CM-001

論文引用ネットワークを利用した類似研究者検索手法
Similar Researcher Search Method Based on Paper Citation Network

中野 翔† 内田 真人†
Sho Nakano Masato Uchida

1. はじめに

共同研究プロジェクトの増加[1]や産学連携推進の気運の高まり[2]に伴い，研究上の関心事が類似した研究者同士のマッチングが重要となっている．また，研究環境の国際化に伴い，国外の研究者とのコネクションも重要となっている．しかし，学術分野の専門分化が進んだことで，自身と研究の興味が類似した研究者を探し出すことは困難となっている．例えば，世界各地の研究者が多数集まる大規模な国際会議においては，多数のセッションが並列で実施されることも多く，その中から自身と研究上の関心事が類似した研究者による発表を探し出すことは容易ではない．そのため，研究者自身がその時点で認知できていなかった類似研究者を検索することができれば，国際会議のように多数の研究者が集まるイベントの中で，自身の研究にとって有用な研究者間のコネクションを確立しやすくなることが期待できる．また，共同研究プロジェクトの促進や，目を通すべき研究論文の精査できるようになることが期待できる．

本論文では，研究内容の類似性を判定する指標として，論文の被引用関係に着目した新しい類似研究者検索手法を提案する．まず，Google Scholar[1] の引用文献表示機能から得られる被引用文献情報を用いて，研究者をノード，研究者間の引用関係をリンクとするような，木構造からなる引用ネットワークを作成する．そして，その引用ネットワーク上から引用関係の強い研究者を抽出し，類似研究者として出力する．本論文では，アンケート調査の結果などから，論文の引用ネットワークに基づく類似研究者検索が効果的であることを示す．また，入力パラメータを調整することで，影響力の大きい研究者や筆頭著者率の高い研究者を優先するなどの，ユーザのニーズに沿った検索を行うことが可能であることを示す．

2. 関連研究

研究者がある研究に関する研究動向を調査する際，関連論文などの「先行研究」を調査する場合と，その研究分野における「先行研究者」を調査する場合が考えられる．前者のアプローチに関しては，IEEE Xplore[2] や CiteSeer[3] などの関連論文を検索できるサービスが充実している．また，関連論文検索に関する研究も多く行われている．これに対して，研究者を人物単位で検索できるサービスは普及しておらず，多くの研究者に利用されているとは言えない．しかし，共同研究の増加などに伴い，研究テーマから先行研究者を探し出すシステムの需要は高く，研究内容から研究者を検索する手法もいくつか提案されている．以下では，これらの関連研究を概観する．

2.1 関連論文検索システム

科学技術・学術基盤調査研究室が 2015 年に発表した調査[3]によると，全世界で発表されている論文量は 1980 年代に比べて現在は約 3 倍となっており，学術的な研究活動は世界的に拡大傾向であるといえる．

研究動向の調査やアイディア着想のために，Google Scholar や CiNii[4] などの大規模学術文献データベースが広く利用されている．このようなデータベースから学術文献を収集する研究者が増加している一方，登録される文献数の増加に伴って，データベースから目的の文献を見つけることは困難になっている．このような問題を解決するために，効率的な関連論文検索手法が多く提案されており，その中でも論文の引用ネットワークを利用した論文検索システムを提案している研究がいくつか存在している．

コンピュータサイエンス分野における学術論文を多く扱っているデジタルライブラリの CiteSeer では，関連文献の検索に CCIDF アルゴリズムを採用している．Huynh ら[4]は，この CCIDF アルゴリズムに共引用・共参照関係を考慮した引用ネットワークを適用することで，文献間の関連度をより高い精度で求めることができる CCIDF+アルゴリズムを提案している．CCIDF+による関連文献の検索は CCIDF による検索に比べ 7〜10%程精度が改善している．この研究からも学術文献間の類似性評価に，論文の被引用関係を適用可能であることがわかる．

Naqvi ら[5]では，引用関係に特定の重みを付けることで引用関係の強度を定義し，複数の引用論文の中から関連性の高い論文を抽出するシステムを提案している．この研究では 2 つの文献の引用関係を評価する際に，キーワード・カテゴリー・著者の 3 つの項目において，両文献に共通している項目の有無に基づいたスコアを与えることで，類似性の高い論文の抽出を可能としている．

2.2 研究者検索システム

全世界の論文量と同様に，主要国の研究者数も増加傾向にある[3]．このような中で研究活動を拡大するためには，より多くの研究者とコネクションを持ち，学術的な連携を行うことが重要である．したがって，自身と研究傾向の近い研究者を検索するシステムの需要は高いと考えられ，研究者のマッチングを支援する研究もいくつか行われている．

堀ら[6]は，組織内の研究者総覧データベース及び CiNii のデータから得られる共著関係パターンから，協調フィルタリングを用いて大学内から共同研究者の候補を検索するシステムを提案している．この研究では，論文のタイトル・概要に含まれる専門用語から特徴ベクトルを作成し，協調フィルタリングを通して共同研究者候補を推薦する．この研究では，研究者の論文情報から合理的な共同研究者候補

† 早稲田大学基幹理工学研究科 情報理工・情報通信専攻
Department of Computer Science and Engineering, Waseda University
[1]https://scholar.google.co.jp/
[2]http://ieeexplore.ieee.org/Xplore/home.jsp
[3]http://liinwww.ira.uka.de/bibliography/Misc/CiteSeer/
[4]https://ci.nii.ac.jp/

を検索することに成功している．しかし，検索の質がクラスタ数に大きく依存しており，共同研究者候補が小規模な組織内の研究者に限定されてしまうという問題もある．

渡辺ら[7]は，ユーザとなる研究者と科研費採択者の研究者情報の DOM (Document Object Model) を作成し，2 者の DOM 間の類似度を評価することで，ユーザの研究内容に近い研究助成金を推薦するシステムを提案している．この研究では，研究助成金の募集期間が開示している採択者一覧および，過去に採択された科研費の研究内容の情報を，科学研究費助成事業データベースである KAKEN[1] から得ることで採択者の研究内容に関する DOM を作成する．この結果，研究者の研究内容に沿った研究助成金を推薦することを可能にしている．しかし，KAKEN のデータベースから採択者情報を取得しているため，国外の研究者はマッチング対象から外れてしまうという問題がある．

3. 提案手法

3.1 提案手法の概要

2.2 節に示した関連研究では，検索対象が国内の研究者に限定されていた．しかし，本論文では国外の研究者も検索対象に含める．これは，国内外問わず多くの研究者が集まる国際会議などでの利用も想定しているためである．また，研究者との新たなコネクションの確立を支援するために，検索システムのユーザ自身が名前や論文を認知していない研究者に焦点を当てる．そこで本論文では，ユーザ自身が執筆した論文における引用論文に関する情報（引用情報）ではなく，被引用論文に関する情報（被引用情報）に注目する．このような被引用情報を取得できる大規模論文データベースとして，本論文では Google Scholar を採用する．Google Scholar には，世界中の膨大な数の研究者や論文の情報が登録されており，論文の被引用情報も参照することができるため，本手法にとって最適なデータベースである．Google Scholar については 3.2 節で詳しく説明する．

本手法では，Google Scholar の引用文献表示機能を利用して，ユーザとなる研究者に関する被引用文献情報を収集する．取得した被引用文献情報に基づいた引用ネットワークを作成することで，引用ネットワークから引用関係の強い研究者を抽出し，類似性の高い研究者として出力する．以下では，ユーザとなる研究者を「起点研究者」と定義する．

本論文が提案する検索手法では「起点研究者」だけでなく，検索結果を操作するための 3 つのパラメータ α, β, γ も入力情報として与える．これらのパラメータは，検索時にユーザが重視したい研究者の要件を設定する目的で用意されている．α は，筆頭著者の優先度を設定するパラメータである．筆頭著者である論文の比率が高い研究者を優先することで，現役研究者の検索順位を上げることが期待できる．値域は $0 \leq \alpha \leq 1$ を取り，値が大きいほど筆頭著者の比率が高い研究者が優先される．β は，引用ネットワークにおいて被引用論文数の多い研究者の重要度を設定するパラメータである．被引用論文数を重視することで，その研究分野における影響力の大きい研究者の検索順位を上げることができる．値域は $0 \leq \beta \leq 1$ を取り，値が 0 に近いほど影響力の大きい研究者が優先される．γ は「起点研究者」の論文を直接引用している回数の多い研究者の，引用ネットワークにおける影響度を操作するパラメータである．値域は $0 \leq \gamma \leq 1$ を取り，値が大きいほど「起点研究者」の論文への引用回数が重視される．これらのパラメータが検索結果に与える影響は 3.3 節及び 3.5 節で詳しく説明する．

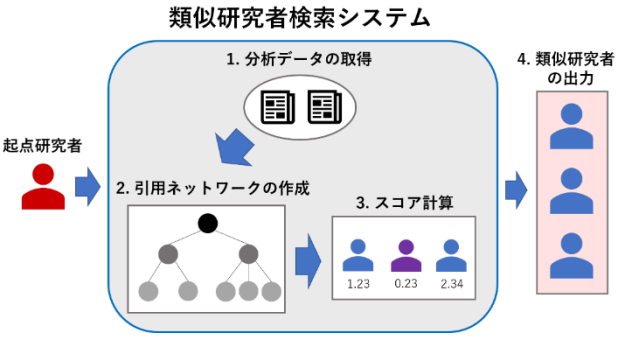

図 3.1　提案手法の概念図

本手法の全体的な概要図を図 3.1 に示す．本手法の全体的な流れは「分析データの収集」「引用ネットワークの作成」「スコア計算・結果の出力」の 4 つに大別することができる．各手順の詳細については 3.3〜3.5 節にて説明する．

3.2 Google Scholar

Google Scholar は，学術用途での文献検索を目的とした学術研究データ検索サービスであり，論文・学術誌・出版物の全文やメタデータにアクセスすることができる．

Enrique ら[8]によれば，Google Scholar に登録されている学術文献の総数は，2014 年 5 月時点で 1.6 億件にも及んでいる．また，学術ウェブ上で利用できる英語で書かれた学術文献の総数は 1.14 億件だと推計されていることから，Google Scholar が極めて大きい規模の学術データを保有していることが示されている．Google Scholar では，検索エンジンサービスと同様に，検索クエリを入力することで学術文献の検索を行うことができる．検索結果として表示されるデータは，文献のタイトルや著者等の基本的なメタデータだけでなく，文献が収録されている学術誌の情報を参照することができる．また，web 上に全文が公開されている場合は，公開されている web ページへのリンクも参照することができる．

Google Scholar の特徴は，登録されている学術データの規模だけでなく，文献引用情報にもアクセスできることが挙げられる．Google Scholar で表示される文献情報には，文献の引用回数と引用文献へのリンクが含まれているため，これらの情報を用いてその文献が持つ引用関係を把握することができる．同様に文献の引用関係に関する情報を参照できる学術データベースとして Web of Science[2] が挙げられる．本論文では，誰でも無料で利用することができる利便性と，文献がデータベースに登録される速報性の高さを考慮して，Web of Science ではなく Google Scholar を採用した．

Google Scholar には学術文献の検索ページだけでなく，研究者ごとのユーザページが存在する．Google Scholar にアカウントを登録している研究者ならば，ユーザページから研究者の発表している論文や出版物，そのユーザの h 指標や i10 指標などを閲覧することができる．各ユーザページの URL には，ユーザの識別子として 12 文字の ID が含まれており，この ID を用いることで研究者を特定することができる．本論文では，この ID を「ユーザ ID」と呼ぶ．

[1] https://kaken.nii.ac.jp/ja/
[2] https://webofknowledge.com/

3.3 分析データの収集

本論文で作成する引用ネットワークは，研究者がノード，ノード間リンクの重みが研究者間の引用論文数となる木構造である．引用ネットワークを作成するために必要なデータは，ノードに対応する研究者のユーザ ID と，研究者間の引用論文数データである．以下では，これらのデータを Google Scholar から収集する方法について説明する．

まず，起点研究者に関するデータを収集する．起点研究者の Google Scholar ユーザページに登録されている論文の中から，引用回数の多い上位 100 件かつ被引用回数 5 件以上である論文の集合をP_0とする．ただし，登録されている論文が 100 件未満である場合は，全論文の中で被引用回数が 5 件以上の論文の集合をP_0とする．被引用回数が 5 件未満の論文をP_0から排除しているのは，影響度が低いと判断される論文を集計対象から外すためである．

次に，起点研究者の引用関係を調べる．Google Scholar では各論文に対して，対象の論文を引用している論文・出版物のメタデータを参照することができる．ここで，P_0に属する全論文に対して，その論文を引用している全論文と全書籍の著者のユーザ ID を取得し，著者ごとに引用回数を集計する．ただし，集計対象となる著者は Google Scholar に登録することでユーザページを保有している研究者のみに限定する．これは，ユーザ ID は Google Scholar にユーザページを保有している研究者のみに与えられているためである．また，引用回数の計算は単純な数え上げではなく，研究者r が著者である引用論文pに対して$A(p,r,\alpha)$を計算し，その総和を求めることで研究者rの「重み付き引用回数」とする．ただし，αは 3.1 節で説明した入力パラメータの一つであり，$A(p,r,\alpha)$は以下のように定義される．

$$A(p,r,\alpha)=\begin{cases}1 & (p\text{の筆頭著者が}r\text{ である})\\ 1-\alpha & (otherwize)\end{cases}$$

集計結果として，P_0に属する論文を引用している研究者の名前と重み付き引用回数のデータを得ることができる．ここで，起点研究者と引用関係のある研究者の集合を$R^{(1)}$とし，その人数を$m^{(1)}$とする．

次に，$R^{(1)}$に属する研究者の中の，起点研究者への重み付き引用回数が多い上位M人に対して，自身の論文を引用している研究者の名前と重み付き引用回数のデータを，同様の手順で収集する．ただし，Mは入力パラメータとして，ユーザが自由に設定できる値であるものとする．

起点研究者も含めた計$M+1$人のデータを収集した時点で，分析データの収集を完了とする．

3.4 引用ネットワークの作成

本論文で作成する引用ネットワークは，「起点研究者」を根とする高さ 2 の木構造である．この木構造では研究者がノードに対応し，研究者間の引用関係がリンクの重みに対応する．引用ネットワークの例を図 3.2 として示す．この例では，起点研究者の論文を引用している研究者が B と C，B の論文を引用している研究者が A と D，C の論文を引用している研究者が B と E と F となっている．なお，木構造の高さを 2 に限定しているのは，データサイズを抑制する一方で，A-B-A といった相互引用関係や，A-C-B-A のような三角形構造を含むことができるためである．この引用ネットワークを作成する方法は以下の通りである．

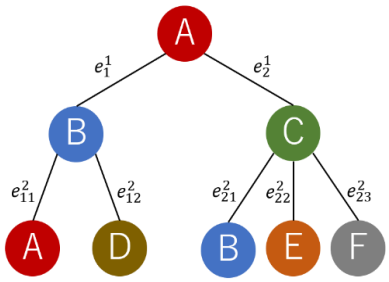

図 3.2 引用ネットワークのイメージ

まず，起点研究者に対応するノードを$r^{(0)}$とし，$r^{(0)}$を引用ネットワークの根とする．$r^{(0)}$は，$R^{(1)}$に属する研究者のノード全てを子ノードに持つ．ここで，$r^{(0)}$の子ノードを$r_i^{(1)}$とする．ただしiは，起点研究者への重み付き引用回数の順位であり，$1\le i\le m^{(1)}$である．さらに，$r^{(0)}$と$r_i^{(1)}$の間にあるリンクを$e_i^{(1)}$とする．起点研究者と同様にデータを収集したM人の研究者に対応するノードも子ノードを持つ．$r_i^{(1)}$に対応する研究者と引用関係にある研究者のノード，つまり$r_i^{(1)}$の子ノードを$r_{ij}^{(2)}$とし，その個数を$m_i^{(2)}$とする．ただしjは，$r_i^{(1)}$に対応する研究者への重み付き引用回数の順位であり，$1\le j\le m_i^{(2)}$である．さらに$r_i^{(1)}$と$r_{ij}^{(2)}$を結ぶリンクを$e_{ij}^{(2)}$とする．ノード間のリンクの重みは，親ノード($r_i^{(1)}$の場合は$r^{(0)}$)に対応する研究者への，子ノードに対応する研究者の重み付き引用回数と定義する．与えられたリンクに対してその重みを返す関数を$w(\cdot)$とし，リンクeの重みは$w(e)$で与えられるものとする．

3.5 スコア計算・出力

引用ネットワークを作成した後，その引用ネットワーク中にノードを持つ研究者ごとのスコアを計算する．本論文では，このスコアが高いほど起点研究者との類似性が高いと考え，検索結果における順位が高くなる．各研究者のスコアは，その研究者の属するノードと親ノードとの間のリンクの重み，つまり重み付き引用回数と同値とする．しかし，図 3.2 の例からもわかる通り，複数のノードが同一の研究者に対応している場合も存在する．このような場合は，各ノードと親ノードとのリンクの重みの合計を，その研究者のスコアとする．また，3.1 節でも述べた通り，複数の検索パラメータによるスコアの補正を行う必要がある．

これらを踏まえ，研究者rのスコア$SCORE(r)$を以下のように定義する．

$$SCORE(r)=\sum_{i=1}^{m^{(1)}}\left\{L\left(r,e_i^{(1)}\right)\times w\left(e_i^{(1)}\right)\times\left(\frac{1}{\sum w(e^{(1)})}\right)^{\beta}\right\}$$
$$+\sum_{i=1}^{M}\sum_{j=1}^{m_i^{(2)}}\left\{L\left(r,e_{ij}^{(2)}\right)\times w\left(e_{ij}^{(2)}\right)\times\left(\frac{w\left(e_i^{(1)}\right)}{\sum w(e^{(1)})}\right)^{\gamma}\times\left(\frac{1}{\sum w\left(e_i^{(2)}\right)}\right)^{\beta}\right\}$$

ただし，関数$L(\cdot)$, $\sum w(e^{(1)})$, $\sum w(e_i^{(2)})$を以下のように定義する．

$$L(r,e)=\begin{cases}1 & (e\text{の子ノードに属する研究者が}r\text{である})\\ 0 & (otherwize)\end{cases}$$

$$\sum w(e^{(1)}) = \sum_{i=1}^{m^{(1)}} w\left(e_i^{(1)}\right)$$

$$\sum w\left(e_i^{(2)}\right) = \sum_{j=1}^{m_i^{(2)}} w\left(e_{ij}^{(2)}\right)$$

ここで，β, γ はそれぞれ3.1節で説明した入力パラメータである．βは，全兄弟ノードのリンクの重みの合計が $\{\sum w(e^{(1)})\}^{1-\beta}$ または $\{\sum w\left(e_i^{(2)}\right)\}^{1-\beta}$ になるように正規化している．$\beta = 0$ の場合，論文の被引用数がそのままリンクの重みとして加算されるため，被引用数の多い論文の影響が大きくなる．βを 1 に近づけるほど，親ノードに当たる研究者の総被引用数で正規化されるため，研究者ごとの引用論文数の格差を小さくすることができる．γは，$R^{(1)}$に属する研究者に対応するノードを親ノードとするリンクの重みを計算する際に，その親ノードの持つ$e^{(1)}$の重みを比重として掛け合わせる．この際，$e^{(1)}$の重みは，$e^{(1)}$の重みの合計が $\{\sum w(e^{(1)})\}^{1-\gamma}$ になるように正規化された値を使用する．$\gamma = 0$ の場合，M個の子ノードの中で，起点研究者に対する引用数が多い研究者と最も引用数が少ない研究者が同等に扱われることになる．$e^{(1)}$の重みを比重として採用することで，「起点研究者」との引用関係がスコアに与える影響の重要度を大きくすることができる．

作成した引用ネットワーク上にノードを持つ全研究者に対して$SCORE(r)$を算出し，その順位を持って検索結果とする．ただし，$SCORE(r)$の順位は降順とし，値が最大の研究者を検索結果の最上位とする．

4. 実験と考察

提案手法による類似研究者検索の性能評価を行うために，現役の大学教員を対象にしたアンケート調査を実施した．また，パラメータα, βが検索結果に与える影響を調べるために対照実験を行った．実験で利用したデータについては 4.1 節，アンケート調査については 4.2 節，各対照実験については 4.3, 4.4 節でそれぞれ説明する．

4.1 実験に用いたデータ

本論文の実験では，早稲田大学基幹理工学部の現役教員 2 名を被験者とし， 3.3 節の手順に従って Google Scholar からデータを収集した．被験者X, Yのデータ収集時に設定したM, P_0に含まれる論文数，およびP_0に含まれる論文の引用回数の合計を表 4.1 に示す．

表 4.1　分析データの概要

	研究者X	研究者Y
M	50	50
P_0に含まれる論文数	23	58
P_0に含まれる論文の引用回数の合計	472	933

4.2 検索手法の性能評価実験

4.2.1 実験概要

提案手法による類似研究者検索手法の性能を評価するために，2 つの被験者実験を行った．

一つ目は，検索結果の正当性を確認するアンケート調査である．本手法では，研究内容の類似性が高く，本人が認知していない類似研究者の発見を目的としている．そこで，被験者に対して提案手法による類似研究者検索を行い，$SCORE(r)$の上位 50 名に対して，研究者の類似性と認知度を，被験者自身が評価する形式のアンケート調査を行った．

二つ目は，起点研究者との類似性の高さが，検索順位に正しく結びついているかを検証するアンケート調査である．被験者に対する検索結果の中から$SCORE(r)$の上位1%, 45～55%, 90～100%の集団を抽出し，それぞれランダムに1名を取り出した，3名をデータセットとする．このデータセットの中から，最も類似性の高い研究者を被験者自身が選択する形式のアンケート調査を行った．以下では，上位1%から選ばれた研究者と，最も類似性の高いとして選ばれた研究者が一致することを「正答」と呼ぶこととする．データセットに対する正答率が高ければ，類似性の高い研究者ほど，より検索順位の上位に現れると考えられる．

4.2.2 実験結果・考察

$SCORE(r)$の上位 50 人に対する認知度・関心度アンケートに対して，各アンケートの選択肢文を表 4.2 に，各アンケートの調査結果を表 4.3, 4.4 に示す．ただし，検索時のパラメータは$\{\alpha, \beta, \gamma\} = \{1.0, 0.0, 0.5\}$とした．

表 4.2　アンケートの選択肢文

認知度	1	名前も論文も記憶にない
	2	名前 or 論文を見た覚えがあるが知り合いではない
	3	共同研究の実績は無いが知り合いの関係である
	4	共同研究の実績がある
関心度	1	全く興味が湧かない研究内容
	2	あまり興味が湧かない研究内容
	3	ある程度興味が持てる研究内容
	4	強く興味が持てる研究内容

表 4.3　認知度アンケートの結果

	研究者X	研究者Y	全体
認知度:1	78%	70%	74%
認知度:2	10%	12%	11%
認知度:3	10%	18%	14%
認知度:4	2%	0%	1%

表 4.4　関心度アンケートの結果

	研究者X	研究者Y	全体
関心度:1	4%	48%	26%
関心度:2	12%	26%	19%
関心度:3	46%	26%	36%
関心度:4	38%	0%	19%

表 4.5　検索結果 － 人数（単位：人）

	研究者 X	研究者Y
$\alpha = 0.3$	34537	19503
$\alpha = 0.6$	37183	32269
$\alpha = 1.0$	38746	34535

表 4.6　アンケートの正答率

	研究者X	研究者Y	全体
$\beta = 1.0$	70%	50%	60%
$\beta = 0.5$	90%	70%	80%
$\beta = 0.0$	90%	80%	85%
全体	83%	67%	75%

各データセットに対する正答率を調査するアンケートに関して，各αにおける検索結果の出力人数を表 4.5 に，各被験者におけるアンケートの正答率をまとめた結果を表 4.6 に示す．ただし，検索時のパラメータは$w_\gamma = 1.0$とし，βは$\{0.3, 0.6, 1.0\}$の 3 パターンからデータセットを各 10 個用意し，被験者ごとに計 30 個作成した．また，αはデータセットごとに$\{0.3, 0.6, 1.0\}$の中からランダムに一つ選ばれるものとする．各被験者に対してデータセットを提示し，3 人の研究者の中から自身と研究分野の傾向が最も近いと考えられる研究者を記入してもらう形式のアンケートを行った．この際，提示されるデータセットの順番はランダムに並び替えられているものとし，選択肢の順番からは検索結果における研究者の順番が推測できないようになっている．

＜認知度・関心度アンケートに関する考察＞

認知度アンケートでは，検索結果の 85%を知り合いではない研究者が占めていることがわかった．特に，名前と論文を認知していなかった研究者の割合は 75%と高く，ユーザが認知していない研究者の検索に成功しているといえる．

関心度アンケートでは，研究者Xに対する検索結果の 84%に対して，研究内容に関心が持てるという回答が得られた．この結果から，研究者Xに対する検索では類似性の高い研究者の検索に成功しているといえる．対して研究者Yに対する検索では，研究内容に関心が持てる割合が 26%と低くなっている．これは，研究者Yの研究テーマが数年の間に変化しており，以前の研究テーマに関する論文に大きく影響されたことが原因だと考えられる．

＜検索順位と類似度に関する考察＞

表 4.5 から，βの設定値が大きいほど正答率が高くなっていることがわかる．また，全体の正答率は 75%となっており，おおむね高い正答率が得られた．正答率が高いということは，実際に類似性の高い研究者が検索結果の上位として選ばれていると考えられ，これらの結果から類似性と$SCORE(r)$の値が正しく相関していることが分かる．

4.3 対照実験：被引用論文数の影響

4.3.1 実験概要

提案手法におけるパラメータの一つであるβが，検索結果に与える影響を測定するために，βの値を複数用意して対照実験を行った．提案手法において，βは影響力の強い研究者，つまり被引用論文数の多い研究者の重要度を設定するパラメータとして定義されている．また，βが 0.0 に近づくにつれて被引用論文数の多い研究者が選ばれやすくなるものとしている．本実験ではβの値によって被引用論文数の多い研究者がどの程度上位に選ばれているのかを比較するための実験を行った．

まず，被験者 2 名を起点研究者として提案手法による類似研究者の検索を行った．この時，βの設定値を$\{0.0, 0.5, 1.0\}$とする 3 パターンを用意し，各パターンに対して検索を行った．また，$\{w_\gamma, \alpha\}$={0.5, 1.0}とした．次に，検索結果として得た研究者リストから上位 100 人を取り出して，各研究者の Google Scholar のユーザページに登録されている論文上位 100 件の引用論文数を調査した．

4.3.2 実験結果・考察

各検索結果の上位 100 人に選ばれた研究者に対して，Google Scholar に登録されている論文 100 件の引用論文数の合計値を集計し，その平均値をまとめた結果を表 4.7 に示す．また，中央値をまとめた結果を表 4.8 に示す．

表 4.7　被引用論文数 － 平均値

	研究者X	研究者Y
$\beta = 0.0$	10523.24	4345.90
$\beta = 0.5$	5243.90	3808.65
$\beta = 1.0$	2464.14	2688.84

表 4.8　被引用論文数 － 中央値

	研究者X	研究者Y
$\beta = 0.0$	6031.0	2148.5
$\beta = 0.5$	1964.0	1429.5
$\beta = 1.0$	904.5	1083.5

表 4.7 と表 4.8 から，βの設定値が小さいほど Google Scholar に登録されている被引用論文数の多い研究者が上位に上がっていることがわかる．$\beta = 0.0$の検索結果と$\beta = 1.0$の検索結果では，被験者Yの場合は被引用論文数の平均値が約 1.6 倍，被験者Xの場合では約 4.3 倍になっている．これらの結果から，βは被引用論文数の多い研究者の重要度を設定するパラメータとして適切に機能しているといえる．

4.4 対照実験：筆頭著者率の影響

4.4.1 実験概要

提案手法におけるパラメータの一つであるαが，検索結果に与える影響を測定するために，αの値を複数用意して対照実験を行った．提案手法において，αは筆頭著者率の高い研究者の優先度を表すパラメータである．本実験ではαの値によって筆頭著者率が高い研究者がどの程度上位に選ばれているのかを比較するための実験を行った．

まず，4.3 節での被験者 2 名を起点研究者として提案手法による類似研究者の検索を行った．この時，αの設定値を$\{0.3, 0.6, 1.0\}$とする 3 パターンを用意し，各αに対して検索を行った．また，$\{\beta, w_\gamma\}$={0.5, 0.5}とした．

次に，検索結果として得た研究者リストから上位 100 人を取り出して，各研究者の Google Scholar のユーザページに登録されている論文上位 100 件における，自身が筆頭著者である論文の割合を調査した．

4.4.2 実験結果・考察

各検索結果の上位 100 人に選ばれた研究者に対して，Google Scholar に登録されている論文 100 件の筆頭著者論文の割合を集計し，その平均値をまとめた結果を表 4.9 とする．また，中央値をまとめた結果を表 4.10 とする．

表 4.9　第一著者論文率 － 平均値

	研究者X	研究者Y
$\alpha = 0.3$	32.137%	31.633%
$\alpha = 0.6$	34.566%	35.408%
$\alpha = 1.0$	37.911%	37.620%

表 4.10　第一著者論文率 － 中央値

	研究者X	研究者Y
$\alpha = 0.3$	26.0%	26.8%
$\alpha = 0.6$	30.0%	30.0%
$\alpha = 1.0$	33.9%	34.2%

表 4.9 と表 4.10 から，αの設定値が大きいほど Google Scholar に登録されている論文の中で筆頭著者となっている比率が高い研究者が，上位に位置づけられていることがわかる．$\alpha = 0.3$の検索結果と$\alpha = 1.0$の検索結果では，両名ともに筆頭著者論文率の平均値が約 6%，中央値が約 8%上昇している．これらの結果から，αは筆頭著者論文率の高い研究者の重要度を設定するパラメータとして適切に機能しているといえる．

5. 結論

5.1 まとめ

本論文では，強い引用関係で結ばれている研究者同士は研究分野での類似性も高いという考えに基づいて，引用ネットワークから研究分野の近い研究者を検索する手法を提案した．本手法では Google Scholar の引用文献表示機能から得られる被引用文献情報を用いて，研究者をノードとする木構造からなる引用ネットワークを作成する．本手法によって，自身と研究テーマの傾向が近い研究者を見つけたい研究者が，国外も含めた自身の属していないコミュニティからも類似研究者を探すことが可能となる．

提案手法の実用性を評価する実験として，提案手法を用いた類似研究者の検索を行い，起点研究者として指定した研究者本人に，検索結果の妥当性を評価してもらう形式のアンケート調査を行った．アンケートでは，75%の割合で妥当な検索結果を得られるということがわかった．このアンケート調査によって，提案手法による類似研究者検索システムの性能の正当性を実証することができた．

また，複数の検索パラメータに対して対照実験を行うことで，各検索パラメータが検索結果に対して機能的な影響を与えることが出来ていることを確認することができた．

5.2 今後の課題

5.2.1 研究トピックを限定した検索手法

研究者が関心を持つ研究トピックは一つに限定されるとは限らず，複数のトピックに関して研究を行うことが多い．また，各トピックに対する関心度にも差異があり，研究動向によって変化していくことが考えられる．

本論文で示した提案手法では，ユーザページに登録されている全論文に対してスコア計算を行っている．そのため 4.2 節で示した研究者Yのように，研究動向が変化しているユーザに対して，十分な精度を得られないことが考えられる．このようなケースに対応するためには，研究トピックを限定した検索手法を行うことが有効である．データ収集時に登録されている論文情報からトピックを推定し，ユーザが指定したトピックのみを収集対象とすることで，更にユーザの関心に沿った検索性能を得ることができると考えられる．

5.2.2 実用化への展望

本論文で示した類似研究者検索手法を実用化することで，国内外の研究者マッチングなどに大きな効果をもたらすことが期待できる．実用例の一つとして，大規模な国際会議にて，同時並行で開かれている複数のセッションから，自身の研究と関連性が高いプレゼンテーションをリコメンドするシステムが考えられる．このようなシステムを開発するためには，対象となる国際会議のセッション発表者一覧を取得するのみならず，ユーザ及び全セッション発表者の被引用文献情報を収集する必要がある．そのため，被引用文献情報をできる限り短時間で収集できる手法を確立することが重要である．

謝辞

本研究を進めるのにあたり早稲田大学基幹理工学部の森達哉教授にご協力を頂いた．また，本研究の一部は，日本学術振興会における科学研究費補助金基盤研究(B)(課題番号 17H01742)による支援を受けている．ここに記し謝意を表す．

参考文献

[1] 文部科学省, “民間企業等との共同研究実施状況”, Tech. rep, URL:http://www.mext.go.jp/a_menu/shinkou/sangaku/sangakub/040 72301/001.htm
[2] 孫 媛, 西澤 正己, 根岸 正光, “日本の論文の共著関係からみた産学連携の現状分析”, “タイトル”, 情報知識学会誌, Vol.15, No.2, pp.89-92 (2005). DOI: 10.2964/jsik_KJ00003381807
[3] 科学技術・学術基盤調査研究室, “Japanese Science and Technology Indicators 2017”, Tech. rep, Aug 2017, URL: http://hdl.handle.net/11035/3178
[4] T. Huynh, K. Hoang, L. Do, H. Tran, H. Luong, S. Gauch, “Scientific publication recommendations based on collaborative citation networks”, 2012 International Conference on Collaboration Technologies and Systems (CTS), May 2012, pp.316-321, DOI: 10.1109/CTS.2012.626106
[5] M. Naqvi, A. U. Mailk, S. Razzaq, M. T. Afzal, “Extraction and visualization of citation network”, International Conference on Computer Networks and Information Technology, July 2011, pp.205-209, DOI: 10.1109/ICCNIT.2011.6020930
[6] 堀 幸雄, 今井 慈郎, 中山 堯, “共同研究の推薦のための協調フィルタリング”, 情報知識学会誌, Vol.18, No.2, pp.99-104 (2008), DOI: 10.2964/jsik.18-99
[7] 渡辺 孝信, 鎌田 真, 市村 匠, “科研費分類構造の類似度評価に基づいた研究助成金マッチング支援システムの開発”, 県立広島大学経営情報学部論集, Vol.8, No.2, pp.103-110 (2016), URL:http://harp.lib.hiroshima-u.ac.jp/pu-hiroshima/metadata/12340
[8] Enrique Orduña-Malea, Juan Manuel Ayllón, Alberto Martín-Martín, Emilio Delgado López-Cózar, “About the size of Google Scholar: playing the numbers”, CoRR, arXiv: 1407.6239, URL: http://arxiv.org/abs/1407.6239