビューするなど定性的な評価を行う必要があろう。調査研究の実施段階では、分析に用いたすべてのプロシーディングペーパーがデータベースに収録されていなかったため、計量書誌学的な手法とは結果を比較できなかったが、今後は共引用分析等の手法で得られた結果との比較も検討課題として挙げられる。萌芽的研究として抽出されたセッションから、そのセッションで発表された論文の著者や、セッションチェア等の研究者情報も取得できるため、これらの研究者情報を用いた共著者分析についても興味が持たれる。また、医学、物理学、化学領域の代表的な学会の年次大会についても、名称の付与されたセッションが開催されていることから提案手法が他の研究領域にも適用できる可能性が示唆された。

謝辞

本調査研究のうち分析用ソフトウェア開発、及び、データ分析に関する部分は、平成 22～24 年度科学研究費補助金（基盤研究（C））研究課題名「萌芽的研究分野の探索手法に関する研究」（課題番号 22500875　研究代表者：古川　貴雄）、平成 26～28 年度科学研究費補助金（基盤研究（C））研究課題名「オープンイノベーションからみた萌芽的研究領域における発展要因の定量分析」（課題番号　26330375　研究代表者：古川　貴雄）の助成の下で行った。

調査にあたり以下の方々に助言をいただいたことを深謝する。

奥和田 久美　　独立行政法人 科学技術振興機構 社会技術技術研究開発センター
シニアフェロー
小柴 等　　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 研究員
七丈 直弘　　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 上席研究官
小笠原 敦　　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター長

参考文献

[1] 科学技術政策研究所, 科学技術の中長期発展に係る俯瞰的予測調査 急速に発展しつつある研究領域調査 - 論文データベース分析から見る研究領域の動向 -, NISTEP REPORT No. 95, 科学技術政策研究所, 2005.5.

[2] 科学技術政策研究所 科学技術動向研究センター, サイエンスマップ 2004 - 論文データベース分析(1999 年から 2004 年)による注目される研究領域の動向調査, NISTEP REPORT No. 100, 科学技術政策研究所, 2007.3

[3] 阪 彩香, 伊神 正貫, 桑原 輝隆, サイエンスマップ 2006 ―論文データベース分析(2001 年から 2006 年)による注目される研究領域の動向調査―, NISTEP REPORT No. 110, 科学技術政策研究所, 2008.6.

[4] 阪 彩香, 伊神 正貫, 桑原 輝隆, サイエンスマップ 2008 ―論文データベース分析 (2003 年から 2008 年) による注目される研究領域の動向調査―, NISTEP REPORT No. 139, 科学技術政策研究所, 2010.5.

[5] 阪 彩香, 伊神 正貫, サイエンスマップ 2010&2012 ―論文データベース分析 (2005 年から 2010 年および 2007 年から 2012 年) による注目される研究領域の動向調査―, NISTEP REPORT No. 159, 科学技術・学術政策研究所, 2014.7.

[6] D. R. Swanson, Two medical literatures that are logically but not bibliographically connected, J Am. Soc. Inform. Sci., 38(4) (1987) 228–233.

[7] T. Furukawa, K. Mori, K. Arino, K. Hayashi, N. Shirakawa, Identifying the evolutionary process of emerging technologies: A chronological network analysis of World Wide Web conference sessions, Technol. Forecast. & Soc. Chang., (in press)

[8] G. S. Day, P. J. H. Shoemaker, A different game, in G. S. Day, and P. J. H. Shoemaker (Eds.), Wharton on Managing Emerging Technologies, John Wiley and Sons Inc., Hoboken, New Jersey (2004). 1–23.

[9] R. N. Kostoff, R. R. Schaller, Science and technology roadmaps. IEEE Transactions on Engineering Management, 48(2) (2001) 132–143.

[10] R. N. Kostoff, R. Boylan, G. R. Simons, Disruptive technology roadmaps. Technol. Forecast. Soc. Change, 71(1) (2004) 141–159.

[11] S. T. Walsh, Roadmapping a disruptive technology: A case study – The emerging microsystems and top-down nanosystems industry. Technol. Forecast. Soc. Change, 71 (1) (2004) 161–185.

[12] L. Zhang, W. Glänzel, Proceeding papers in journals versus the 'regular' journal publications. Journal of Informetrics, 6 (1) (2012) 88–96.

[13] J. Y. Halpern, D. C. Parkes, Journals for certification, conferences for rapid dissemination. Communications of the ACM, 54 (8) (2011) 36–38.

[14] E. Mansfield, Academic research and industrial innovation. Res. Policy 20 (1) (1991) 1–12.

[15] L. Fleming, O. Sorensen, Science as a map in technological search. Strategic Research Journal 25 (8-9) (2004) 909–928.

[16] S. Lee, B. Yoon, C. Lee, H. Park, Business planning based on technological capabilities: Patent and technology-driven roadmapping. Technol. Forecast. Soc. Change, 76 (6) (2009) 769–786.

[17] H. Small, Co-citation in the scientific literature: A new measure of the relationship between two documents. J Am. Soc. Inform. Sci., 24(4) (1973) 265–269.

[18] H. Small, Visualizing Science by Citation Mapping. J. Am. Soc. Inform. Sci. 50 (9) (1999) 799–813.

[19] M. Gmür, Co-citation analysis and the search for invisible colleges: A methodological evaluation, Scientometrics 57 (1), (2003) 27–57.

[20] W. Glänzel, B. Thijs, Using ‘core documents’ for the representation of clusters and topics. Scientometrics, 88 (1) (2011) 297–309.

[21] W. Glänzel, B. Thijs, Using ‘core documents’ for detecting and labelling new emerging topics. Scientometrics, 91 (2) (2012) 399–416.

[22] Y. Kajikawa, J. Yoshikawa, Y. Takeda, & K. Matsushima, Tracking emerging technologies in energy research: Toward a roadmap for sustainable energy. Technological Forecasting and Social Change, 75 (6), (2008) 771–782.

[23] Y. Kajikawa, & Y. Takeda, Structure of research on biomass and bio-fuels: A citation-based approach. Technol. Forecast. and Soc. Change, 75 (9), (2008) 1349–1359.

[24] Y. Kajikawa, Y, T. Takeda, Citation network analysis of organic LEDs. Technol. Forecast. Soc. Change, 76 (8) (2009) 1115–1123.

[25] N. Shibata, Y. Kajikawa, Y. Takeda, K. Matsushima, Detecting emerging research fronts based on topological measures in citation networks of scientific publications. Technovation 28 (11) (2008) 758–775.

[26] N. Shibata, Y. Kajikawa, Y. Takeda, I. Sakata, K. Matsushima, Detecting emerging research fronts in regenerative medicine by the citation network analysis of scientific publications. Technol. Forecast. Soc. Change 78 (2) (2011) 274–282.

[27] S. H. Chen, M. H. Huang, D. Z. Chen, Identifying and visualizing technology evolution: A case study of smart grid technology. Technol. Forecast. Soc. Change 78 (6) (2012) 1099–1110.

[28] S. H. Chen, M. H. Huang, D. Z. Chen, S. G. Lin, Detecting the temporal gaps of technology fronts: A case of smart grid. Technol. Forecast. Soc. Change 79 (9) (2012) 1705–1719.

[29] P. D. Bruza & M. Weeber (Eds.), Literature-based discovery, Springer-Verlag, Heidelberg (2008).

[30] D. R. Swanson & N. R. Smalheiser, An interactive system for finding complementary literatures: a stimulus to scientific discovery. Artificial Intelligence, 91(2) (1997) 183–203.

[31] N. R. Smalheiser, D. R. Swanson, Using ARROWSMITH: a computer-assisted approach to

formulating and assessing scientific hypotheses. Computer Methods and Programs in Biomedicine, 57(3) (1998) 149–153.

[32] D. R. Swanson, N. R. Smalheiser, A. Bookstein, Information discovery from complementary literatures: categorizing viruses as potential weapons. J Am. Soc. Inform. Sci. Tech., 52 (10) (2001). 797–812.

[33] N. R. Smalheiser, Predicting emerging technologies with the aid of text-based data mining: the micro approach, Technovation 21 (10) (2001) 689–693.

[34] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, R. A. Harshman, Indexing by latent semantic analysis. J Am. Soc. Inform. Sci. Tech., 41 (6) (1990) 391–407.

[35] M. D. Gordon & S. Dumais, Using latent semantic indexing for literature based discovery. J Am. Soc. Inform. Sci. Tech., 49 (8) (1998) 674–685.

[36] D. M. Blei A. Y. Ng, M. I. Jordan, Latent Dirichlet Allocation. The Journal of Machine Learning Research, 3 (2003) 993–1022.

[37] T. L. Grifiths, M. Steyvers, Finding scientific topics. Proceeding of the National Academy of Sciences 101, (suppl. 1), 5228–5235 (2004).

[38] D. M. Blei, J. D. Lafferty, Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning. ACM, (2006) 113–120.

[39] D. M. Blei, Probabilistic topic models. Communications of the ACM 55 (4) (2012) 77–84.

[40] A. Daud, J. Li, L. Zhou, F. Muhammad, Conference Mining via Generalized Topic Modeling, in Machine Learning and Knowledge Discovery in Database, LNCS 5781 (2009) 244–259.

[41] R. N. Kostoff, Text mining using database tomography and bibliometrics: A review. Technol. Forecast. Soc. Change, 68 (3), (2001) 223–253.

[42] R. N. Kostoff, J. A. del Rio, J. A. Humenik, E. O. Garcia & A. M. Ramirez, Citation mining: Integrating text mining and bibliometrics for research user profiling. J. Am. Soc. Inform. Sci. Tec., 52 (13) (2001) 1148–1156.

[43] R. N. Kostoff, Literature-related discovery (LRD): introduction and background. Technol. Forecast. Soc. Change, 75 (2), (2008) 165–185.

[44] R. N. Kostoff, M. B. Briggs, J. L. Solka, R. L. Rushenberg, Literature-related discovery (LRD): Methodology. Technol. Forecast. Soc. Change, 75 (2), (2008), 186–202.

[45] R. N. Kostoff, Systematic acceleration of radical discovery and innovation in science and technology. Technol. Forecast. Soc. Change, 73 (8) (2006) 923–936.

[46] R. N. Kostoff, J. A. Block, J. L. Solka, M. B. Briggs, R. L. Rushenberg, J. A. Stump, D. Johnson, T. J. Lyons & J. R. Wyatt, Literature-related discovery (LRD): Lessons learned, and future research directions. Technol. Forecast. Soc. Change, 75 (2) (2008) 276–299.

[47] R. N. Kostoff, Literature-related discovery and innovation—update, Technol. Forecast. Soc. Change,

79 (4) (2012), 789–800.

[48] M. H. Coletti & Bleich, H. L, Medical subject headings used to search the biomedical literature. Journal of the American Medical Informatics Association, 8 (4) (2001) 317–323.

[49] D. R. Swanson, N. R. Smalheiser, & V. I. Torvik, Ranking indirect connections in literature‐based discovery: The role of medical subject headings. J Am. Soc. Inform. Sci. Tech., 57 (11) (2006) 1427–1439.

[50] N. A. Behkami, T. U. Daim, Research Forecasting for Health Information Technology (HIT), using technology intelligence. Technological Forecasting and Social Change, 79 (3) (2012) 498–508.

[51] R. N. Kostoff, & C. G. Lau, Combined biological and health effects of electromagnetic fields and other agents in the published literature, Technol. Forecast. Soc. Change, 80 (7) (2013), 1331–1343.

[52] Y. H. Tseng, C. J. Lin, Y. I. Lin, Text mining techniques for patent analysis. Information Processing and Management 43 (2007) 1216–1247.

[53] A. L. Porter, R. J. Watts, T. R. Anderson, Mining PICMET: 1997–2003 papers help you track management of technology developments. Proceedings of Portland International Conference on Management of Engineering and Technology, 2003, pp. 188–193.

[54] J. Kwakkel, S. W. Cunningham, T. R. Anderson, Remining PICMET: 1987–2008, Proceedings of Portland International Conference on Management of Engineering and Technology, 2009, pp. 22–36.

[55] M. Bengisu, R. Nekhili, Forecasting emerging technologies with the aid of science and technology databases. Technol. Forecast. Soc. Change 73 (7) (2006) 835–844.

[56] E. Mörchen, M. Dejori, D. Fradkin, J. Etienne, B. Wachmann, M. Bundshus, Anticipating annotations and emerging trends in biomedical literature. Proceedings of 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2008, pp. 854–962.

[57] Y. G. Kim, J. H. Suh, S. C. Park, Visualization of patent analysis for emerging technology. Expert Systems with Applications 34 (3) (2008) 1804–1812.

[58] Y. Jo, C. Lagoze, C. L. Giles, Detecting research topics via the correlation between graphs and texts. Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, 2007, pp. 390–379.

[59] C. Chen, CiteSpaceII: Detecting and visualizing emerging trends and transient patterns in scientific literature. J. Am. Soc. Inform. Sci. Tec. 57 (3) (2006) 357–377.

[60] G. Salton, C. Buckley. Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24 (5), (1988) 513–523.

[61] T. U. Daim, G. Rueda, H. Martin, P. Gerdsri, Forecasting emerging technologies: Use of bibliometrics and patent analysis. Technol. Forecast. Soc. Change 73 (8) (2006), 981–1012.

[62] P. J. Shoemaker, T. Vos. Gatekeeping theory. Routledge (2009).

[63] M. Gibbons. Mode 2 society and the emergence of context-sensitive science. Science and public policy, 27(3) (2000) 159–163.

調査担当者

調査設計・実施・分析・取りまとめ

古川 貴雄　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 客員研究官
共立女子大学 家政学部 准教授
森 薫　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 技術参与
慶應義塾大学 大学院政策・メディア研究科 研究員
有野 和真　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 客員研究官
(2014 年 3 月まで)

調査設計・分析・取りまとめ

林 和弘　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 上席研究官
白川 展之　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 客員研究官
独立行政法人 新エネルギー・産業技術総合開発機構 技術戦略研究センター 研究員
慶應義塾大学 大学院政策・メディア研究科 特任講師
公益財団法人 未来工学研究所 連携研究員
野村 稔　文部科学省 科学技術・学術政策研究所 科学技術動向研究センター 客員研究官

付録 1　時系列セッションネットワークの生成手順

本調査研究で提案した時系列セッションネットワークの生成方法と、開発した分析用ソフトウェアの関係を図 A1 に示す。以下に、データベース構築、分析データ、テキスト処理・ネットワーク生成、ネットワーク可視化について説明する。

(1)　データベース構築

最初に、カンファレンスプログラムから、カンファレンスの名称、開催年、セッション、各セッションで発表された論文の題名、著書の情報を抽出した。次に、学術文献データベースから対応する論文の著者別所属組織とアブストラクトを収集し、これらの情報をデータベースに格納した。

(2)　分析データ

データベースから各論文のアブストラクト、論文の発表されたセッション、開催年のデータを抽出し、CSV 形式のファイルに出力して分析データとした。

(3)　テキスト処理・ネットワーク生成

アブストラクトを記述するテキストデータに、名詞や動詞の語形変化を吸収するステミング処理を適用してから、各単語について*tf-idf*を求めて論文間類似度を計算した。ステミング等の処理は統計処理言語のR[4]のtmパッケージに含まれる機能を利用した。さらに、論文間類似度とセッション、カンファレンス開催年の情報を用いて、セッション間を接続するネットワークを生成した。

(4)　ネットワーク可視化

生成されたネットワークはネットワーク可視化ソフトウェアGraphViz[5]を用いて可視化した。Microsoft Windows 7 に含まれるPowerShellを用いて、RとGraphVizを連携させ、これらの処理を自動化した。

[4] The R Project for Statistical Computing, http://www.r-project.org/
[5] Graphviz – Graph Visualization Software, http://www.graphviz.org/

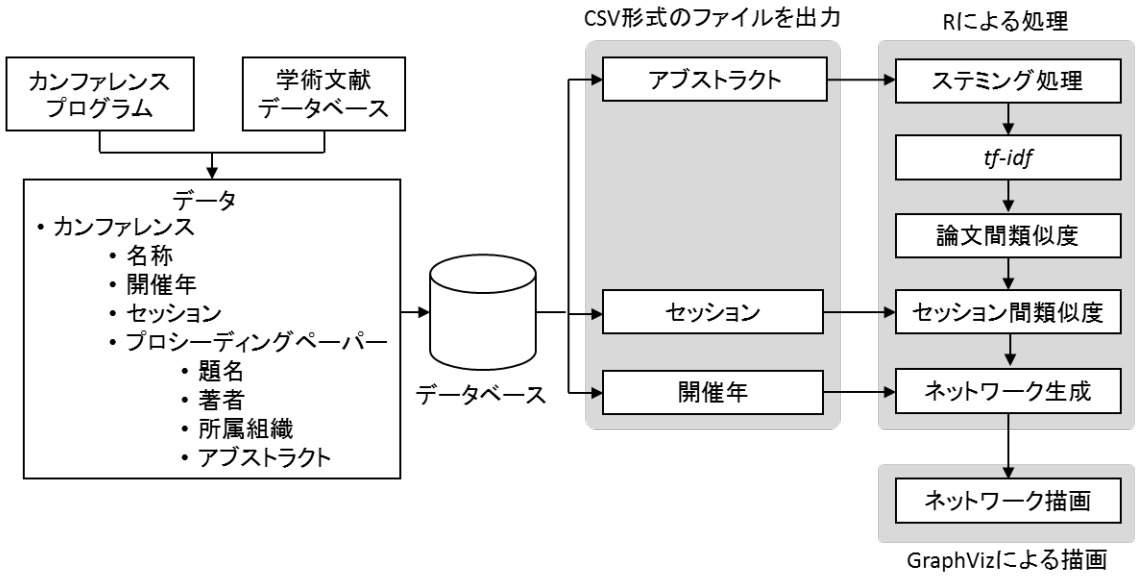


図 A1　時系列セッションネットワークの分析方法と開発したソフトウェアの関係

付録 2　ソフトウェアとサンプルデータ

本調査研究で作成したソフトウェアとサンプルデータは科学技術・学術研究所ホームページのデータ・資料として公開している。以下では、ソフトウェアを使用するための外部プログラムの設定と実行方法、サンプルデータファイルについて説明する。

本調査研究で作成したソフトウェアは、表 A2 に示す条件で動作を確認した。

表 A2　オペレーティングシステムと外部ソフトウェア

Microsoft Windows 7 Professional (64bit)	Service Pack 1
R	Version 2.15.3
GraphViz	Version 2.28.0

付録 2.1　環境設定とソフトウェアの操作方法

R と GraphViz を連携させるために、PowerShell の設定を変更する。

(1) PowerShell の起動

Windows 7 のすべてのプログラム > アクセサリ > Windows PowerShell > PowerShell

(2) ExecutionPolicy の設定

PowerShell から以下のコマンドを入力し、Scope: LocalMachine を Unresticted に設定する。

```
Get-ExecutionPolicy –list
Set-Execution Policy Unrestricted
Get-ExecutionPolicy –list
```

R を起動し、データ処理に使用するパッケージ(tm, igraph, SnowballC, proxy)をインストールしておく。以下のコマンドを R に入力し、セッションネットワークのデータを生成し、PowerShell を介して GraphViz を起動し、ネットワーク図を自動生成する。

```
# 作業ディレクトリの設定 (c:¥¥...はプログラムの展開先である。)
setwd(“c:¥¥...¥sample¥¥SessionNetwork¥¥WWW”)

# R スクリプト本体の読み込み
source(“..¥¥Scripts.R”)

# セッション・論文データの読み込みと初期化
```

```
GlobalInitialize(“c:¥¥...¥¥sample¥WWW2002-2011.csv”, “WWW”)

# ネットワーク生成とネットワーク図描画
SaveSelectedSample()
```

ネットワーク生成時のしきい値は Scripts.R の SaveSelectedSample()内にある SaveSelectedNetwork(…)で設定している。SaveSelectedNetwork(…)では、ネットワークの基点となるセッション名、セッション間接続のしきい値(threshold)、出力フォルダ・ファイル名(fileName)、学会識別タグ(filterConfName)を設定している。

付録 2.2　サンプルデータ

サンプルデータファイル(sample¥WWW2002-2011.csv)には、国際会議名略称、開催年、セッション名、論文名、アブストラクト、著者情報を記述している。

DISCUSSION PAPER No. 110

国際学会に注目した萌芽的研究領域の発展過程分析
－ World-Wide Web Conference の事例分析 －

2014 年 11 月

文部科学省　科学技術・学術政策研究所
科学技術動向研究センター

古川 貴雄　森 薫　有野 和真
林 和弘　白川 展之　野村 稔

〒100-0013
東京都千代田区霞が関 3-2-2 中央合同庁舎第 7 号館 東館 16 階
TEL：03-3581-0605　FAX：03-3503-3996