# 引用コンテキストに基づく<br>高影響論文の推定に関する研究

筑波大学
図書館情報メディア研究科
2019年3月
神澤 義人

# 目 次

# 図 目 次

# 表 目 次

# 第1章　はじめに

## 1.1　背景

学術論文は，過去から現在に至るまで継続的に研究の対象とされてきた．特に学術論文間の引用活動は、それらが文献間の関係を表現し，学術の発展の経緯や学術領域の俯瞰に繋がるとして，計量書誌学やネットワーク分析の主題として多く取り上げられてきた．例えば，研究成果である文献と，文献の著者についてのスコアを算出する手掛かりとして，引用活動は着目されている．ジャーナルインパクトファクター（JIF）や CiteScore のような評価指標は，文献間の引用被引用の関係数に基づいて算出される，文献を評価するための指標である．また，h 指数やそのフォロワーである h-index ファミリーと呼ばれる派生指数は，ある研究者が公刊した論文の被引用数に基づいて算出される，論文の著者に対するスコアである．一方，ネットワーク分析では，論文をノード，引用関係をノード間を結ぶエッジとして，引用関係ネットワークが構築される．これらの研究活動は，急速に増え続ける論文の発行本数と，多様化し続ける研究分野に対応し，学術領域について客観的に評価する．

本論文では，学術分野内部や社会に対して高い影響力を持った論文を推定する手法を提案することを目的としている．その際．単純な被引用数だけではなく，論文を構成する文章について，そのコンテキストに着目することで，引用の目的を考慮した上での推定を目指した．

## 1.2　高影響論文の定義

高影響論文という言葉には曖昧さがある．ここでは，影響の高さとは，そもそも何について言及しているものなのか概観する．その後，本論文における高影響論文の定義を述べる．

まず，社会的な影響が挙げられる．インターネットやデジタルリポジトリの発達により，論文への言及がニュースや SNS で積極的に行われるようになっている．研究発表の内容によっては，バズるという，Buzz を由来とした新語が表現するような，情報拡散が発生する．このようなインターネットでの言及に基づく社会的な影響の指針としてオルトメトリクスという指標が開発された．インターネットをベースとした情報拡散行動は，目に見える社会的な影響の大きさとして，その論文の評価軸として認知されつつある．

学術的な影響を測るには，先述の通り，他の文献から受ける引用活動が大きな手掛かりとなっている．一般的には，被引用数がその論文が与えた影響の大きさを知る手がかりとして用いられる．ただし単純な被引用数を基準にすると問題が発生することが知られている．被引用数は経年によって増加する可能性のある値で，これは減少することのない単調増加の性質をもつ．そのため，新しい論文と古い論文を比較するときに格差が生まれる．さらに，被引用数を基にした評価は，その論文が属する学術領域の活発さや人口や引用文化といった様々な要因に左右されるために，全ての論文に対して一律な価値を提供ものではない．

JIF を始めとする評価指標も，開発と改良が繰り返されており，その有効性については多く議論がなされている．加えて，JIF は論文が掲載されている学術雑誌について言及するもので，掲載されているそれぞれの論文の価値や影響力について直接言及するものではない．

論文が社会や学術領域に与える影響や潜在的な価値を知る方法は複数挙げられるが，その一つに，有識者やコミュニティが選出を行う，論文の表彰制度がある．

表彰制度は学術コミュニティによって千差万別であるが，大きく分けて 2 つに分類することができる．論文の発表と合わせて賞が決定するものと，過去の発表について賞が送られるものである．前者については，学術会議や学術雑誌などにおいて，その年の会議に投稿された論文から，優秀なものを委員らが選出，会議開催に合わせて表彰を行う制度である．受賞した論文は，Best Paper のように称される．論文そのもの以外にも，会議におけるキーノートやポスター，実演など，論文に付随する発表活動の優秀さを讃える制度を設けるコミュニティも存在する．

後者は，後年になってから当時発表された論文や著者を表彰する制度である．例えば，データベースに関する話題を中心として採録する Very Large Data Base(以下，VLDB) は，10 年前の VLDB Conference での発表から，特に影響が大きかった論文とその著者を取り上げて，表彰を行う VLDB 10-Year Awards を設けている．

ここで，本論文における高影響論文の定義をする．本論文における高影響論文とは，その論文の被引用活動や内容について，学術的社会的な影響を知る人物らによって，価値や影響を認められた論文を指すこととする．すなわち，本論文では学術コミュニティによる学会やワークショップ，学術雑誌等の場において，なんらかの賞与を受けた論文を，高影響論文と見做すこととする．

本研究では，主に上に述べた 10-Year Awards のように，論文が発表された後の影響を考慮した上で，評価を受けた論文を高影響論文として取り扱った．図 1.1 に，過去に発表された論文が，被引用活動と人手による選出によって受賞し，高影響論文とされる過程を示す．

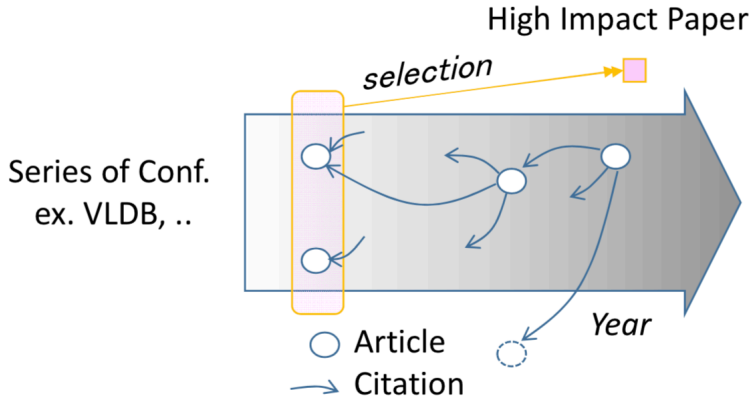


図 1.1: 高影響論文選出までの過程

## 1.3　研究の貢献

本研究の貢献は，高影響論文の推定について，実現への足がかりを提示することである．高影響論文の推定手法を確立することにより，学術分野における主要な論文を機械的に得ることが可能になる．これにより，研究者のサーベイ活動の補助し，対象とする学術分野の動向を俯瞰し，研究活動の評価指標の手掛かりを得ることに繋がっていく．更に，初学者や分野外の人間にとっては，関心を寄せる学術領域について知る足掛かりとして機能することが期待される．

本論文の第 1 章では，研究の背景と目的，言及対象の定義について述べた．第 2 章では関連研究を紹介し，続く第 3 章では，本論文の提案手法について説明する．第 4 章では，提案手法実現に向けて作成を試みた引用関係についての実験データについて説明する．以降，第 5 章では，実験方法の説明とその結果について記し，第 6 章で考察，最終章にてまとめとする．

# 第2章　関連研究

学術論文の引用関係に着目した研究はこれまで数多く取り組まれている．大槻らは，学術論文の引用ネットワークの分析において，各論文の被引用期間に着目し，重要度の算出方法について提案した [1]．また，ページランクアルゴリズムを引用ネットワークに適用することで，クラスタリングによって識別された各分野の主要論文の動的な抽出を試みた [2]．江藤は共引用関係を，本文中で引用される位置関係を基に分類し，それぞれの場合において被引用論文間の類似度の算出・比較を行った [3]．難波らは，論文間の参照情報を考慮したサーベイ論文作成支援システムの開発 [4] や，参照の理由を考慮して関連論文を組織化する手法の提案に取り組んだ [5]．論文が参照されている箇所の自動抽出を行い，提案する参照タイプを自動で識別する機能を開発した．この技術を基礎とし，参照構造を用いて論文間の類似度を測定，組織化する手法を開発した．学術分野を題材とした研究では，桂井らが時間変化による単語共起を検出することにより，研究トレンドの可視化を試みている [6]．情報推薦では，Hao らにより，あるニュース記事に対する他記事の自動推薦を目的として、コンテンツの類似性を見出すために暗黙的/明示的セマンティクスを活用する手法がそれぞれ提案された [7]．分布間距離を拡張した Word Mover's Distance を提案し，言語の多様性に対応した．計量書誌学の分野では，柴田らは，学術論文の引用分類について，過去の研究事例を取り上げつつ再定義を試みた [8]．引用活動の概念的な整理を行い，その結果，引用分類についての基礎的なスキーマの獲得したことを報告した．また，オントロジー研究では，Shotton により引用関係について表現が可能な CiTO が提案されている [9]．文献が持つ引用関係を機械可読状態にし，引用情報をマークアップした結果，文献自身が引用について表現可能とすることを目指している．

# 第3章　提案手法

本章では，本論文で扱う学術論文の概要を述べた後，提案手法について詳細に説明する．

本論文では，高影響論文の推定方法を提案する．高影響論文とは前述の通り，人手により選出された，受賞論文を指すこととする．

## 3.1　学術文献の構造

高影響論文を推定する手掛かりとして，本研究では学術論文が持つ文章構造と文献の引用活動の関係に着目した．学術論文は，著者らの研究の背景や成果，見解を整頓し，章立てによって構造化された専門性の高い文章である．章構造の作り方は基本的に論文著者が自由に行うことができる．ただし，論文発表は学術コミュニティに対する研究者らが行う情報発信手段の最たるものであるため，実際には，研究者らには研究意図と成果の報告を論理立てた構造で記述することが求められている．

そして，学術論文では，メインコンテンツと周辺研究の関連性の説明や手法の裏付け等を目的として，本文中で他の文献を引用する．引用によって生まれた記述は，巻末の References に対応した引用マークがつけられて，該当箇所を既存の知恵を用いて論文に組み入れたことがわかるようにされる．加えて，被引用文献が論文の文脈を無視して引用されることは，論理的な文章を構築する目的から外れてしまうため，例外的な事例であると仮定する．つまり，論文中に現れる引用は，確固たる著者の目的を前提として行われると考えられる．そこで筆者は論文の引用活動について，単に論文間を結ぶ線としてだけでなく，本文を基に分析する事により，文献の引用意図が可能ではないかと考えた．さらに，ある論文について，限定された範囲で複数の引用活動を調査することで，その論文が領域内で果たしている役割や貢献について，おおよそ推測することができるのではないかと仮説をたてた．

## 3.2　データ要件

## 3.3　Section の分類

論文のテキストは，Section という単位を使って構造化されている．前述の通り，論文の Section 構造は著者らが構築することが可能であるが，内容が伝わるような論理的な構造が求められる．そのため，論文の Section 構造は，特に同一の研究領域において極端に既存の構造パターンから外れる論文の数は少ないと考えられる．

論文の Section 構造を構築するにあたり，よく用いられるパターンに IMRAD がある．IMRAD は，最も知られている論文構造方式の一つである．IMRAD は，Introduction, Methods，Results and Discussion の頭文字が由来であり，これはそのまま学術論文に求め

られる情報構造と説明順番となっている．実際の論文ではこの他，Abstract や RelatedWorks，Conclusion，References のような Section が追加されて，文章が構造化される．

本研究では，論文を構成している Section に基づき引用構造の分類を行い，引用論文-被引用論文ペアについてデータセットを作成する．

データセットを作成するにあたり，データセット内の"References"を除く Section を集計し，以下の SectionTag を付与することで分類する．

1. A:Abstract
2. I:Introduction
3. R:RelatedWorks
4. F:FutureWorks
5. C:Conclusion
6. M:Methods

(1)-(5) のいずれにも該当しなかった Section については，論文の主張を特に担う Methods として分類をした．また，SubSection 以下は，Section が持つ目的から大きく離れることはないと仮定し，データセット上では上位層に吸収する．Abstract における引用の可否は，学術分野やコミュニティごとの引用作法により異なるが，今回実験データを作成するにあたり複数の論文で登場が確認されたため SectionTag の一つとして採用した．

IMRAD Format に則らず集計を行うのには，2 つの理由がある．第一に，引用が頻繁に行われる箇所は Introduction，RelatedWorks と想定することが可能であり，それぞれの Section での引用は目的が異なると考えられるため．第二に，多様な Section 名に対応した分類を，IMRAD における M，R，D について行うのは，今回は困難であると判断したためである．

またデータセット内の引用ペアは引用が行われた Section 単位でユニークである．例えばある文献について 1 論文中で Introduction と RelatedWorks でそれぞれ引用が行われていた場合，それぞれの Section の情報を持つ引用ペアを 1 つずつ作成する．

## 3.4 被引用期間

受賞論文と非受賞論文の特徴比較する要素として，被引用期間 (Cited Period) を導入する．被引用期間は論文が発表された年から受賞論文が決定するまでの期間，連続して引用を受けた最長の年数と定義する．この値を特徴量に加えることで，対象の論文がどのように引用を受けたのか，時間的な特徴を考慮することが可能となる．図 3.1 に，架空のある論文について年間の被引用数を棒グラフで示す．該当の論文は，2000 年から 2005 年にかけてと，2009 年に引用を受けている．この場合の被引用期間は，最長の連続被引用年数である 6 となる．

論文の被引用傾向を考慮した事例として，大槻ら [1] の研究が挙げられる．Otsuki は，文献の年間の被引用数について，最大被引用数の 10%以上の被引用数が継続した年数を Period として定義した．本研究では，扱う論文データの範囲を特定の学術コミュニティに

限定しており，年ごとの最大被引用数に限りがあるため，被引用期間の定義から割合を撤廃した．

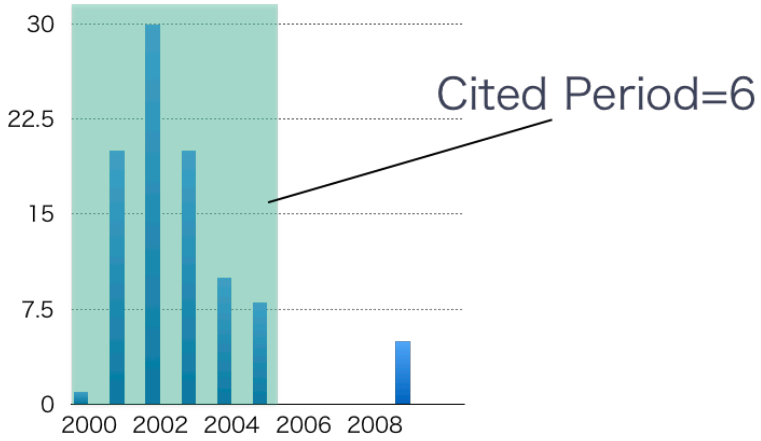


図 3.1: 被引用期間の例

## 3.5 高影響論文の推定

### 3.5.1 論文の特徴量

論文ペアデータセットに基づき，データセット内の各論文について特徴量を作成する．その論文がデータセット内の他の論文から，どの Section で引用を受けているのか，また発表後何年目に引用を受けているのかについてまとめる．論文の引用ペアは論文ペア１：１に対して，Section ごとに作成する．そのため，引用が行われた Section の数で分割した値を各章の特徴量について加える．Section タイプごとの変数は最終的に正規化する．これは，引用回数にとらわれず，引用を受けた Section タイプについて論文ごとの傾向を表現するためである．そして，対象の論文が発表された年を 0 として，最大 10 年分の年間被引用数を保存する．総被引用数は，論文の発行からデータセット内で行われた 10 年以内の被引用のみを合算した値とする．Awards の受賞対象となるまでの期間に限定することで，新旧論文間に経過年数による引用を受ける機会の格差を生まないように配慮し，年代別の比較を可能とするためである．

### 3.5.2 高影響論文の推定について

3.5 で作成した特徴量に基づいて，受賞論文と非受賞論文を混在するデータについて，二値分類問題を解く．受賞論文と近しい特徴量を持つ論文は受賞側として分類される．よって，過去の受賞実績に近い特徴を持つ，受賞には至らずとも十分な影響力を持つであろう論文を推定することに繋がる．今回は，二値分類問題の解法として良く用いられるロジスティック回帰分析，ランダムフォレストによる決定木（RF），多層パーセプトロン（MLP）の三手法について同様の実験を行い，5 分割交差検定を実施した結果の平均を比較する．

# 第4章　実験データ

本章では，今回の実験に使用したデータセット及び実験データの構築手法について，失敗事例を含め説明する．

## 4.1　データセットの構築

今回，高影響論文の分類問題に取り組むにあたり，分類のファクターとなりうる高影響論文を内包する論文の集合が求められた．要求される条件は次のとおりである:

- 論文を発行する学術コミュニティが取り扱う分野には方向性がありつつも，ある程度幅の広い領域の論文を受け入れていること
- 歴史が長く，過去論文のコミュニティ内部における引用の変遷について，観測が期待できること
- 人手により，発表後に与えた影響や功績を讃える制度が存在していること

そこで，今回は VLDB の国際会議や予稿集で発表された論文を収集し，データセットを作成した．VLDB は Conference で発表された論文の中から，特に高影響を与えたとされる論文とその著者を 10 年越しに表彰する"10-Year Awards" という制度を 1995 年から運用している．本論文では，10-Year Awards を受賞した論文を，学術分野に少なくない影響を与え，コミュニティにも認知されている高影響論文とみなした．

以下に，今回使用したデータセットの概要を記す．

今回は，VLDB のホームページから，1982 年から 2007 年の Conference（以下，VLDB Conf.）で発表された論文と，2008 年から 2017 年前期までに発表された Proceedings of VLDB という予稿集（以下，PVLDB）で発表された論文の PDF ファイルを収集した．PDF ファイルは pdftojson[1] を使用してテキストと文書構造の抽出を行った．その結果，十分なテキストを正確に抽出できたのは，2000 年以降の PDF だった．pdftojson で抽出したテキストは，章番号とタイトル以下に対応する本文を格納した本文ファイルと，論文タイトルと References 一覧をまとめたテキストファイルに変換した．論文タイトルと References 情報は，引用被引用の関係にある論文の引用ペアである．

## 4.2　引用ペアに対する Crossref による書誌情報適用の試み

論文の引用数や生のデータから引用ペアを集計する上で避けて通れない問題が，表記ゆれである．名前の省略や，記号，また論文ごとに References に記載する論文 1 本についての情報の有無や，誤字の想定など，考慮するべき情報は際限がなくなる．そこで今回は，

[1] pdftojson. https://github.com/ldenoue/pdftojson

表 4.1: 収集した VLDB 論文の PDF

| | Publications | Text Ext. | Award Year | Awards Paper |
|---|---|---|---|---|
| 2000 | 96 | 72 | 2010 | (1) |
| 2001 | 105 | 99 | 2011 | 1 |
| 2002 | 130 | 109 | 2012 | 1 |
| 2003 | 128 | 116 | 2013 | 0 |
| 2004 | 150 | 141 | 2014 | 1 |
| 2005 | 152 | 140 | 2015 | 1 |
| 2006 | 134 | 134 | 2016 | 1 |
| 2007 | 152 | 152 | 2017 | 1 |
| 2008 | 169 | 169 | 2018 | 1 |
| 2009 | 180 | 180 | - | - |
| 2010 | 186 | 186 | - | - |
| 2010-2011 | 177 | 177 | - | - |
| 2011-2012 | 222 | 222 | - | - |
| 2012-2013 | 237 | 237 | - | - |
| 2013-2014 | 240 | 240 | - | - |
| 2014-2015 | 244 | 244 | - | - |
| 2015-2016 | 190 | 190 | - | - |
| 2016-2017 | 220 | 220 | - | - |
| Total | 3,112 | 3,050 | - | 7 |

Crossref REST API[2] を用いて Crossref が保有する書誌情報を取得した．Crossref REST API は，Query に関係が高いと判別した順番に，Crossref が保有する書誌情報検索結果を json 形式で POST してくれる．今回は PDF から抽出した References 情報を Query として，Crossref から得られた結果を，引用ペアに追加した．Crossref REST API はクエリを解釈し近しい文献の書誌情報を提供するので，タイトルや著者名に多少の表記ゆれがあっても同一の書誌情報を得られる．そのため，Crossref REST API で取得した書誌情報を基本的な手がかりにすることによって，引用情報の集計を実現した．また，書誌情報には DOI が含まれており，同一の DOI を持つ引用ペアを統合することで，表記ゆれの問題の解決を図った．

今回はテキスト抽出に成功した 2000-2017 年前期までの期間に VLDBConf と PVLDB で発表された論文について，引用ペア情報を作成した．2000 年から 2017 年前期までの期間に VLDB から発表された刊行物の PDF3,112 本からは，3,050 件のテキストデータが抽出できた．しかし，2000 年の Awards Paper は，テキスト抽出に失敗したため，今回は引用情報の抽出対象から除外した．ただし，データセット中の他の論文から引用されるため，実験データ中に 2000 年の受賞論文を含む引用データは存在する．表 4.1 に以上のことをまとめる．

表 4.2: Crossref REST API の返り値と VLDB の PDF に基づくデータセット

| Base PDF | Pairs(Rows) | Pairs with Correct Crossref's Return (based on Published Year) | Pairs with ( Cited Year - Published Year ) <0 |
|---|---|---|---|
| 3,112 | 76,694 | 37,710 | 32,724 |
| Unique DOI | Cited DOI | Cited DOI from the DOI in Dataset | |
| 2,293 | 24,893 | 1,179 | |

### 4.2.1 適用結果の確認

引用ペアへの Crossref から得た書誌情報の適用結果について検証を行う．Crossref から取得した書誌情報の検証は特に重要である．なぜなら，実際の分析で取り扱う実験データを抽出する時に，複数の引用関係を統合する鍵は，Crossref から得られる DOI だからである．表 4.2 に今回作成した実験データについて示す．

今回のデータセットにおいて，Coference において 2000 年以降に発表されたものから，PVLDB vol.10 上で 2017 年前期までに発表された，合計 3,112 本の PDF を用いている．なぜ 1999 年以前の PDF を含めた全 4,412 件の PDF を使用しなかったかというと，pdftojson を用いて本文を取得できるデータが微々たるものであったためである．データセットは 2 つの論文の DOI だけでなく，Crossref に基づく発表年や論文のタイトルを保存している．加えて，検証を行うために，Crossref に与えたクエリ（PDF から取得した論文タイトルと，対応する References の 1 行）と，実際の発表媒体を記したカラムを用意している．その結果，データセット全体では 76,694 件の引用ペアが収集できた．

ここから，Crossref REST API で取得した書誌情報に基づいて，データセットを検証していく．データセットを作成するにあたり，Crossref REST API で取得した検索結果の上位 1 件を，各論文に書誌情報として適用している．データセットに対して適用した書誌情報は，上位 1 件の結果に含まれる文献のタイトル，発行年，DOI である．全ての適用結果を目視で確認することは困難なため，PDF から抽出したクエリを基に Crossref REST API で得られた発行年と，実際の論文の発表年を比較し，その一致について調査した．その結果，約 38,000 件の引用ペアについて，発表年の一致が認められた．ただし，同年に発表された，書誌情報が類似する論文が複数存在している場合には，異なる論文の書誌情報が適応されている可能性がある．同時に，被引用論文の発行年との差がマイナスになった引用ペアが，32,724 件確認できた．本来の References 情報とは一致しない書誌情報から取得した発表年を用いているために，このような不整合が発生していると考えられる．

また，データセットの基となった PDF は 3,112 本あるが，引用ペアのソース側に登場する DOI の件数は，2,293 件だった．この差は，PDF からの本文抽出の失敗や，Crossref から受け取ったデータに欠落が確認された引用ペアを排除したために，生まれたものである．実験データに組み込まれた 2,293 件の論文は，自身の本文中で複数の引用活動を行っている．それらを累計すると，計 24,893 件の DOI が引用されていた．2,293 件の論文の DOI のうち，1 度でも実験データ内の別の論文から引用を受けていたのは，1,179 件だった．

以上の通り，作成した実験データの状態について確認した．VLDB で発表された全ての論文が DOI を登録していることはなく，Crossref が書誌情報を保有していない論文も少な

[2]Crossref REST API. https://github.com/CrossRef/rest-api-doc

くなかった.

## 4.3 人手での実験データ構築

今回は実験を行うにあたり，受賞論文と非受賞論文の被引用特徴量を必要最低数作成した．特徴量の作成は人手で行った．実験で使用するのは，現在までに発表されている 1999 年から 2008 年の VLDB 10-Year Arard 受賞論文 9 本と，それぞれの受賞論文と同年に発表された非受賞論文合計 9 本である．非受賞論文の 9 本は，上の条件に加え，2019 年 1 月 22 日現在 Google Scholar[3] 上で確認できる範囲で，VLDB Conference あるいは PVLDB で発表された論文から，発表後 10 年以内に 1 度以上の引用を受けていることを条件に，著者が無作為に選出した.

表 4.3 に，今回被引用特徴量を作成した論文の一覧を掲載する.

対象の論文 18 本について被引用特徴量を作成する．被引用特徴量として，10 年間に VLDB Conference，PVLDB，VLDB Journal で発表された論文から受けた総被引用数（Cited Count．以下，CC），3.4 節で定義した被引用期間（Cited Period．以下，P），3.3 節で提案した SectionTag-[A,I,R,F,C,M] の割合を採用する．SectionTag の割合は，CC の値を用いて 0-1 の範囲で正規化されている．引用が行われている Section が，SectionTag-[A,I,R,F,C,M] のいずれに該当するかは，Section 名を元に著者が判断し，Section 名のみでの判断が困難である場合は，引用マーク周辺の文章を実際に確認し，文意を考慮した上でタグ付けを行った.

[3]Google Scholar https://scholar.google.co.jp/

表 4.3: 対象論文

| VLDB 10-Year Awards 受賞論文 | | |
|---|---|---|
| Awarded Year | Published Year | Paper Title |
| 2018 | 2008 | WebTables:exploring the power of tables on the web |
| 2017 | 2007 | Multi-Probe LSH: Efficient Indexing for High-Dimensional Similarity Search |
| 2016 | 2006 | The new Casper: Query processing for location services without compromising privacy |
| 2015 | 2005 | C-Store: A Column-oriented DBMS |
| 2014 | 2004 | Efficient Query Evaluation on Probabilistic Databases |
| 2012 | 2002 | Approximate Frequency Counts over Data Streams |
| 2011 | 2001 | Generic Schema Matching With Cupid |
| 2010 | 2000 | Don't Be Lazy, Be Consistent: Postgres-R, A New Way to Implement Database Replication |
| 2009 | 1999 | Database Architecture Optimized for the New Bottleneck: Memory Access |
| VLDB 10-Year Awards 非受賞論文 | | |
| | 2008 | Scalable query result caching for web applications |
| | 2007 | Graph Indexing: Tree + Delta >= Graph |
| | 2006 | Providing Resiliency to Load Variations in Distributed Stream Processing. |
| | 2005 | KLEE: A Framework for Distributed Top-k Query Algorithms. |
| | 2004 | Efficient Indexing Methods for Probabilistic Threshold Queries over Uncertain Data |
| | 2002 | StatStream: Statistical Monitoring of Thousands of Data Streams in Real Time. |
| | 2001 | Surfing Wavelets on Streams: One-Pass Summaries for Approximate Aggregate Queries. |
| | 2000 | Offering a precision-performance tradeoff for aggregation queries over replicated data |
| | 1999 | Cache Conscious Indexing for Decision-Support in Main Memory |

# 第5章　実験・結果

4.3 節で作成した被引用特徴量を用いて，VLDB 10-Year Awards の受賞論文と非受賞論文について分析実験を行う．今回は，ロジスティック回帰分析，ランダムフォレストによる決定木（RF），多層パーセプトロン（MLP）の三手法を用いて二値分類問題に取り組む．それぞれ 5 分割交差検定を以って実験データの学習と分類を試み，三手法の結果の平均を比較する．それぞれの手法の実装には，python の機械学習ライブラリである scikit-learn[10] を用いた．

一般的に分類問題では，その精度を表す指標として，Precision，Recall，F 値が参照される．Precision は，正解として分類したデータのうち，実際に正解であった割合を表し，Recall は，正解データのうち，実際に正解と判別された割合を表す．F 値は Precision と Recall の調和平均である．今回の分類問題においては，特に，Recall と同様の意味を持つ真陽性率（True Positive Rate．以下，PBR）と，偽陽性率（False Positive Rate．以下，FPR）を考察の材料として使用する．TPR は，正解データを正解データとして正しく分類される割合を示す値である．今回の実験においては，受賞論文を正しく受賞論文として認識できた精度を表している．FPR は，不正解データのうち，正解データとして誤って分類されてしまった割合を示す値である．今回の実験では，非受賞論文でありながらも，受賞論文であると認識されたデータの割合を表している．そのほか，正解データのうち不正解データとして分類されてしまった割合を偽陰性率（False Negative Rate．以下，FNR），不正解データを不正解データとして正しく分類される割合を真陰性率（True Negative Rate．以下，TNR）と呼称する．

## 5.1　二値分類結果：ロジスティック回帰分析

図 5.1 に，ロジスティック回帰分析の二値分類を，実験データについて 5 分割交差検定を行った平均値を正規化し，混同行列として示す．この混同行列は，左上に TPR，右上に FNR，左下に FPR，右下に TNR の値を示しており，高い値を持つほどに色濃く描画を行なっている．また，Accuracy と，正規化前の TP，FN，FP，TN の四値と，F 値について，それぞれの平均値と標準偏差を表 5.1 に示す．

## 5.2　二値分類結果：ランダムフォレスト

図 5.2 に，ロジスティック回帰分析の二値分類を，実験データについて 5 分割交差検定を行った平均値を正規化し，混同行列として示す．また，Accuracy と，正規化前の TP，FN，FP，TN の四値と，F 値について，それぞれの平均値と標準偏差を表 5.2 に示す．

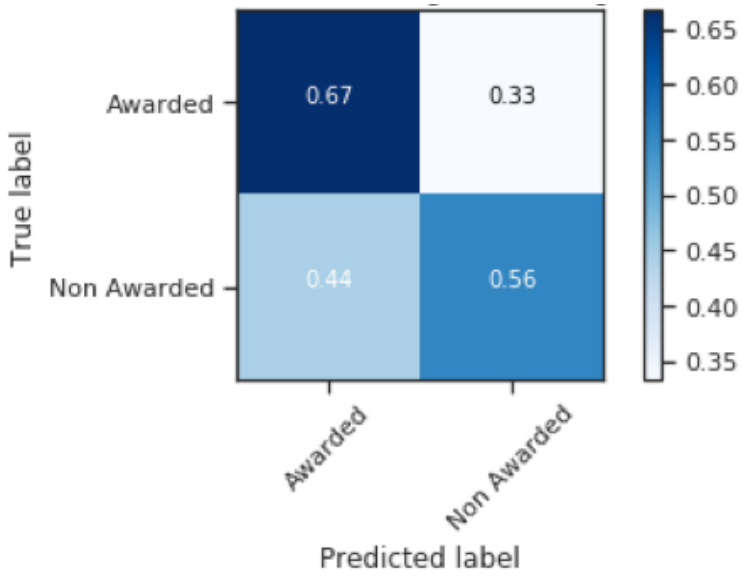


図 5.1: 混同行列： ロジスティック回帰分析

表 5.1: ロジスティック回帰分析結果

| parameter | 平均値 | 標準偏差 |
|---|---|---|
| Accuracy | 0.65 | +/-0.30 |
| TP | 1.20 | +/-0.75 |
| FN | 0.60 | +/-0.80 |
| FP | 0.80 | +/-0.75 |
| TN | 1.00 | +/-0.63 |
| F-measure | 0.63 | +/-0.37 |

## 5.3 二値分類結果： 多層パーセプトロン

図 5.3 に，ロジスティック回帰分析の二値分類を，実験データについて 5 分割交差検定を行った平均値を正規化し，混同行列として示す．また，Accuracy と，正規化前の TP，FN，FP，TN の四値と，F 値について，それぞれの平均値と標準偏差を表 5.3 に示す．

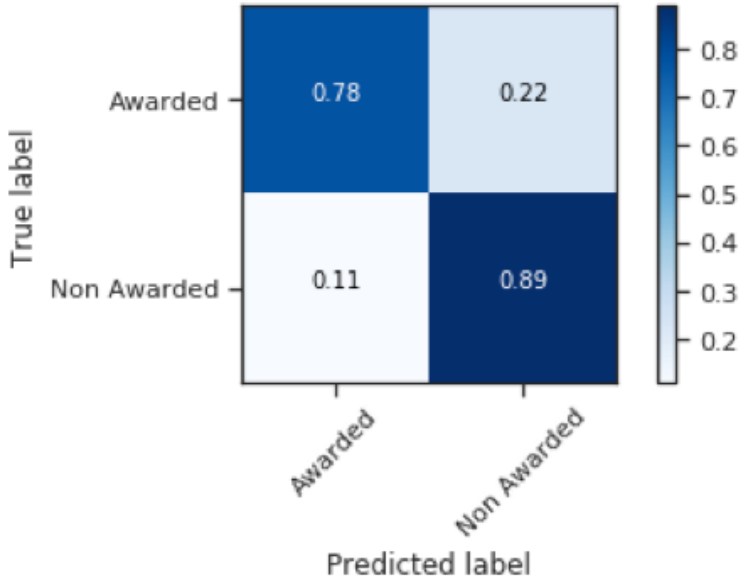


図 5.2: 混同行列： ランダムフォレスト

表 5.2: ランダムフォレスト結果

| parameter | 平均値 | 標準偏差 |
|---|---|---|
| Accuracy | 0.85 | +/-0.20 |
| TP | 1.40 | +/-0.80 |
| FN | 0.40 | +/-0.80 |
| FP | 0.20 | +/-0.40 |
| TN | 1.60 | +/-0.49 |
| F-measure | 0.76 | +/-0.39 |

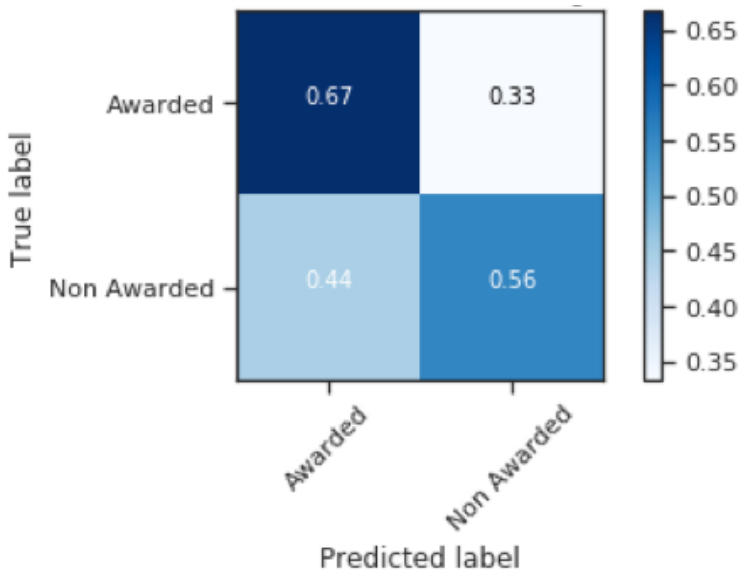


図 5.3: 混同行列： MLP

表 5.3: 多層パーセプトロン結果

| parameter | 平均値 | 標準偏差 |
|---|---|---|
| Accuracy | 0.65 | +/-0.30 |
| TP | 1.20 | +/-0.75 |
| FN | 0.60 | +/-0.80 |
| FP | 0.80 | +/-0.75 |
| TN | 1.00 | +/-0.63 |
| F-measure | 0.63 | +/-0.37 |

# 第6章　考察

## 6.1　PDF と DOI を用いた実験データの構築について

今回，実験データを構築するにあたり，正解データとなる受賞論文を含む論文集合と，それらの本文情報の獲得が必要とされた．論文の本文情報を直接提供しているケースは少ない．論文は PDF あるいは html での公開が一般的であり，研究対象としようとした場合，素直に本文情報を扱えるデータの数は限られている．また，PDF からテキストを抽出する過程においても，紙を読み込んで OCR した論文 PDF や，PostScript で作成された PDF から本文情報の抽出を試みた時，今回の試みでは思うような抽出を実現できなかった．引用関係と本文の両方について一から取り扱うことは，実際のところいくつもの壁が存在する．今回の場合は，10 年という期間の考慮を前提としていたのに対し，データ収集の段階で一定の時期から過去のデータが取り扱えないことが判明したために，ただでさえ少ないデータ数がより減少した．また，学術文献への DOI 付与状況が想定していたよりも完全でなかったことも，実験データの品質を下げる要因となった．

## 6.2　人手で作成した実験データの評価と考察

表 6.1，6.1 に，作成した実験データについて掲載する．

受賞論文と非受賞論文それぞれの特徴量の傾向について比較を行う．CC の平均値を比較すると，受賞論文 29.555556 に対して非受賞論文 11.222222 と，受賞論文という評価を裏付けるような高い値を獲得している．ただし，最小値や最大値，標準偏差を比較してみるとわかるように，非受賞論文は全体的に CC が低めであるのに対し，受賞論文は CC の標準偏差が 21.8581 と，その内訳に大きなばらつきがあることが確認できる．その一方で，被引用期間 P は，正解データ不正解データ共に値の高低差はあるものの，受賞論文の特徴量が持つ P の平均値が高く示された．

表 6.1: 受賞論文データについての特徴量

| | A | I | R | F | C | M | Cited Count | Cited Period |
|---|---|---|---|---|---|---|---|---|
| count | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| mean | 0.014466 | 0.347359 | 0.330769 | 0.002028 | 0.013221 | 0.292157 | 29.555556 | 7.555556 |
| std | 0.029644 | 0.168668 | 0.154063 | 0.004145 | 0.029653 | 0.147270 | 21.858128 | 2.920236 |
| min | 0.000000 | 0.148148 | 0.027778 | 0.000000 | 0.000000 | 0.083333 | 6 | 2 |
| 25% | 0.000000 | 0.211818 | 0.310909 | 0.000000 | 0.000000 | 0.211452 | 18 | 7 |
| 50% | 0.000000 | 0.299876 | 0.320219 | 0.000000 | 0.000000 | 0.284102 | 21 | 8 |
| 75% | 0.008197 | 0.450170 | 0.373577 | 0.000000 | 0.005952 | 0.386364 | 30 | 10 |
| max | 0.087500 | 0.666667 | 0.620741 | 0.011111 | 0.090909 | 0.516100 | 70 | 10 |

表 6.2: 非受賞論文データについての特徴量

| | A | I | R | F | C | M | Cited Count | Cited Period |
|---|---|---|---|---|---|---|---|---|
| count | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| mean | 0.000000 | 0.188463 | 0.506695 | 0.000000 | 0.000000 | 0.304841 | 11.222222 | 4.666667 |
| std | 0.000000 | 0.141640 | 0.115076 | 0.000000 | 0.000000 | 0.104398 | 6.887993 | 3.041381 |
| min | 0.000000 | 0.000000 | 0.341276 | 0.000000 | 0.000000 | 0.166667 | 2 | 1 |
| 25% | 0.000000 | 0.143939 | 0.444444 | 0.000000 | 0.000000 | 0.224443 | 6 | 3 |
| 50% | 0.000000 | 0.190476 | 0.492381 | 0.000000 | 0.000000 | 0.316952 | 11 | 5 |
| 75% | 0.000000 | 0.222222 | 0.600000 | 0.000000 | 0.000000 | 0.333333 | 14 | 6 |
| max | 0.000000 | 0.434281 | 0.666667 | 0.000000 | 0.000000 | 0.500000 | 25 | 10 |

SectionTag については，I と R に明確な特徴が現れている．I の値の平均は，受賞論文が被受賞論文を上回り，R の値は，被受賞論文が受賞論文を上回っている．これは，10-Year Award の受賞論文が同コミュニティ内の研究論文において，研究やアイデアの発端となるような役割を持つことが，他の論文に比べて多いのではないかと，考察することが可能である．同様に，非受賞論文は，同コミュニティ内の研究論文において，近しい領域の周辺研究として紹介される機会が多いのではないかと考えられる．

また，非受賞論文が A,F,C での被引用が確認されなかったことが示されている．これは，今回のデータにおいて，論文の概要及び今後の研究課題に関する項で，非受賞論文への言及が行われていないということである．受賞論文でもこれらの項について値を獲得した特徴量は少ないため，明確な言及は避けるが，受賞論文が他の研究において，研究の総括や今後の研究方針について影響を与えている可能性を考慮する材料にはなるだろう．加えて，受賞論文の特徴量は対象期間内において変動をしないが，非受賞論文のデータは無作為な選出のために，この値の変動や評価については，今後の分析により変動する可能性は大いに秘めている．

図 6.1 に，特徴量各値の相関について可視化したヒートマップを示す．白に近いほど相関があり，黒に近づけば相関がないというように読み取る．この図では，CC-P 間の相関と，それらと F との相関が特にはっきりと示されている．また，SectionTag 同士の相関については，F-R 間の相関が比較的はっきりと示されて，I と R，I と M の間については特に相関が見られないという結果だった．対象の論文について，Introduction で引用を行う場合は，研究の主題に特に影響をするような研究であり，本筋とは多少外れる周辺分野の研究について，言及が行われることがあると推察される RelatedWorks とは，引用目的が異なると考えられる．そのため，I と F の相関の無さについては，肌感として納得がいく結果である．

## 6.3 二値分類の結果比較

5 分割交差検定により得られたロジスティック回帰分析，ランダムフォレスト，多層パーセプトロンの三手法の分析結果について，比較・考察する．

今回の実験では，いずれの手法を用いても受賞論文と非受賞論文の完璧な分類を行うことができないことが確認された．受賞論文と非受賞論文の SectionTag の割合が完全に分離

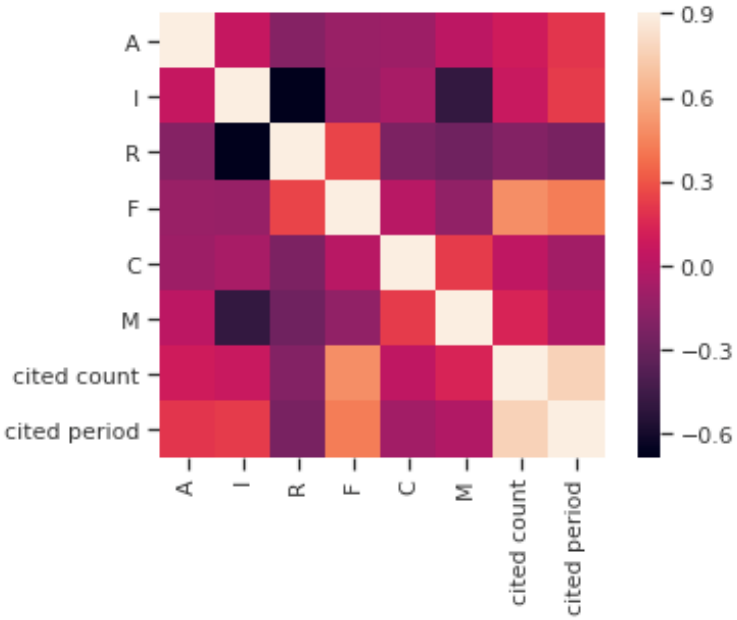


図 6.1: 特徴量の相関

した特徴量ではない上，被引用回数や被引用期間についてもデータによりばらつきがあった．そのため，いずれの手法においても False が発生した．いずれの実験結果について，TP，FP，FN，TN の標準偏差が+/-0.4 から+/-0.8 程度を示していることから，データ数の少なさと特徴量のばらつきが分析結果に少なくない影響を与えているものと考察する．ただ，いずれの場合でも，TPR と TNR の平均値が低いこともない．結果として，受賞論文と非受賞論文は，提案する特徴量を用いることで概ね分類が可能であると示唆された．また，ランダムフォレストにより，他の二手法よりも優れた Accuracy と TPR，F 値を獲得することができた．

このランダムフォレストについて，どの特徴量に重きを置いた分類が行われていたかを以下に示す．

1. I : 0.284008
2. R : 0.191464
3. CC : 0.162599
4. M : 0.114616
5. P : 0.085693
6. F : 0.083680
7. A : 0.052333
8. C : 0.025606

表 6.3: FP に分類された特徴量：RF

| title | A | I | R | F | C | M |
|---|---|---|---|---|---|---|
| Graph Indexing: Tree + Delta >= Graph | 0.0000 | 0.434281 | 0.341276 | 0.000000 | 0.000000 | 0.224443 |
| | cited count | cited period | | | | |
| | 13 | 6 | | | | |

この値は，5 分割交差検定を伴わずに同様の実験データで実施した二値分類の結果より獲得したものである．

特徴量の重要度が最も高いのは，SectionTag-I で，次いで R，CC と続く．単純な被引用数だけではなく，Introduction と RelatedWorks での引用が多くなされていることが，受賞論文と非受賞論文を分類するにあたり，特に重要なパラメータであることが示された．

ここで，本研究の目的である高影響論文の推定について，実験結果から言及する．上に示した通り，本論文で提案した特徴量を用いることで，高影響論文である受賞論文と，非受賞論文の二値分類を行うことが可能である．だが，分類問題に取り組む過程で，非受賞論文であるにも関わらず，受賞論文として誤った分類をされたデータがいくつか存在する．混同行列のうち FP に属するデータがこれにあたる．今回のように，高い TPR を記録した分類器は，受賞論文の特徴を正確に捉えた学習が行えたと考えられる．加えて，TNR も十分に高く，非受賞論文の判別も行えていることが確認できている．このような結果を前提にした時，FP に分類されたデータは，10-Year Awards のような賞に該当していなくとも，受賞論文に近しい構造で，引用を同コミュニティから受けている論文であると見なすことが可能ではないかと考察する．表 6.3 に，同様の実験データを用いた，5 分割交差検定を伴わないランダムフォレストによる分類結果で，FP と判別された特徴量を示す．

以上の解釈と考察により，非受賞論文の中から，過去のコミュニティから受けた引用構造によって，影響力が保証された論文を，高影響論文として推定する．

# 第7章　まとめ

本研究では，学術領域内の高影響論文を推定するため，本文中で引用活動が出現する Section に注目して，データセットの作成及び，コンテキストに着目して受賞論文と非受賞論文の分類を試みた．今回は人により選出された 10 Year-Awards を受賞した論文を，学術領域に高い影響を与えた論文とみなし，これらを含んだ論文集合を基として，データセットの構築を試みた．まず，DOI に基づき引用情報を付与してデータセットを構築した結果として，当初の目的の実験データを抽出するには力不足なデータセットが組み上がることとなった．これは，長期間の活動がある研究コミュニティでも，過去の論文から情報抽出が適切に行えないケースが確認できたことと，複数の引用ペアを累計するために獲得を試みた書誌情報が，DOI の未登録などの理由により，適切に付与されなかったことなどが，原因として考えられる．

次に，分類問題を実践するために，人手で受賞論文と非受賞論文についての特徴量を集計・作成し実験データとした．受賞論文と非受賞論文の特徴量は，データ数こそ限られているものの，二値分類を試みた結果，概ね分類可能であった．特にランダムフォレストを用いた二値分類で最も良い結果が得られ，Introduction と RelatedWorks での被引用が，分類の上で重要な要素となっていることが判明した．また，受賞論文の特徴を正確に学習し，非受賞論文を非受賞論文として分類可能な場合に，FP と判別された非受賞論文が，受賞論文に近しい影響力を持つ可能性を考察した．以上のことから，論文中の引用の構造を用いた引用の分類を考慮した被引用特徴から，高影響論文を推定することが可能である．

今回は 10 年間全ての期間を特徴量として採用したが，発表後 9 年間，8 年間，と期間を限定した特徴量を作成し，同様の結果を得られた場合，未来の受賞論文の推定に繋がる可能性をここで言及する．

また今後は，データセットの拡大を検討する必要がある．VLDB に限っては，1999 年以前の論文の特徴量を作成する際に，当時の PDF から本文情報を抽出する課題を解決しなければならない．数の限られた受賞論文については人手で特徴量を正確に集計し，他の非受賞論文は，エラー処理を工夫しつつ抽出精度の高い特徴量を機械的に作成することが想定される．その他，10-Year Awards のような制度を持つ学術コミュニティは VLDB のみではないため，複数の学術コミュニティで検証を行い，本論文の提案手法の普遍性について，検証していくことが求められる．

# 謝辞

本論文の執筆に辿り着くまでに，多くの障害と挫折がありました．最後までご指導くださった佐藤哲司教授と，共に研究課題について議論し，叱咤激励くださった研究プロジェクトに属する先達の皆様に感謝申し上げます．きっと，私は，私の知る以上に多くの人に支えられていたのだと思います．人付き合いが悪く口下手な私を根気よく相手し，同じ時間空間を共有してくれた佐藤・関研究室の皆さんに感謝申し上げます．段取りの悪い私を時に窘めつつも，朗らかに学生生活を共にしてくれた友人達，特に三年次編入の同期には，ここでは挙げきれないほどに感謝をしています．そして見守り続けてくれた家族に感謝申し上げます．

# 参考文献


[1] 大槻明, 川上あゆみ, 林剛, 川村雅義. 引用論文の分散値を重み付けとして考慮したページランクアルゴリズムによる主要論文の抽出. 情報知識学会誌, Vol. 21, No. 2, pp. 213–219, 2011.

[2] Akira Otsuki. Dynamic extraction of key paper from the cluster using variance values of cited literature. *CoRR*, Vol. abs/1310.4904, , 2013.

[3] Masaki Eto. A new co-citation measure based on structures of citing papers. *IPSJ Transactions on Databases*, Vol. 49, No. SIG7(TOD37), pp. 1–15, mar 2008.

[4] 難波英嗣, 奥村学. 論文間の参照情報を考慮したサーベイ論文作成支援システムの開発. 自然言語処理, Vol. 6, No. 5, pp. 43–62, 1999.

[5] 難波英嗣, 神門典子, 奥村学. 論文間の参照情報を考慮した関連論文の組織化. 情報処理学会論文誌, Vol. 42, No. 11, pp. 2640–2649, nov 2001.

[6] 桂井麻里衣, 小野峻佑. 語の共起のバースト検出に基づく研究トレンドの可視化. pp. G7–2, 2017.

[7] Hao Peng, Jing Liu, and Chin-Yew Lin. News citation recommendation with implicit and explicit semantics. In *ACL*, 2016.

[8] 柴田大輔, 芳鐘冬樹. 学術文献における引用分類の観点. 情報知識学会誌, Vol. 26, No. 3, pp. 277–296, 2016.

[9] David Shotton. Cito, the citation typing ontology. *Journal of Biomedical Semantics*, Vol. 1, No. 1, p. S6, Jun 2010.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.