修士論文

内容ベースフィルタリングと引用ネットワークを用いたハイブリッド法による科学論文レコメンデーション手法の考案と評価

相原 健史

主指導教員 Dam Hieu Chi

北陸先端科学技術大学院大学
先端科学技術研究科
(知識科学)

令和 2 年 3 月

目 次

図 目 次

表 目 次

第1章　序論

1.1　研究背景

1.1.1　論文数の増加

近年，開発途上国の研究活動において大きな発展が見られる．研究力を図る指標として，多くの世界大学ランキングでは，論文の質を示す被引用数と，研究のアウトプット量を示す論文数を用いている [1]．ここでは，図 1.1 に各国の科学および工学分野の論文数の推移を示す[1]．

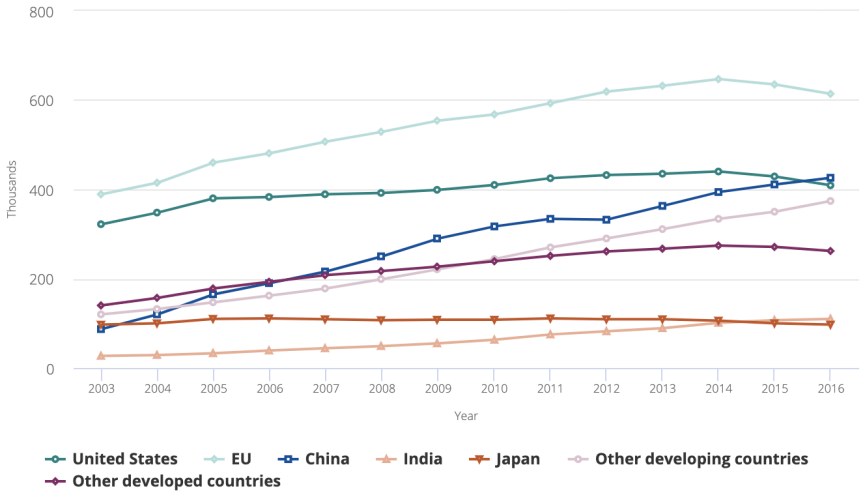

図 1.1: 各国の科学および工学分野における論文数の推移

EU，アメリカ，その他の先進国では 2014 年までは論文数は増加しており，2014 年以降では論文数に減少が見られる一方，中国，インド，その他の発展途上国では論文数は増加し続けていることが読み取れる．特に中国，インドは目を瞠るものがある．表 1.1 に各国の 2006 年と 2016 年の科学および工学分野の発表論文数を示す[1]．

中国では 2006 年に 189,760 本，2016 年に 426,165 本，インドでは 2006 年に 38,590 本，2016 年に 110,320 本の論文が公開され，10 年間で中国は約 2.25 倍，インドでは約 2.86 倍

[1]National Science Foundation による Science & Engineering Indicators 2018 より引用

表 1.1: 各国における 2006 年と 2016 年の科学および工学分野の発表論文数

Rank	Country	2006	2016
-	World	1,567,422	2,295,608
1	China	189,760	426,165
2	United States	383,115	408,985
3	India	38,590	110,320
4	Germany	84,434	103,122
5	United Kingdom	88,061	97,527
6	Japan	110,503	96,536
7	France	62,448	69,431
8	Italy	50,159	69,125
9	South Korea	36,747	63,063
10	Russia	29,369	59,134

の論文が公開されている．世界全体で見ると 2006 年は 1,567,422 本の論文が公開されているのに対し，2016 年には 2,295,608 本の論文が公開されており，10 年間で約 1.46 倍の数の論文が発表されている．このことから，先進国では科学および工学分野の研究活動に衰退の色が見られるが，発展途上国では今後も発展の傾向にあることが示唆される．また，世界全体での発表論文数は今後も増加する傾向にあると考えられる．

1.1.2 大学生・大学院生の研究活動

大学や大学院の卒業・終了要件要件として研究を課されている学生は多い．一例として，筆者の本研究における大まかなプロセスを図 1.2 に示す．

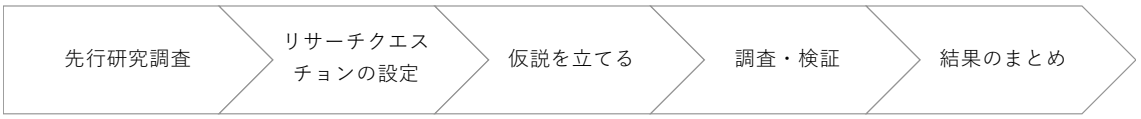

図 1.2: 研究の流れ

まず，先行研究調査では，興味がある分野に関する論文や文献などを読み，まだ明らかにされておらず，調査する価値がある事柄を見極めていく．次に先行研究調査で疑問に思ったことや調べたいことを，リサーチクエスチョンとして設定する．リサーチクエスチョンとは，研究において明らかにしたいことに対する問いであり，これを解決することを研究の目的とするものである．その後，リサーチクエスチョンに対して，仮説を立てる．仮説は，検証可能な範囲で設定し，仮説が正しいかどうかを検証するためのものである．そして，調査・検証をし，その結果をまとめるというプロセスを辿る．

上記のいずれも，研究においては重要なプロセスであるが，大学生・大学院生にとって先行研究調査は特に重要である．大学生や大学院生は，研究に対していわゆる「ビギナー」であり，大学院生である筆者を含め，専門分野の知識を十分持っているとは言い難い．先行研究の調べ方，研究論文の読み方やまとめ方など，先行研究調査の方法に関しても熟練されていない．そのため，リサーチクエスチョンを設定や仮説を立てるためにも，一般の研究者より多くの時間を先行研究調査に割く必要があると考えられる．

一方で，前節 1.1.1 に示したように世界全体での論文の発表数は増加の一途をたどっており，論文検索の難易度は格段に上がっている．学生にとってはこのような状況において，先行研究調査をより効率的に行う必要があることが示唆される．

1.1.3 論文のレコメンデーション

学生のみならず研究者であっても，莫大な論文情報から目的の論文を探し出すのは難しい．必要な論文や論文内の情報にたどり着くために，検索スキルを身につけたり，論文の読み方を工夫したりする，人間側からのアプローチ [2] がある一方で，コンピュータにこの問題をさせようと，様々な論文のレコメンデーション手法が考案されている．「レコメンデーション」とは，対象のユーザに必要な情報などを選択して提示することであり，平たく言えば「おすすめ」機能である．コンピュータが論文の内容や筆者の情報，論文を検索している人の検索情報などを用いて，レコメンドすべき論文を選択する．レコメンドするアイテムの特徴を利用する手法 [3]，レコメンデーションを利用するユーザの評価を用いる手法 [4]，また，論文の引用ネットワークや著者同士のソーシャルネットワークに基づく手法 [5][6] などが存在する．また，これらの手法を組み合わせたハイブリッド法があり，近年では組み合わせる手法や組み合わせ方によって，多くのアルゴリズムが考案されている．この論文レコメンデーションにより，研究者は情報過多から軽減され，関連する論文を容易に見つけられるようになることが期待されている．そこで，本研究では新たな論文のレコメンデーション手法を提案する．従来よりも優れたレコメンデーション手法を提案することによって，より効率的な先行研究調査を行えようにすることを目指す．

1.1.4 評価手法の妥当性

現在までに様々な論文のレコメンデーション手法が考案されているが，その手法が正しいかどうかを判別することは容易ではない．図 1.3 に論文レコメンデーションの評価手法として用いられる，論文の引用データによる評価手法の概要を示す．

コンピュータ上のレコメンデーションを行うシステム (以下レコメンダシステムと呼称) は，現在読まれている論文 O(レコメンデーションをするための始点となる論文) の内容，

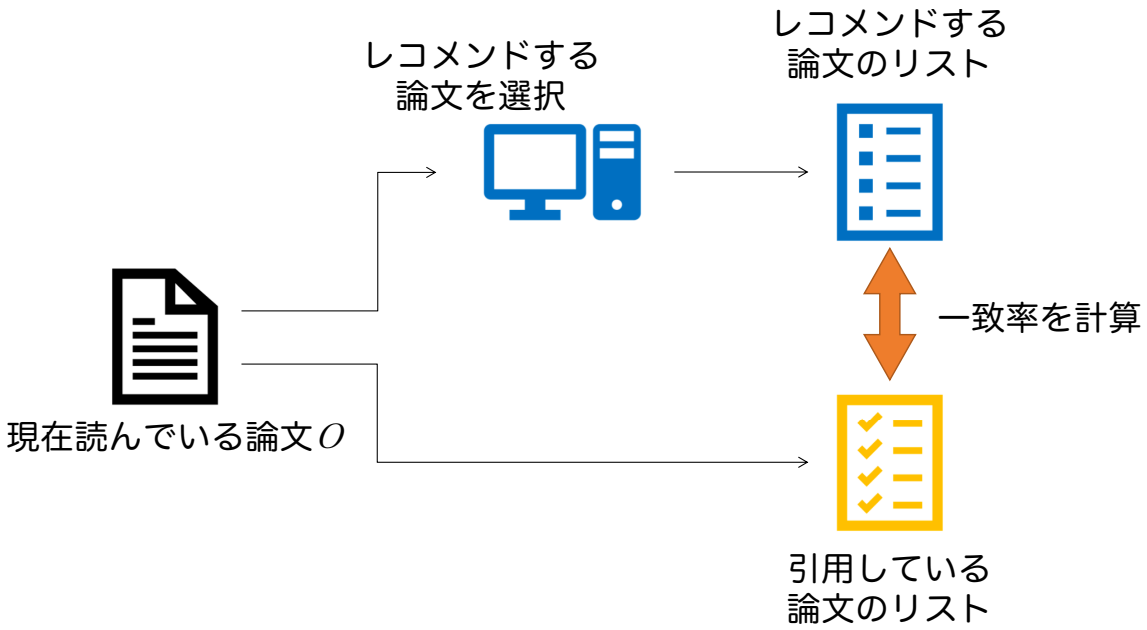

図 1.3: 引用データを用いた論文レコメンデーションの評価

筆者の情報，論文が掲載されている学術雑誌の情報などを用いて，レコメンドする論文を選択する．この時点で，選択された論文はランク付けされ，レコメンドされるべきものが上位にくるようにリスト化された状態になっている．そして，レコメンドする論文のリストと対象の論文の引用リストを比較し，一致率を求めることで，正しくレコメンドされているかを評価する．

ここで重要なのが，引用リストを正解データとして比較しているということである．引用リストを正解データとするのであれば，引用リストに載っているもののみをレコメンドすれば良いのであり，レコメンダシステムは必要ない．一方で，引用リストに載っていない論文でもレコメンドされるべき論文は存在する．例えば，対象の論文 O よりも後に発表された論文は引用リストに載せることは不可能だが，内容的に類似しており，より新しく優れた手法を用いている論文 R があるとする．レコメンダシステムはこの状況を考慮し，R を読むべきだと判断する可能性がある．しかし，引用リストを正解ラベルとしてしまうと，読むべきだと判断された論文 R をレコメンドすると，不正解だと判断されてしまう．このような現象は，後に発表された論文に限らず起こりうる．なぜなら，引用リストに含まれる論文は，通常，O の中で引用または言及された論文に限られ，言及されなかった論文は内容的に類似しており優れていたとしても，引用されないからである．このように，論文レコメンダシステムは必ずしも教師あり学習として評価することはできない．そのため，教師なし学習という観点での評価方法を検討する必要がある．

また，状況によっても，レコメンドされるべき論文は変化する．図 1.4 は，派生関係にある，2 本の論文の関係を表している．

論文 B は論文 A から派生した論文であり，論文 A と論文 B は親と子の関係にある．こ

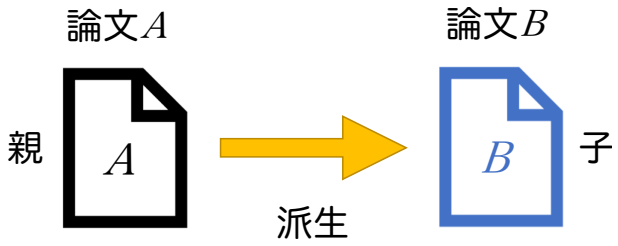

図 1.4: 派生関係にある 2 本の論文

こで，通常の研究者は自身の研究のために，最新の論文を読みたいという動機があるとする．この場合は，この研究者に対して，より後に出版された論文 B がレコメンドされるべきである．一方で，大学生や大学院生などのビギナー研究者は，その学問分野自体を理解したいという目的で論文を読むことがある．このとき，大元の論文を読むことで理解がより進むのであれば，より前に出版されている論文 A がレコメンドされるべきである．このように，レコメンドされる論文がユーザによって変わるのであれば，評価手法も変えなければならない可能性がある．

以上のような背景から，論文レコメンダシステムの正しい評価手法について再考する必要があると考えた．そこで本研究では，大学生・大学院生および研究者が論文を書くプロセスをモデル化し，論文レコメンダシステムをどのような評価手法を用いて評価すれば良いのかを検討する．

1.2 研究目的

本研究の目的は，論文検索にかかる時間を最小化することで，効率的な先行研究調査を可能にすることである．そのために，論文レコメンダシステムの正しい評価手法について検討する．また，同時に論文レコメンデーションのアルゴリズムを考案する．そして，実際にレコメンダシステムの構築および評価を行い，本研究で用いた評価手法によって適切に評価されているか議論する．

1.3 本論文の構成

本論文の構成を以下に示す．

第 1 章 序論

第 1 章では，研究背景および研究目的について述べた．

第 2 章 関連知識

第 2 章では，関連知識について述べる．

第 3 章 評価手法と研究手法

第 3 章では，本研究の提案手法と評価手法について述べる．

第 4 章 検証実験と考察

第 4 章では，検証実験の詳細と結果について述べる．

第 5 章 まとめと今後の課題

第 5 章では，まとめおよび今後の課題について述べる．

第2章　関連知識

本章では，関連知識として既存の様々な論文レコメンデーション手法の概観を述べる．また，論文レコメンダシステムに関連する研究について述べる．

2.1　レコメンデーション手法

現在まで，多くのレコメンデーション手法が考えられているが，それらは，

- 内容ベースフィルタリング (Content-Based Filtering),
- 協調フィルタリング (Collaborative Filtering),
- グラフベース法 (Graph-Based method),
- ハイブリッド法 (Hybrid recommend method)

の 4 つのカテゴリに分類することができる [7]．本節ではそれぞれの手法について説明する．

2.1.1　内容ベースフィルタリング

内容ベースフィルタリングは，アイテムの特徴とユーザのプロファイル情報に基づいてレコメンドを行う手法である [3]．図 2.1 は内容ベースフィルタリングを用いた論文レコメンダシステムの概略図である．

ユーザ X はレコメンドを行う対象のユーザである．内容ベースフィルタリングではまず，コンピュータが論文データから論文の特徴を学習しておく．この特徴の学習とは，論文の本文の内容やキーワードなどを対象に，テキストマイニングの手法を用い論文のクラスタリングなど行うことで，論文同士の類似度を求めておくことである．そして，ユーザ X のプロファイル情報を基に，学習済みの論文の特徴データからユーザの研究分野や興味に基づいた論文をマッチングさせることで，レコメンドする論文を求める．重要なのは，論文の「内容」を基に類似する論文を求めていることであり，これが内容ベースフィルタリングと呼ばれる所以である．

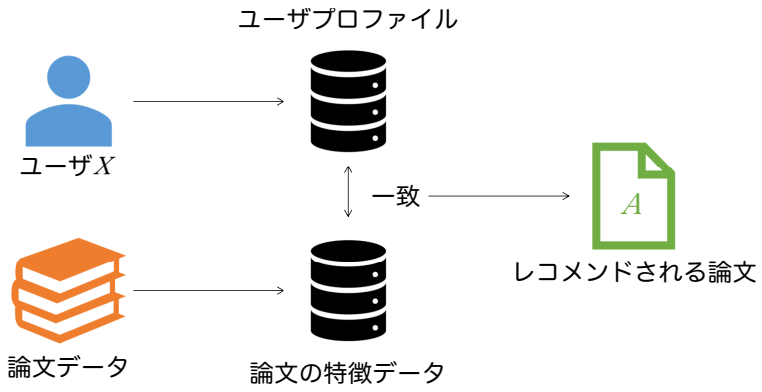

図 2.1: 内容ベースフィルタリングの概略図

2.1.2 協調フィルタリング

協調フィルタリングは，アイテムに対する他のユーザの評価に基づいてレコメンドを行う手法である [4]．図 2.2 は協調フィルタリングを用いた論文レコメンダシステムの概略図である．

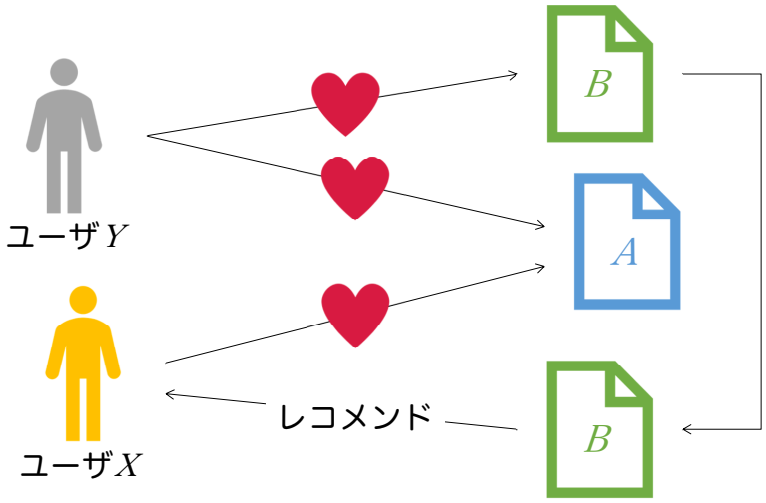

図 2.2: 協調フィルタリングの概略図

2 人のユーザ X と Y がいることを想定し，既にユーザ Y が論文 A と論文 B を高く評価しているとする．その後，レコメンドを行う対象のユーザ X は論文 A を高く評価した．このとき，レコメンダシステムによって，同じ論文 A を高く評価しているユーザ X とユーザ Y は似ているユーザだと判断されると，ユーザ X がまだ評価していない論文 B がレコメンドされる．協調フィルタリングのメリットは，アイテムの特徴などに依存することがないため，汎用的に用いることができることである．一方，デメリットとしては，ユーザによる評価を用いるので，誰にも評価されていないアイテムはレコメンドできないという弱点を持っている．

2.1.3　グラフベース法

グラフベース法は，論文の引用ネットワークや著者同士のソーシャルネットワークに基づく手法である．ここで言うグラフとは，ノードとエッジによって構成されるグラフ (図 2.3) のことであり，関数のグラフとは異なることに注意が必要である．

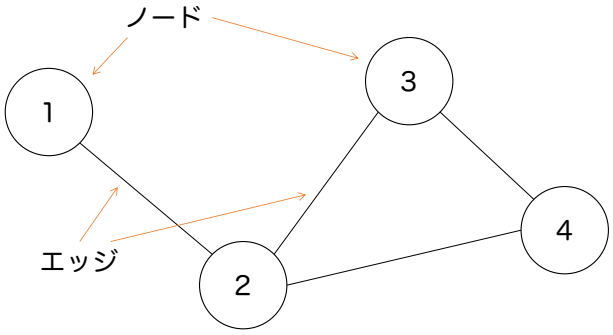

図 2.3: ノードとエッジ

論文や著者などをノードとして配置し，論文や著者同士の関係をエッジで結ぶと，それぞれの関係性をグラフとして表すことができる．グラフベース法では，このグラフの構造を解析した情報を利用して，関連する論文を見つけることでレコメンデーションを行う．グラフの構造を解析するときには，グラフの分散表現を獲得するための，グラフ埋め込み [8] と呼ばれる手法が用いられる．

2.1.4　ハイブリッド法

ハイブリッド法は，名前の通り 2 つ以上のレコメンデーション手法を組み合わせた手法を指す．単一の手法 (内容ベースフィルタリング，協調フィルタリング，グラフベース法) を用いたレコメンダシステムが存在する一方で，ハイブリッド法によるレコメンダシステムも多く存在する．ハイブリッド法の利点は，様々なレコメンデーション手法と多くの情報を組み合わせて使用できることに加え，結果としてもより優れたパフォーマンスを発揮することが証明されている [7]．表 2.1 は内容ベースフィルタリングと協調フィルタリングの比較表[1]である．

内容ベースフィルタリングは，アイテムの特徴を基にレコメンドを行うという性質から，レコメンダシステムのユーザ数に関わらずレコメンドが可能であったり，類似するアイテムをレコメンドしたりすることが可能である．しかし，アイテムに関する情報が得られない場合は，内容ベースフィルタリングでのレコメンドは困難であることや，ユーザ自身が知っているアイテムの特徴にレコメンドの対象が制限されやすいため，意外性のあるアイテムがレコメンドされにくいという問題がある．一方，協調フィルタリング

[1] 神嶌敏弘 推薦システムのアルゴリズム [9] より本文の説明を加えて作成

表 2.1: 内容ベースフィルタリングと協調フィルタリングの比較

項目	内容	協調	説明
多様性	×	○	ユーザ自身が知らないアイテムをレコメンド可能か
ドメイン知識	×	○	アイテムに関する情報がなくてもレコメンド可能か
スタートアップ問題	△	×	新たなユーザやアイテムでもレコメンド可能か
利用者数	○	×	他にユーザがいない場合でもレコメンド可能か
被覆率	○	×	全てのアイテムがレコメンドされうるか
類似アイテム	○	×	類似するアイテムをレコメンド可能か
少数派の利用者	○	×	少数派ユーザでも適切にレコメンドされるか

では，ユーザによる評価を基にレコメンドを行うため，まだ評価されていないアイテムをレコメンドすることができないという特徴があるが，アイテムの特徴によらずレコメンドできるため，内容ベースフィルタリングが苦手とするものを協調フィルタリングは得意とする．このように，複数の手法を組み合わせることで，単一の手法のみを使うときの弱点を補い合うことができる．

論文レコメンダシステムとしては，Gupta & Varma(2017)[10] は CCA(正準相関分析)によって，2 つの特徴量ベクトルを組み合わせることで，単一の手法よりも高い予測精度でレコメンドを行うことができている．

2.2 グラフベース法におけるグラフ構造の解析手法

2.2.1 概観

節 2.1.3 では，グラフベース法について簡単に説明したが，グラフ構造の解析には，グラフ埋め込み (Graph Embedding) と呼ばれる手法が用いられる．これは，グラフ上のノードやエッジなどの特徴を，ベクトル空間として埋め込むための手法である [11]．ベクトルに変換することにより，機械学習やディープラーニングなどの特徴量として利用しやすくなることが利点として挙げられる．2014 年に DeepWalk[12] という手法が提案され，これをきっかけとしてグラフ埋め込みの研究が盛んに行われるようになった．本節では DeepWalk を例にとり，DeepWalk のアルゴリズムについて解説する．

2.2.2 DeepWalk

DeepWalk とは，Deep Learning と Random Walk の 2 つの言葉を組み合わせた名前である．次の 2 つのステップを踏むことで，グラフ構造をベクトル空間に変換していく．

1. ランダムウォークによってグラフ上のノードをランダムに移動したときの配列情報を得る

2. Word2Vec によって配列情報をベクトルに変換する

2.2.3 ランダムウォーク

ランダムウォークとは，次の位置が確率的に無作為に決定される運動のことである．グラフのあるノードから出発し，次に移動する位置をランダムに決める．移動を複数回繰り返し，通ってきたノードの情報がそのまま配列情報とする．

グラフの例を図 2.4 に示す．このグラフでは A〜F までの 6 個のノードがある．A というノードを出発点として，ランダムウォークによって 5 回移動を繰り返した結果，A, B, D, F, E, B という配列が生成されたとする (図 2.5)．

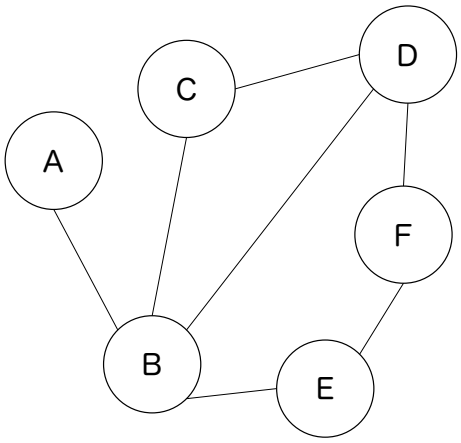

図 2.4: グラフの例

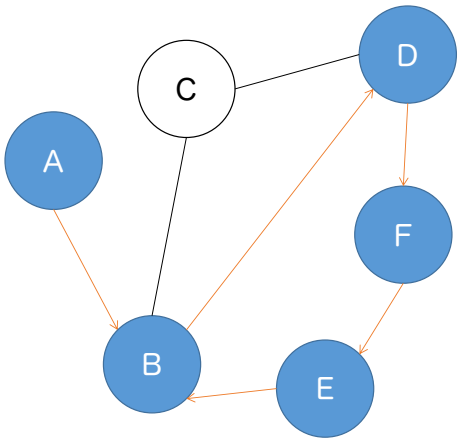

図 2.5: グラフ上でランダムウォークをする例

このランダムウォークで得られた配列情報を文章とみなすことで，自然言語処理で使われるベクトル化手法を適用できるようになる (表 2.2)．DeepWalk では，この配列情報を Word2Vec という手法によりベクトルに変換する．

表 2.2: ランダムウォークで生成された配列と文章の比較

ランダムウォーク	A	B	D	F	E	B
文章	I	am	walking	in	the	park

2.2.4 Word2Vec

Word2Vec[13] とは，ニューラルネットワークを用いて単語の予測問題を解かせることで，単語のベクトル表現を得る手法である．ニューラルネットワークが単語の予測問題を解くとき，ニューラルネットワークは各ニューロンで重みを学習する．この重みを，単語ベクトルとして使用することができる．図 2.6 は Word2Vec が解く問題の例である．

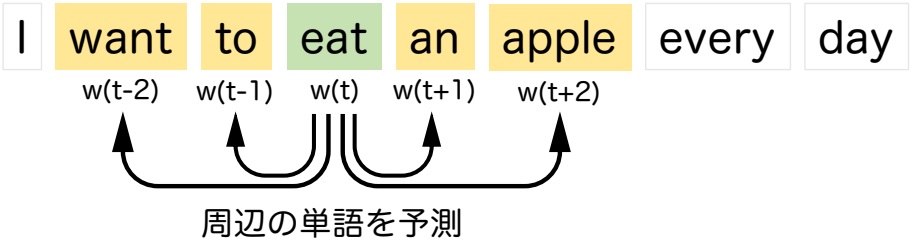

図 2.6: Word2Vec が解く問題の例

文章中のある 1 単語に注目する．この単語を w(t) とすると，その前後の単語は $\cdots$,w(t-2), w(t-1), w(t+1), w(t+2), $\cdots$ で表すことができる．ニューラルネットワークに w(t) を入力として与え，その周辺の単語 $\cdots$,w(t-2), w(t-1), w(t+1), w(t+2), $\cdots$ を予測させる．入力はイテレーション毎に $t := t + 1$ として，入力の単語を一つずつ後ろ方向にずらすことによって，学習データに含まれる単語について順に学習を行う．

続いて，Word2Vec で用いるニューラルネットワークのモデル[2]を図 2.7 に示す．

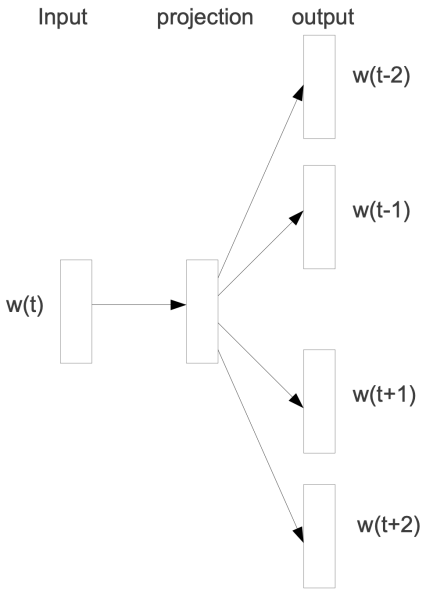

図 2.7: Word2Vec のニューラルネットワーク

[2]Mikolov et al. Distributed Representations of Words and Phrases and their Compositionality[13] より引用

Word2Vec のニューラルネットワークは，入力層，1 層の隠れ層，出力層のみからなる．学習させ終えたニューラルネットワークの入力層と隠れ層の間の重みを，単語のベクトルとして使用する．

2.2.5 Word2Vec の入力としてランダムウォークの配列を使う

Word2Vec よって文章中の単語をベクトルとして表現することができるが，グラフ構造のベクトル化も同様にして行う．グラフ上でランダムウォークをすることによって得られた配列を文章とみなし，Word2Vec のニューラルネットワークに入力として与える．こうすることで，それぞれのノードについてのベクトル表現が得られ，これを機械学習の特徴量などに利用できる．

第3章　評価手法と研究手法

本章では，論文レコメンダシステムの正しい評価手法について考える．また，本研究で提案する論文レコメンダシステムについて述べる．

3.1　研究の手順

本研究は以下の手順で進める．

1. レコメンダシステムを使用する場合の論文執筆モデルの構築
2. モデルを基にした評価手法の考案
3. 論文レコメンダシステムの構築
4. 評価手法の妥当性の検証

3.2　レコメンダシステムを使用する場合の論文執筆モデル

節 1.1.4 で既に問題点を述べているように，論文レコメンダシステムを正しく評価できなければ，いくら最適化されていたとしても正しいレコメンドはできず，システムには限界が訪れる．引用データを正解ラベルとして利用することは，予測精度がどれだけ良くなっても，無意味な学習になってしまう．

本研究では，論文レコメンダシステムにおける正しい評価手法を考えるために，研究者が論文レコメンダシステムを使用し，論文を執筆するまでのモデルを考える．このモデルを基に，論文レコメンダシステムがレコメンドに成功する，理想的なケース (図 3.1) と，論文レコメンダシステムがレコメンドに失敗する，望ましくないケース (図 3.2) を想定し，この 2 つの場合を定量的に評価するための評価式を考える．

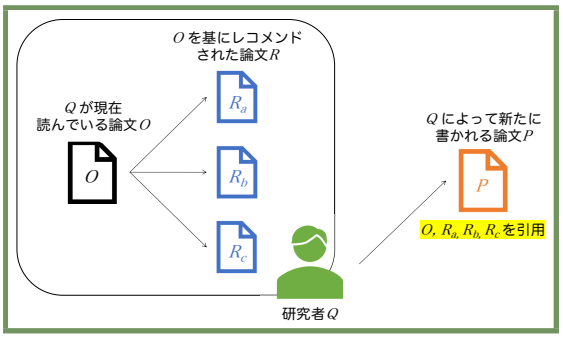

図 3.1: 理想的なケース

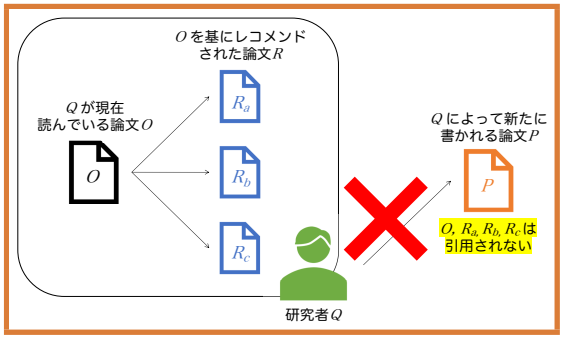

図 3.2: 望ましくないケース

3.2.1 論文の執筆

まず，研究者 Q が論文を書くことを考える (図 3.3).

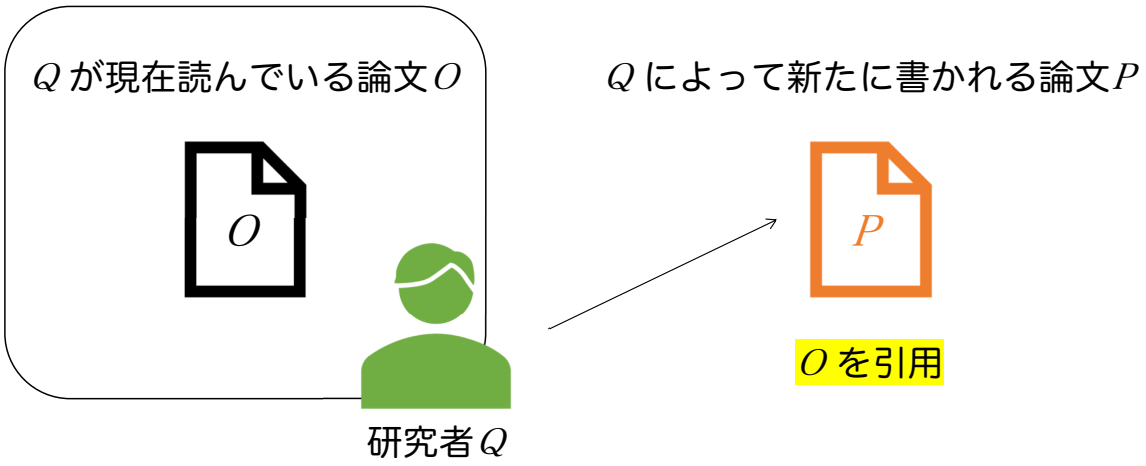

図 3.3: 研究者 Q の論文執筆モデル図

研究者 Q が現在読んでいる論文を O とする．研究者 Q は論文 P を書こうとしている．このとき，研究を進める上で論文 O が重要な役割を果たしていたと仮定すると，Q によって新たに書かれる論文 P では，論文 O が引用されていると考えられる．これを論理式で表すと次のようになる．

$$\text{Read } O \Rightarrow \text{Cite } O \tag{3.1}$$

3.2.2 理想的なケース

次に，研究者 Q が理想的なレコメンドが可能な論文レコメンダシステムを使って論文を書くことを考える (図 3.4)．ここで言う理想的とは，使用するレコメンダシステムは必

ず，研究者 Q が論文を書くにあたり重要な論文をレコメンドできるという意味である．

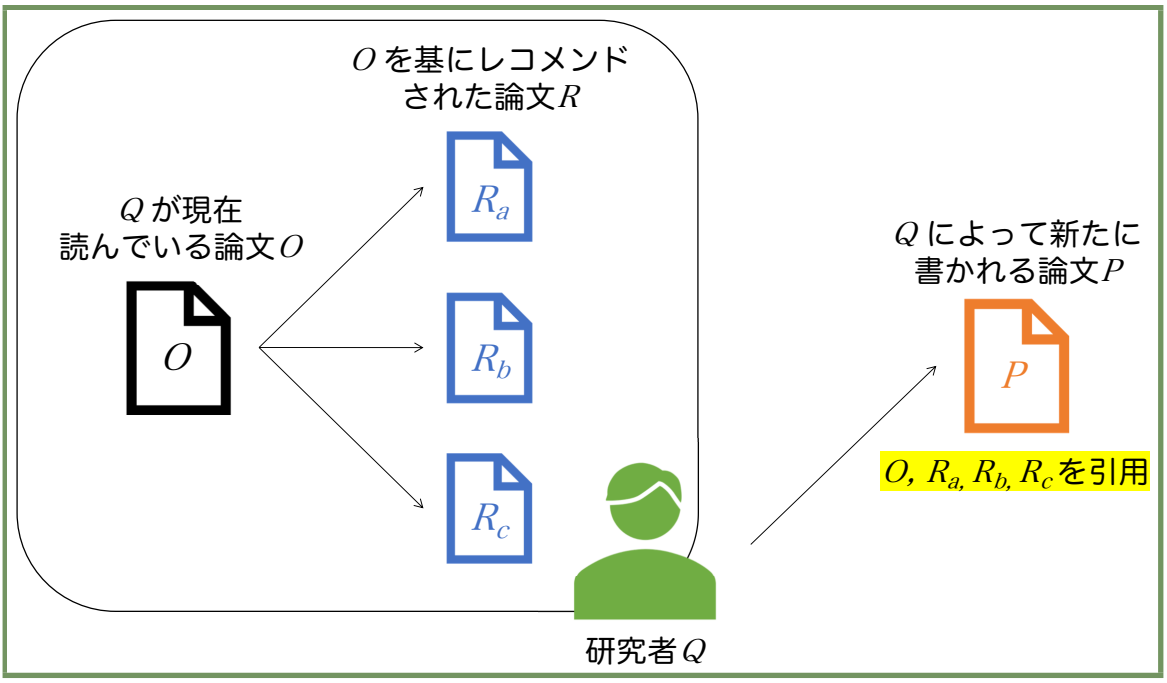

図 3.4: 理想的なレコメンドを行うケース

図 3.4 で使っているのは理想的な論文レコメンダシステムであるため，レコメンドされた論文 $R_a, R_b, R_c \in R$ は研究者 Q が論文 P を書くにあたり重要な論文である．研究者 Q は論文 O に加え，論文レコメンダシステムによってレコメンドされた論文 R を読み，論文 P を書く．このとき，Q によって新たに書かれる論文 P では，論文 O, R が引用される．これは論理式では次のように表される．

$$\text{Read } O \Rightarrow R \text{ is recommended} \Rightarrow \text{Cite } R \tag{3.2}$$

つまり，

$$\text{Read } O \Rightarrow \text{Cite } R \tag{3.3}$$

が成り立つ．このとき，論文 O は研究者 Q の研究において重要な論文だと仮定しているので，Read $O \coloneqq$ Cite O が成り立つ．よって (3.3) 式より，

$$\text{Cite } O \Rightarrow \text{Cite } R \tag{3.4}$$

が導かれる．

3.2.3　望ましくないケース

次に，研究者 Q は論文レコメンダシステムを使うが，レコメンドは失敗してしまうケースを考える (図 3.5).

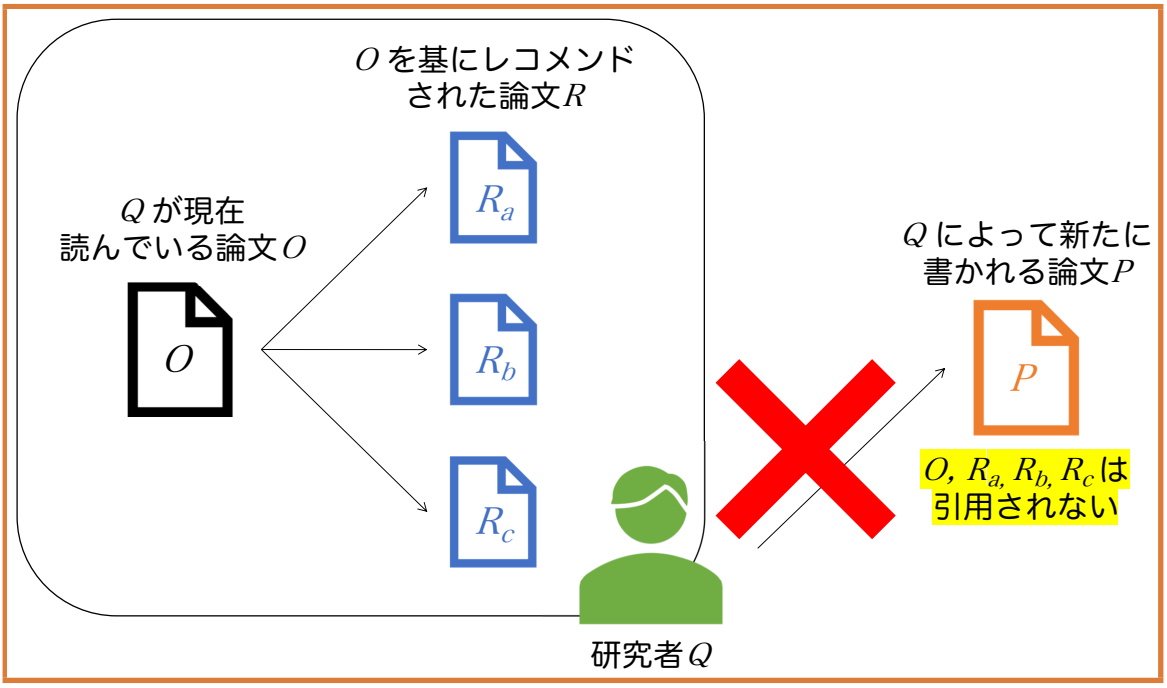

図 3.5: レコメンドが失敗するケース

レコメンドが失敗してしまうということは，研究者 Q はレコメンドされた論文 $R_a, R_b, R_c \in R$ を読むが，研究者 Q の研究に不必要な論文であったために，引用がされないという場合である．これは論理式では次のように表される．

$$\text{Read } O \Rightarrow R \text{ is recommended} \Rightarrow \neg\text{Cite } R \tag{3.5}$$

つまり，

$$\text{Read } O \Rightarrow \neg\text{Cite } R \tag{3.6}$$

が成り立つ．このとき，理想的なケースの場合と同様に，論文 O は研究者 Q の研究において重要な論文だと仮定しているので，Read $O \fallingdotseq$ Cite O が成り立つ．よって (3.6) 式より，

$$\text{Cite } O \Rightarrow \neg\text{Cite } R \tag{3.7}$$

が導かれる．

3.3　条件付き確率による定量化

3.3.1　理想的なケース

前節 3.2 では，論文レコメンダシステムの理想的なケースと望ましくないケースを想定し，理想的な論文レコメンダシステムを想定し，モデル化を行った．この 2 つのケースを定量的に評価するためにはどうすればよいのかを考えたい．

まず，(3.4) 式，Cite $O \Rightarrow$ Cite R が真である場合を考える．論文の全体集合を U としたときに，論文 O を引用している論文と論文 R を引用している論文の関係は図 3.6 のように表すことができる．

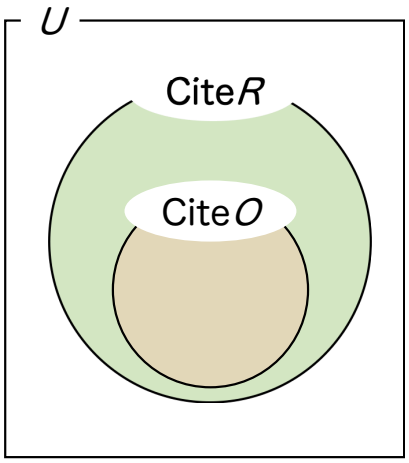

図 3.6: 論文 O を引用している論文と論文 R を引用している論文の関係 (理想的なケース)

しかし，図 3.6 は理想的なケースであり，必ずレコメンドが成功する場合を想定したときに成り立つ関係である．したがって，実際にはこののような関係にはならず，図 3.7 のようになる．

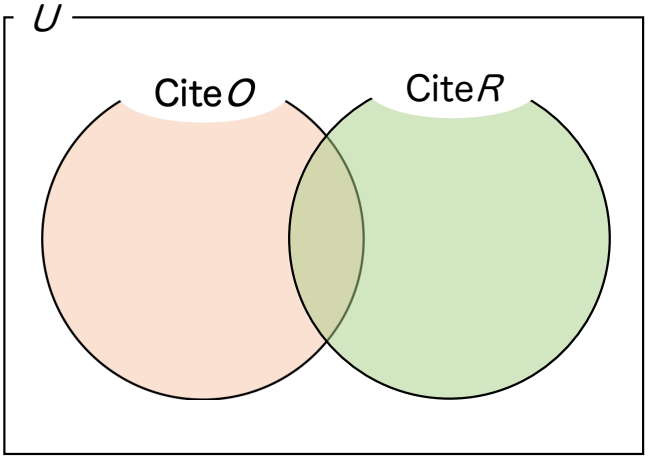

図 3.7: 論文 O を引用している論文と論文 R を引用している論文の関係 (実際のケース)

Cite O の集合が Cite R の集合から外に飛び出した形となっている．Cite O の飛び出している部分はレコメンドが失敗する確率が上がるほど，大きくなると考えられる．また逆に，Cite O の中の Cite R の割合は，レコメンドが成功する確率が上がるほど，大きくなると考えられる．そのため，この「Cite O が起こったときの Cite R の割合」を式として表したい．

一般に，事象 A が起こる確率は次の式によって表される．

$$Pr(A) \tag{3.8}$$

また，事象 B が起こるという条件のもとで，事象 A が起こる条件付き確率は次の式によって表される．

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)} \tag{3.9}$$

以上を踏まえると，図 3.7 の「Cite O が起こったときの Cite R の割合」は Cite O が起こるという条件のもとで，Cite R が起こる条件付き確率として考えられる．よって，

$$Pr(\text{Cite } R \mid \text{Cite } O) = \frac{Pr(\text{Cite } R \cap \text{Cite } O)}{Pr(\text{Cite } O)} \tag{3.10}$$

という数式で表される．また，(3.4) 式，Cite $O \Rightarrow$ Cite R の対偶を考え，

$$\neg\text{Cite } R \Rightarrow \neg\text{Cite } O \tag{3.11}$$

(3.11) 式より同様に条件付き確率を考え，

$$Pr(\neg\text{Cite } O \mid \neg\text{Cite } R) = \frac{Pr(\neg\text{Cite } O \cap \neg\text{Cite } R)}{Pr(\neg\text{Cite } R)} \tag{3.12}$$

が導かれる．(3.10) 式または (3.12) 式のどちらかを用いることで，論文レコメンダシステムの妥当性を評価することができると考えられる．ただし，ある特定の論文を考えたときに，それを引用する論文は論文全体の集合の一部であり，確率が非常に小さくなるために有意な結果が得られない可能性を考慮し，本研究では (3.12) を評価式として用いることとする．

3.3.2 望ましくないケース

次に，望ましくないケースの条件付き確率を考える．(3.7) 式，Cite $O \Rightarrow \neg$Cite R より，同様に条件付き確率を考え，

$$Pr(\neg\text{Cite } R \mid \text{Cite } O) = \frac{Pr(\neg\text{Cite } R \cap \text{Cite } O)}{Pr(\text{Cite } O)} \tag{3.13}$$

が導かれる．よって，(3.13) を用いて，論文レコメンダシステムがレコメンドに失敗する確率を求めることができると考えられる．

3.4 論文レコメンダシステムの構築

CCA を用いた論文レコメンデーション [10] では，DeepWalk と Doc2Vec による特徴量を正準相関分析によって組み合わせている．しかし，DeepWalk ではニューラルネットワークを用いており，高いパフォーマンスが得られる一方で，内部の実装ではニューラルネットワークを用いており，得られた結果の解釈が難しいという問題点がある．そこで，本研究では DeepWalk の代わりに GraRep[14] と呼ばれるグラフ埋め込みの手法を用いる．解析的な手法によってベクトル表現を得ることができるため，解釈性を担保することができると考えられる．Doc2Vec と GraRep によって得られた 2 つのベクトル情報から，正準相関分析によって組み合わせた特徴量ベクトルを作り，コサイン類似度を用いて論文のレコメンドを行う．

3.4.1 Doc2Vec

Doc2Vec[15] とは，Word2Vec を拡張することで文章のベクトル化を行えるようにした手法である．使用するニューラルネットワークの構成は Word2Vec と変わらないが，入力を文章 ID とし，文章中の単語を予測させることによって，文章 ID に対するベクトル表現を得ることができる．

3.4.2 GraRep

GraRep とは，最適化問題を特異値分解を用いて解析的に解くことによって，グラフ構造をベクトル化する手法である．表 3.1 に GraRep のアルゴリズムを示す．

GraRep では，まず最初にグラフ上の各ノードから移動したときの遷移行列を求める．遷移行列とは，各ノードからノードへの移動が行われる確率を表す行列である．次に，各ステップごとのベクトル表現を順に求めていく．ここでは，遷移行列から対数による確率行列に変換し，特異値分解によって得られた U^k, Σ^k から $U^k_d(\Sigma^k_d)^{\frac{1}{2}}$ を計算し，これを最終的なベクトル表現とする．最後に，各ステップのベクトルを結合することで，k 回移動したときのベクトル表現が得られる．GraRep を使う利点として，DeepWalk よりも良い

表 3.1: GraRep のアルゴリズム

Input

グラフの隣接行列 S

最大遷移ステップ数 K

対数シフトパラメータ β

各ステップにおけるベクトルの次元数 d

1. グラフ上で k 回移動したときの遷移行列 A^k を求める

$A = D^{-1}S$ を計算

(ただし，D は A の次数行列を表す)

$$D_{ij} = \begin{cases} \sum_p S_{ip}, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

$A^1, A^2, \ldots, A^K$ をそれぞれ計算

2. k ステップのベクトル表現を求める

For $k = 1$ to K

2.1 対数確率行列を求める

$\Gamma_1^k, \Gamma_2^k, \ldots, \Gamma_N^k$ ($\Gamma_j^k = \sum_p A_{p,j}^k$) をそれぞれ計算

$X_{i,j}^k = \log(\frac{A_{i,j}^k}{\Gamma_j^k}) - \log(\beta)$ を計算

X^k の負の要素を 0 に置き換える

2.2 特徴量ベクトル W^k を求める

$[U^k, \Sigma^k, (V^k)^T] = SVD(X^k)$

$W^k = U_d^k (\Sigma_d^k)^{\frac{1}{2}}$

End for

3. k ステップベクトルを連結する

$W = [W^1, W^2, \ldots, W^K]$

Output

グラフのベクトル表現 W

精度でラベル推定を行うことができる [16] ことや，ニューラルネットワークを使わず遷移行列から解析的に解くため，結果の解釈がしやすいということが挙げられる．

3.4.3 CCA

CCA(正準相関分析) とは，2 種類の観測データに共通して含まれる情報を取り出す手法 [17] である．同じ対象を観測した 2 つの多次元変数 $Q = \{q_1, q_2, q_3, \ldots, q_n\}^T \in \mathbb{R}^{nxd_q}$ と $V = \{v_1, v_2, v_3, \ldots, v_n\}^T \in \mathbb{R}^{nxd_v}$ があり，d_q, d_v をそれぞれ Q, V の次元数とする．このとき，Q, V を線形変換した値 QW_q と VW_v の相関係数が最大になるような W_q, W_v を解くことが CCA のタスクである．

$$(W_q, W_v) = \underset{(W_q, W_v)}{\operatorname{argmax}} \operatorname{corr}(QW_q, VW_v) \tag{3.14}$$

本手法では (3.14) によって得られた，W_q と W_v をそれぞれ Doc2Vec，グラフ埋め込みによる手法 (DeepWalk または GraRep) の線形変換された特徴量とし，

$$W = \alpha W_q + (1 - \alpha) W_v \tag{3.15}$$

を最終的な特徴量として使用する．

3.4.4 論文のレコメンド

ベクトル化した特徴量を元に，どの論文をレコメンドすべきか決定しなければならない．本手法では，論文間のコサイン類似度を求め，値が高いものからレコメンドされるようにする．

第4章　検証実験と考察

本章では，本研究における検証実験の結果と考察を述べる．

4.1　検証実験の目的

下記の 8 つの手法について，評価および比較をすることで，本レコメンデーション手法の有効性を明らかにする．また，本評価手法によって評価を行い，論文執筆モデルを基に考案した評価手法が妥当であるかを議論する．

1. Random
2. Doc2Vec
3. DeepWalk
4. CCA($\alpha = 0.1$, DeepWalk+Doc2Vec)
5. CCA($\alpha = 0.05$, DeepWalk+Doc2Vec)
6. GraRep
7. CCA($\alpha = 0.1$, GraRep+Doc2Vec)
8. CCA($\alpha = 0.05$, GraRep+Doc2Vec)

なお，Random という手法は全ての論文中からランダムにレコメンデーションを行うレコメンダシステムである．ランダムなレコメンデーションはレコメンドの結果に意味を持たないため，この結果と他の手法による結果とを比較することで，本評価手法の妥当性を測定する．

4.2 実装環境

本研究では，Python 3.7.5[1] と Jupyter Lab[2] を用いて実装および検証を行った．また，使用パッケージに関しては，以下の通りである (表 4.1)．

表 4.1: 実装に用いたパッケージ一覧

Algorithm	Package	Version
Doc2Vec	gensim [1]	3.8.1
DeepWalk	GitHub Repository [2]	e7bab0648c4ed13dfc5401c3558687d30da9fad1
	gensim [3]	3.8.1
GraRep	GitHub Repository [4]	aa4d1bf09af983a946aab55afabd3e37deec8c3b
CCA	scikit-learn [5]	0.22
Others	NumPy [6]	1.17.4
	NetworkX [7]	2.4

[1] https://radimrehurek.com/gensim/models/doc2vec.html
[2] https://github.com/phanein/deepwalk
[3] https://radimrehurek.com/gensim/models/word2vec.html
[4] https://github.com/benedekrozemberczki/GraRep
[5] https://scikit-learn.org/stable/modules/generated/sklearn.cross_decomposition.CCA.html
[6] https://numpy.org/
[7] https://networkx.github.io/

4.3 実行環境

プログラムの実行は，北陸先端科学技術大学院大学が所有する大型計算機[8] 上で，Docker[9] コンテナ内の Jupyter Lab を使用して行った．

4.4 使用データ

使用したデータに関しては，Arnetminer[18] の引用ネットワークデータセット V11[10] を使用した．表 4.2 は Arnetminer Citation Network Dataset V11 が持つフィールドの一覧表である．論文の識別には id，DeepWalk および GraRep による引用ネットワークか

[1]https://www.python.org/
[2]https://jupyter.org/
[8]https://www.jaist.ac.jp/iscenter/mpc/
[9]https://www.docker.com/
[10]https://www.aminer.org/citation

らのベクトル化には references フィールド，Doc2Vec のベクトル化には indexed_abstract フィールドをそれぞれ用いた．引用数が 10 未満の論文，参考文献リストまたはアブストラクトが欠損しているデータを取り除いた，248,471 本の論文を対象にレコメンダシステムの構築を行った．また，1 本の論文に対してレコメンドする論文の数は 10 本とした．

表 4.2: Arnetminer Citation Network Dataset V11 のフィールド一覧

フィールド名	フィールドタイプ	説明
id	string	paper ID
title	string	paper title
authors.name	string	author name
author.org	string	author affiliation
author.id	string	author ID
venue.id	string	paper venue ID
venue.raw	string	paper venue name
year	int	published year
keywords	list of strings	keywords
fos.name	string	paper fields of study
fos.w	float	fields of study weight
references	list of strings	paper references
n_citation	int	citation number
page_start	string	page start
page_end	string	page end
doc_type	string	paper type: journal, book title...
lang	string	detected language
publisher	string	publisher
volume	string	volume
issue	string	issue
issn	string	issn
isbn	string	isbn
doi	string	doi
pdf	string	pdf URL
url	list	external links
abstract	string	abstract
indexed_abstract	dict	indexed abstract

4.5 実験結果

本評価手法では引用されうる全ての論文の引用リストを調べる必要があるため，計算量が大きく，時間がかかる．そのため，サンプリングによる評価を行う．1 回につき 100 本の，論文 O と論文 R を引用可能な論文をサンプリングし，評価式を用いて計算を行う．これを 100 回繰り返し，得られた数値の平均値によって比較する．なお，以下の表中では，Doc2Vec，DeepWalk，GraRep をそれぞれ D2V，DW，GR として一部を省略している．

4.5.1 理想的なケース

表 4.3 に各手法による $Pr(\neg\text{Cite } O \mid \neg\text{Cite } R)$ の平均と標準偏差をそれぞれ示す．

表 4.3: 各手法による $Pr(\neg\text{Cite } O \mid \neg\text{Cite } R)$ の平均と標準偏差

Method	平均	標準偏差
Random	0.99999	0.00006
Doc2Vec	0.99519	0.00195
DeepWalk	0.99007	0.00370
CCA(0.1, DW+D2V)	0.98977	0.00369
CCA(0.05, DW+D2V)	0.99033	0.00382
GraRep	0.99012	0.00344
CCA(0.1, GR+D2V)	0.99407	0.00243
CCA(0.05, GR+D2V)	0.99384	0.00269

考察

$Pr(\neg\text{Cite } O \mid \neg\text{Cite } R)$ は理想的なケースを想定しているため，表はレコメンドが成功する確率を表している．そのため，1 に近ければ近いほど，レコメンドが成功していると考えられる．だが，Random の結果よりも，その他の手法の平均値が低くなってしまっている．この結果より，論文 O と論文 R の間には何らかの関係があるということが考えられる．図 4.1 は論文 O と論文 R の関係を表す図である．

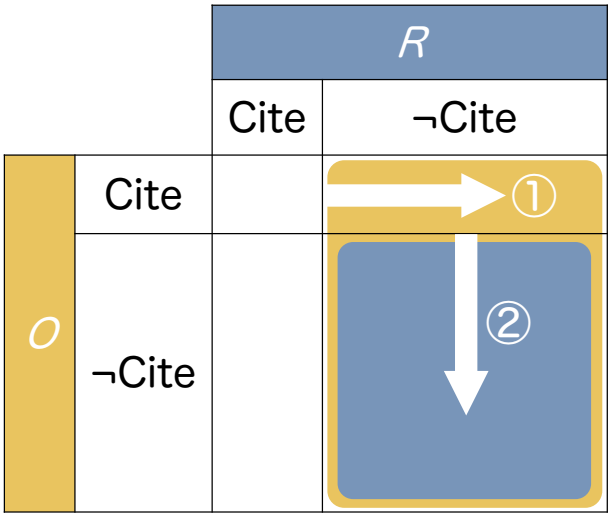

図 4.1: 論文 O と論文 R の関係

論文レコメンダシステムが行っているのは，$Pr(\text{Cite } R)$ を大きくすることである．図中の①がこれに当たる．もし，Cite O と Cite R が独立であった場合，論文レコメンダシステムによって $Pr(\text{Cite } R)$ が大きくなったとしても，$Pr(\text{Cite } O)$ には影響しないはずである．しかし，Random よりも，その他の手法の平均が小さくなっている．これは図中の②のように，論文 O が引用される確率が小さくなっており，Cite O と Cite R は独立ではないという事実を表している．この要因として，派生関係にある論文の引用は排他的に行われるということが考えられる．図 4.2 のように，派生の親となる論文 A と論文 A から派生した論文 B があるとする．ここで，論文 A と論文 B を引用可能な論文 P という論文を書くことを想定する．この場合，論文 A と論文 B は内容的に似ているため，双方の論文の差異に触れる必要がなければ，親となる論文 A を引用すれば十分な場合が多い．そのため，派生関係にある論文の引用は排他的となり，表 4.3 のような結果になったと考えられる．また，この引用の排他性により，本評価手法をはじめとする引用データを使用する評価手法では，内容の類似性が高い論文がレコメンドされているかどうかを評価することが難しい可能性がある．

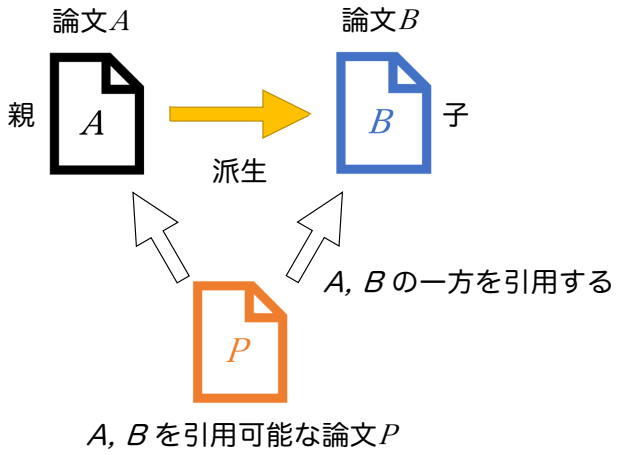

図 4.2: 派生関係にある 2 本の論文とそれらを引用可能な論文

4.5.2 望ましくないケース

表 4.4 に各手法による $Pr(\neg\text{Cite } R \mid \text{Cite } O)$ の平均と標準偏差をそれぞれ示す.

表 4.4: 各手法による $Pr(\neg\text{Cite } R \mid \text{Cite } O)$ の平均と標準偏差

Method	平均	標準偏差
Random	1.00000	0.00001
Doc2Vec	0.99463	0.00236
DeepWalk	0.99015	0.00374
CCA(0.1, DW+D2V)	0.99039	0.00340
CCA(0.05, DW+D2V)	0.99071	0.00414
GraRep	0.98921	0.00380
CCA(0.1, GR+D2V)	0.99353	0.00274
CCA(0.05, GR+D2V)	0.99385	0.00262

考察

$Pr(\neg\text{Cite } R \mid \text{Cite } O)$ は望ましくないケースを想定しており，表はレコメンドが失敗する確率を表していると考えられる．そのため，Random では，ほぼレコメンドが失敗しているということが読み取れる．また，DeepWalk，GraRep を単体で用いたときの確率の平均値が低いことから，グラフベース法を単体で用いたときはレコメンドが失敗しにくいということがわかる．このことから，グラフベース法を単体で使用したレコメンダシステムは，引用データのみを入力データとして用いているため，引用データを使用して評価を行う本評価手法においては，レコメンドが失敗しにくいという結果が出ていると考えられる．一方で，グラフベース法に内容ベー

スフィルタリングを組み合わせた手法によって，より内容的に近い論文がレコメンドされると，レコメンドの失敗確率が上がっている．つまり，節 4.5.1 において示唆された，派生関係にあり，内容的に似ている論文の引用は排他的に行われるという点を裏付ける結果が得られた．

第5章　まとめと今後の課題

5.1　まとめ

本研究では，論文レコメンダシステムがどうあるべきかということを出発点として，確率的に比較する評価手法を提案した．具体的には，論文レコメンダシステムを使うときの論文執筆のモデルを考え，理想的なケースと望ましくないケースを想定し，それぞれの場合についての評価式を考案した．また，CCA により GraRep と Doc2Vec の特徴量を組み合わせた論文のレコメンド手法を提案した．

本評価手法による定量的な評価を，8 つの論文レコメンデーション手法について行った結果，派生関係にある論文の引用は排他的に行われることが示唆された．また，引用ネットワークを使用したレコメンダシステムでは，このことを裏付ける結果が得られた．さらに，引用データを用いる現在のレコメンダシステムの評価方法では，派生関係にある親と子の論文のうち一方しかレコメンドすることができないという可能性が示唆された．

5.2　今後の課題

論文の継承関係などを利用し，文脈によって評価するポイントを変化させることが可能な評価手法を考案することが課題として挙げられる．大学生や大学院生などのビギナー研究者であれば，継承関係の親となる論文がレコメンドされたときに評価値が高くなり，また，一般の研究者であれば，最新の論文がレコメンドされたときに評価値が高くなる，といった評価方法を考案することで，目的に合わせて論文のレコメンダシステムを学習させることができると考えられる．

また，本評価手法では条件付き確率によって評価を行ったが，論文全体の集合に対して，ある 1 本の論文を引用する割合はとても小さい．そのため，確率の値が極端に大きくなる，または小さくなることで，比較したときの差異がわかりにくいという問題点がある．そこで，確率のとり方を変えるなどすることで，差異をわかりやすくする工夫が必要だと考える．

さらに，ある論文がどれだけの論文から引用しているのかを計算するためには，全ての論文の引用情報を調べる必要があるため，計算量が大きくなりやすい．実装を工夫す

ることで計算量を抑え，高速化を行うことで，論文の評価手法として利用しやすくする必要があると考えている．

参考文献

[1] 小泉周. 研究力の測り方. 学術の動向, Vol. 23, No. 12, pp. 12_64–12_67, 2018.

[2] 飯尾淳. 情報を集める技術・伝える技術: 情報社会の一員として備えておくべき基礎知識. 近代科学社 Digital, September 2019.

[3] Michael J. Pazzani and Daniel Billsus. Content-Based Recommendation Systems. In Peter Brusilovsky, Alfred Kobsa, and Wolfgang Nejdl, editors, *The Adaptive Web: Methods and Strategies of Web Personalization*, Lecture Notes in Computer Science, pp. 325–341. Springer, Berlin, Heidelberg, 2007.

[4] Daniel Billsus and Michael J. Pazzani. Learning Collaborative Information Filters. In *ICML*, 1998.

[5] Feng Xia, Haifeng Liu, Ivan Lee, and Longbing Cao. Scientific Article Recommendation: Exploiting Common Author Relations and Historical Preferences. *IEEE Transactions on Big Data*, Vol. 2, No. 2, pp. 101–112, June 2016.

[6] Quan Zhou, Xiuzhen Chen, and Changsong Chen. Authoritative Scholarly Paper Recommendation Based on Paper Communities. In *2014 IEEE 17th International Conference on Computational Science and Engineering*, pp. 1536–1540, December 2014.

[7] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia. Scientific Paper Recommendation: A Survey. *IEEE Access*, Vol. 7, pp. 9324–9339, 2019.

[8] Primož Godec. Graph Embeddings — The Summary. https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007, June 2019.

[9] 神嶌敏弘. 推薦システムのアルゴリズム. September 2016.

[10] Shashank Gupta and Vijay K. Varma. Scientific Article Recommendation by using Distributed Representations of Text and Graph. In *WWW*, 2017.

[11] Peng Cui, Xiao Wang, Jian Pei, and Wenwu Zhu. A Survey on Network Embedding. *arXiv:1711.08752 [cs]*, November 2017.

[12] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. DeepWalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '14*, pp. 701–710, New York, New York, USA, 2014. ACM Press.

[13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

[14] Shaosheng Cao, Wei Lu, and Qiongkai Xu. GraRep: Learning Graph Representations with Global Structural Information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management - CIKM '15*, pp. 891–900, Melbourne, Australia, 2015. ACM Press.

[15] Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. *arXiv:1405.4053 [cs]*, May 2014.

[16] 浅谷公威. ネットワークの表現学習. 人工知能学会誌, Vol. 31, No. 4, pp. 587–593, July 2016.

[17] 赤穂昭太郎. 正準相関分析入門. 日本神経回路学会誌, Vol. 20, No. 2, pp. 62–72, 2013.

[18] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. ArnetMiner: Extraction and mining of academic social networks. In *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*, p. 990, Las Vegas, Nevada, USA, 2008. ACM Press.

謝辞

本修士論文の作成ならびに研究にあたり，テーマ決定から研究手法，結果の解釈など，多くの助言，ご指導をいただきました北陸先端科学技術大学院大学の Dam Hieu Chi 准教授に心から感謝いたします．データ科学を学ぶのは大学院が初めてで知識のなかった私ですが，大学での講義をはじめ，データ科学の勉強会やゼミ活動にて手厚いご指導を頂きました．また研究室では，思うように研究が進まなかった私を何かと気にかけ，優しいお言葉をかけてくださり，精神面で救われた場面が何度もありました．なんとか研究が形になり，こうして論文を書くことができたのは，Dam 先生のご指導なくしては達成できませんでした．心より感謝の意を表すとともに，これまでの貴重な研究室での学びを今後に生かしていきたいと考えております．

また，サービスサイエンスの研究や学生生活の面で多くのサポートをいただいた，山梨県立大学の杉山歩准教授にも深く感謝しております．杉山先生には私の学部時代よりご指導いただいており，大学院に入ってからも就職活動や研究活動などさまざまな相談に乗っていただきました．サービスサイエンスの研究においても，論文の作成から学会発表までご指導いただき，とても貴重な経験を得ることができました．心より感謝いたします．

研究活動ならびに学生生活において，研究室の皆様や友人にも大変お世話になりました．良い研究室メンバー，友人に恵まれたことで，非常に充実した大学院生活を送ることができました．深く感謝いたします．

最後に，私の修士課程への進学に理解を示し，あらゆる面で学生生活を支援していただいた家族に心からの感謝を捧げます．