様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　６　月　１２　日現在

機関番号：　３２６０８
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０３７５

研究課題名（和文）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

研究課題名（英文）Quantitative analysis of evolutionary process for emerging research fields from the aspect of open innovation

研究代表者

古川　貴雄（Furukawa, Takao）

共立女子大学・家政学部・教授

研究者番号：７０２６２６９９

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：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、基本データ生成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。先端領域における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の関係を用いて、研究領域の発展過程を示す時系列ネットワークを用いた分析手法を示した。

研究成果の概要（英文）：This study discussed a method to analyze the evolutionary process of emerging research for aiming to apply science technology load mapping. Co-citation analysis as a representative bibliometric technique and text analysis were compared from the aspect of basic data structure, relationship among scientific articles, stability of analysis, coverage of information retrieval, required time for creating basic data. The text analysis that can process fuzzy data was feasible for extracting knowledge in advanced disciplines. This study proposed a method using time-series networks visualizing the evolutionary process of research fields, by using relationships between conference sessions and abstracts of proceeding papers.

研究分野：　計量書誌学

キーワード：　萌芽的研究　計量書誌学　データベース

１．研究開始当初の背景

萌芽的な研究領域の発展はイノベーションの源泉であり、このような研究領域に対して戦略的な目標を設定し、公的研究開発投資により研究を推進するといったイノベーション政策が必要とされている。例えば、ヒトゲノムプロジェクトは、1990 年代末に米国の NIH(National Institute of Health)とDOE(Department of Energy)によって開始され、公的研究開発投資による大学や公的研究機関におけるゲノム関連研究の推進だけでなく、ベンチャー企業等民間セクターを含めたゲノムシーケンサーの実用化・商業化を加速した。さらに、高性能ゲノムシーケンサーの登場は、ゲノム創薬などの新たな医療分野の市場・雇用の創出や、難病の治療といった社会的課題の解決への貢献が期待されている。

２．研究の目的

研究成果は学術論文や特許として公開されており、研究の発展過程の分析に共引用分析などの手法が用いられている。しかし、ある論文が論文誌に投稿されて受理され、さらに発表された論文が引用されるまでには程度の時間がかかるという問題がある。さらに、文献データベースにインデックスされ引用情報が利用可能になるまでの時間も無視できない。

本研究では、速報性の高い国際学会のプロシーディングのデータを分析することにより、萌芽的研究領域の発展過程における、知識の生成・蓄積、共有・再利用、知識の流動の様子を定量的に分析する。さらに、知識の生成から流動に至るパターンについて類型化を試みる。

３．研究の方法

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研究領域を例に、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。

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。

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
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
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。
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。
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。
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。

４．研究成果

計算機科学の研究領域では、他の研究領域と比較して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。

(1) 論文・セッションの類似度

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*) 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。*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。

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

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。

(2) 時系列ネットワーク生成アルゴリズム

以下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。

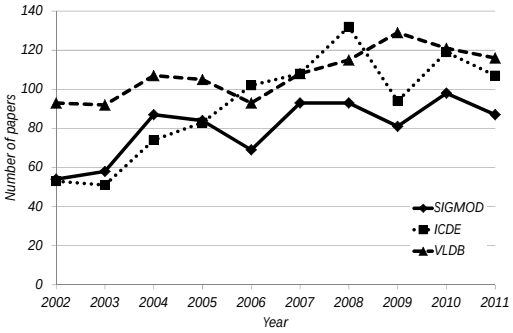
① 基準年からルートノードとなるセッションを選択する。基準年以外の全セッションノードを接続されるセッションノードの候補とする。
② 各セッション候補について、ルートセッションとの類似度を計算する。
③ セッションペアの類似度が設定値よりも大きい場合、セッションノード間を接続するエッジを挿入する。接続されたセッションは候補セッションノードから除く。
④ 新たに接続されたリーフノードを選択し、リーフノードが含まれる年のセッションを候補セッションノードから除く。
⑤ 各候補セッションノードについて、リーフセッションノードとの類似度を計算する。
⑥ ステップvで計算したセッション間類似度が設定値よりも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

(3) 時系列ネットワークの分析例

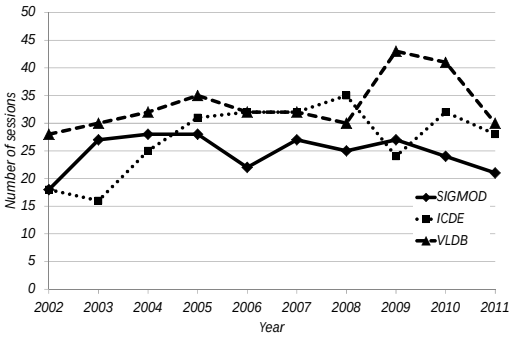
情報科学におけるデータベース研究の発展過程を分析するために、先端研究事例の発表される以下の国際学会に注目した。

- SIGMOD: ACM SIGMOD (Special Interest Group on Management of Data) Conference
- ICDE: IEEE International Conference on Data Engineering
- VLDB: International Conference on Very Large Data Base

これらの国際学会で発表された論文、及び、企画されたセッション数の推移を図１に示す。



(a) 論文数



(b) セッション数

図 1 SIGMOD, ICDE, VLDB における論文数とセッション数の推移

次に、以下のセッションに注目して分析を行った。

- Service-Oriented Computing, Data Management in the Cloud(SIGMOD 2011)
- Web Applications and Cloud Computing (ICDE 2011)
- MapReduce and Hadoop(VLDB 2011)

① クラウドにおけるサービス指向コンピューティング・データマネジメント

図 2 に、2002～2011 年に開催された SIGMOD について、Service-Oriented Computing, Data Management in Cloud というセッションに注目して生成したネットワークを示す。図には、国際学会の名称と開催年、セッション名、アブストラクトから抽出された tf-idf の高い単語を示している。さらに、異なったセッション間で共有されている単語を矢印上に示した。また、出力エッジの多いセッションノードは赤く、入力エッジの多いセッションノードは青く表示した。紫のセッションノードは入力エッジも出力エッジも多いことになる。

このネットワークから、Service-Oriented Computing, Data Management in Cloud という研究は、2004 年の Indexing and Tuning、2008 年の Special Platforms、2006 年の Replication, Cashing and Pub/Sub から派生したことがわかる。さらに、これらのセッションを遡ると、2002 年の Path Indexing と 2003 年の

Subscription Systems に到達している。

抽出された単語から、2010 年の Cloud Computing & Internet Scale Computing は、高性能データベースのための基盤技術、広域分散システム、メッセージ転送手法に関する研究と解釈できる。従って、サービス指向コンピューティングという研究は、インデキシング、広域分散システムのためのプラットフォーム、多くのユーザーを対象としたインターネットサービスといった要素技術が集まって形成されたと考えられる。

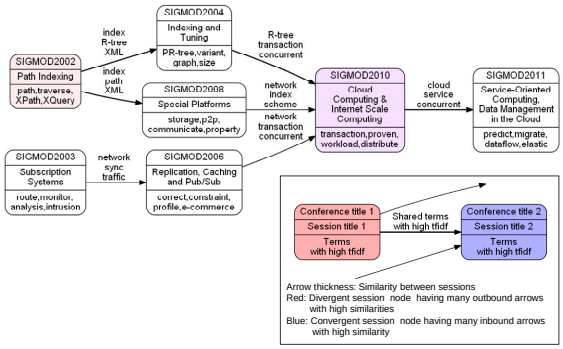


図 2　2011 年の Service-Oriented Computing, Data Management というセッションに至る時系列ネットワーク

② MapReduce と Hadoop

図 3 に、2002～2011 年に開催された VLDB について、MapReduce and Hadoop というセッションに注目して生成したネットワークを示す。2011 年の MapReduce and Hadoop というセッションノードから遡ると、2010 年の Cloud Computing, 2009 年の MapReduce というセッションノードに辿りつくことがわかる。一方、2007 年の Information Extraction and Text や 2002 年から 2004 年にかけて XML や Query Processing といった単語を含むセッションにも接続されている。

図の上側の流れは、大規模並列、スケーラブルコンピューティングのための基盤技術を示しており、下側の流れは、2000 年代初期の XML クエリ処理から派生した非構造化データの情報抽出に関する研究を示している。従って、MapReduce に関する研究は、これらの技術に関する研究が統合されて発展した研究と考えることができる。

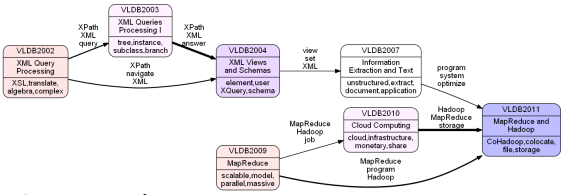


図 3　2011 年の MapReduce and Hadoop というセッションに至る時系列ネットワーク

このような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。

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。

５．主な発表論文等

〔雑誌論文〕（計　1 件）

[1] Takao Furukawa, Kaoru Mori, Kazuma Arino, Kazuhiro Hayashi, Nobuyuki Shirakawa, Identifying the Evolutionary Process of Emerging Technologies: A Chronological Network Analysis of World Wide Web Conference Sessions, Technological Forecasting & Social Change, 査読有, Vol. 91, pp. 280–294, Feb. 2015.

６．研究組織

(1)研究代表者

古川 貴雄（FURUKAWA, Takao）
共立女子大学・家政学部・教授
研究者番号：70262699

(2)研究分担者

林 和弘（HAYASHI, Kazuhiro）
文部科学省・科学技術・学術政策研究所・上席研究官
研究者番号：00648339

白川 展之（SHIRAKAWA, Nobuyuki）
文部科学省・科学技術・学術政策研究所・主任研究官
研究者番号：20556071

長谷川 誠（HASEGAWA, Makoto）
東京電機大学・工学部・教授
研究者番号：80303171