Vol.2015-DBS-162 No.23
2015/11/26

手掛かり語による学術論文の引用意図分類の一手法

吉次 優[1,a)]　太田 学[1,b)]　高須 淳宏[2,c)]

概要：学術論文では一般に研究の根拠や考え方などが書かれた論文を参照しており，それらの被引用論文を読むことで，元の論文をより深く理解することができる．しかし，被引用論文は複数あることが多く，それら全てを閲覧，確認することは困難である．そこで，その支援を目的として，手掛かり語を用いて閲覧論文中の引用箇所（引用を表す文字列）の引用意図を推定し，被引用論文中の適切な被引用箇所（閲覧論文に引用されるべき箇所）を特定する手法が提案されている．しかし，引用意図の分類精度が十分でなかったため，本研究では引用意図分類精度の向上を図った．

キーワード：手掛かり語，学術論文，引用意図

A Citation Intention Classification Method for Academic Papers Using Clue Words

Yu Yoshitsugu[1,a)]　Manabu Ohta[1,b)]　Atsuhiro Takasu[2,c)]

Keywords: clue words, academic papers, citation intention

1. はじめに

学術論文では多くの場合，その研究の根拠や用いた手法に関する情報などが書かれた論文を引用する．そのため，被引用論文を読むことで閲覧中の論文をより深く理解することができる．しかし，被引用論文は 1 つの論文に対して複数あることが多く，それら全てを引用意図に応じて閲覧するのは閲覧者にとって大きな負担である．

この支援に関する研究として，被引用論文から適切な被引用箇所を自動で特定し，論文閲覧者に提示する手法が提案されている．石井ら [1] の方法では，まず閲覧論文中から引用箇所を特定し，人手で収集した手掛かり語を用いてその引用意図を分類する．次に，分類した引用意図に適すると考えられる被引用論文中の節を限定する．限定した節の各文に含まれる語と，引用箇所に含まれる語や人手で収集した手掛かり語とを照合し，一致した語数の多かった文とその周辺を適切な被引用箇所として提示している．本研究では，この石井らの提案した引用箇所の引用意図分類について検討する．石井らの引用意図分類の問題点として，手掛かり語を人手で収集しているため汎用性が低いことや，引用意図分類の精度が十分でなく，被引用箇所探索時の節の限定など後の手順に対する影響が大きいことが挙げられる．そこで，引用箇所の引用意図分類のための手掛かり語の収集と選出方法を新たに提案し，分類精度の向上を図る．

以下に本稿の構成を示す．2 節で論文の分類に関する関連研究について紹介し，3 節で先行研究である石井らの手法について概要を述べる．4 節では本研究で提案する引用箇所の分類手法について説明し，5 節で引用箇所の引用意図分類実験によって提案手法の評価を行い，6 節で本研究のまとめと今後の課題について述べる．

[1] 岡山大学大学院自然科学研究科
Graduate School of Natural Science and Technology, Okayama University
[2] 国立情報学研究所
National Institute of Informatics
[a)] yoshitsugu@de.cs.okayama-u.ac.jp
[b)] ohta@de.cs.okayama-u.ac.jp
[c)] takasu@nii.ac.jp

2. 関連研究

島ら [2] は，研究内容に着目した論文の自動分類方法について提案した．研究分野ごとに分類されることの多い論文を，研究内容によってより詳細に分類することで，論文検索の効率化を図った．論文全体の構成要素を，研究の目的や歴史，関連研究などの「序論部」，研究目的を実現するための方法や研究に用いる用語の定義などの「提案手法」，評価実験の方法や実験データ，実験結果などの「実験」，実験結果の分析や提案手法の問題点，研究のまとめや今後の課題などの「結論部」の 4 種類と定義した．論文の各節をこの 4 種類に対応付け，「提案手法」に当たる部分を「研究内容に関する記述」とみなす．この「研究内容に関する記述」に当たる節の文から形態素解析器 MeCab[3] を用いて名詞を抽出し，特徴量として用いた．

榊ら [4] は，論文ネットワークとカテゴリネットワークを構築し，これら 2 つを組み合わせた制約付きネットワーク上で論文をクラスタリングした．アブストラクトの文書ベクトルのコサイン類似度を各論文間の類似度とした論文ネットワークと，人手で論文のカテゴリ分類を行い，同じカテゴリの論文同士を結んだカテゴリネットワークをそれぞれ構築する．両者を組み合わせることで，近い類似度を持つ論文同士が同じカテゴリに属しているかどうか，という制約を付与された制約付きネットワークが構築され，このネットワーク上でクラスタリングすることで，結果が類似度とカテゴリの両方の影響を受け，良い精度が得られると考えた．論文ネットワークまたはカテゴリネットワークの片方のみによるクラスタリングと比較して，制約付きクラスタリングはそれらの分類精度を上回った．

難波ら [5]，[6] は，論文中で参照されている被参照論文の参照理由を考慮することで，論文間の類似度を測り，関連論文を組織化する研究を行った．論文が他の論文を参照する際の目的を cue word（照応詞，接続詞等）を用いて解析し，参照・被参照関係に参照タイプというリンク属性を付与した．ここで，参照タイプとは参照の理由についての属性で，既存研究の理論や手法を用いて新規の理論を提唱する場合などの「論説根拠型」，既存研究の問題点の指摘，関連研究との比較などの「問題点指摘型」，これら 2 つに当てはまらない「その他型」の 3 種類が定義された．

3. 石井らの引用意図分類

先行研究である石井らの研究 [1] における引用意図分類について述べる．石井らは引用意図のクラスを 6 種類定義し，各クラスに含まれる手掛かり語の出現回数などを考慮して引用箇所を分類する手法を提案した．

3.1 引用意図の分類クラス

石井らは，引用意図の分類クラスを “Group”，“Method”，“Result”，“Data”，“Equation”，“Other” の 6 種類と定義した．これらは NTCIR-9(the 9th NII Test Collection for IR Systems) [7] の論文を分析して定められたもので，詳細は以下の通りである．

- Group
  タスクやフォーラム，ワークショップの詳細を引用したい場合
- Method
  著者が研究で使用する手法，または著者の手法の比較対象となる既存手法を引用したい場合
- Result
  既存研究の実験結果を引用したい場合
- Data
  既存研究で用いられた実験データなどを引用したい場合（ただし，論文中に使用したデータの件数が記述されているもの）
- Equation
  計算式の詳細を引用したい場合（ただし，計算式そのものが引用されているもの）
- Other
  上記 5 つのいずれのクラスにも当てはまらない場合

3.2 石井らの分類方法

NTCIR-9 の論文から人手で収集した手掛かり語を用いて，以下のような手順で引用箇所を分類する．ここで，引用箇所とは論文中で引用を表す文字列を持つ 1 文と定義されている．また，表 1 は石井らが用いた手掛かり語と引用意図のクラスごとの手掛かり語の件数をまとめたものである．

① 分類対象の引用箇所に，Equation の手掛かり語が 1 つ以上含まれていれば Equation に分類する．
② ① の条件を満たさず，かつ Data の手掛かり語が 1 つ以上含まれていれば Data に分類する．
③ ① と ② の両方の条件を満たさず，かつ Group，Method，Result のいずれかの手掛かり語が 1 つ以上含まれていれば，最も手掛かり語の数が多いクラスに分類する．
④ ①〜③ の条件を全て満たさなかった場合，Other に分類する．

この分類方法は手掛かり語の数や各引用意図の出現回数を考慮して定められている．Data や Equation のように比較的出現頻度の低いクラスは手掛かり語が含まれているかどうかで判断している．また，Group，Method，Result のように比較的出現頻度の高いクラスでは手掛かり語が多く，他の引用意図を持つ引用箇所の文中にも出現することがある．そのため，最も手掛かり語の多いクラスに分類する．

表 1　石井らが用いた手掛かり語
Table 1　Clue words Ishii et al. defined

引用意図	手掛かり語	件数 [件]
Group	campaign, challenge, forum, goal, participated, task, track, work, workshop	9
Method	adopt, algorithm, applied, apply, approach, approached, based, category, characterized, classify, clustering, detail, details, develop, divided, employ, employed, engine, error, evaluate, experiment, explored, extract, feature, focused, function, group, idea, improve, method, model, modeling, including, label, mechanism, obtained, performed, platform, predict, program, propose, proposed, provided, re-rank, role, search, suggestion, system, technique, toolkit, trained, use, using, way	54
Result	achieve, analysis, best, effect, effective, participants, performance prove, reason, reported, research, result, show, waste	14
Data	collection, library	2
Equation	calculate, formula	2

3.3　石井らの分類方法の問題点

石井らの分類方法の問題点は，手掛かり語を人手で収集している点や，分類方法の手順が表 1 の手掛かり語の件数に依存している点である．この方法では分類対象の論文が変わった場合の対応が難しいため，極力人為的な作業を減らす方が望ましい．また，石井らの方法による引用意図分類の実験結果は十分な精度とは言えなかった．図 1 に石井らの手法の分類結果を示すが，分類クラスによって精度はまちまちで，F 値の平均は 0.6 を下回っている．

4.　手掛かり語の抽出と引用意図分類方法

3.1 節で説明した引用意図のクラスのうち，Other 以外の 5 クラスについて，手掛かり語の選出方法及び選出した手掛かり語を用いた引用意図の分類方法について述べる．

4.1　分類に用いる手掛かり語

NTCIR-9 に投稿された論文のうち，SpokenDoc タスクの論文 9 件から 66 文，GeoTime タスクの論文 9 件から 71 文，INTENT タスクの論文 12 件から 99 文の合計 236 文の引用箇所を抽出した．これらは，石井らが実験に用いたデータから，Other に属する引用箇所を除いたものである．具体的には，まず pdftotext[8] を用いて各論文 PDF ファイルを TXT ファイルに変換する．この TXT ファイルから「[1]」や「[2, 7, 12]」のような表現（以下，引用表現と呼ぶ）を含む文を引用箇所として抽出する．引用表現の定義は“$[[0-9]+([,|-]+[0-9]+)*]$”とした．それぞれの引用箇所を，英文形態素解析器 TreeTagger[9] を用いて形態素解析し，名詞，動詞，形容詞を手掛かり語の候補として抽出する．候補となった単語の出現回数を分類クラスごとにカウントし，比較のため以下の 5 種類の方法で手掛かり語を選出する．

(1) 1 クラスにしか現れない単語をそのクラスの手掛かり語とする．
(2) 引用箇所に複数回出現している単語のうち，全出現回数の過半数を 1 クラスが占めている場合，その単語をそのクラスの手掛かり語とする．
(3) 引用箇所に複数回出現している単語のうち，出現回数が最大のクラスが 1 つに定まる場合，その単語をそのクラスの手掛かり語とする．
(4) 全出現回数の過半数を 1 クラスが占めている場合，その単語をそのクラスの手掛かり語とする．
(5) 出現回数が最大のクラスが 1 つの場合，その単語をそのクラスの手掛かり語とする．

また，表 2 に上記の各方法で選出したクラスごとの手掛かり語の件数をまとめる．表 2 より，Method の手掛かり語は多く，Data と Equation の手掛かり語は少ない．これは，引用箇所に含まれる語から手掛かり語を選出しているため，各引用意図に属する引用箇所の数が多ければ手掛かり語も多くなる．

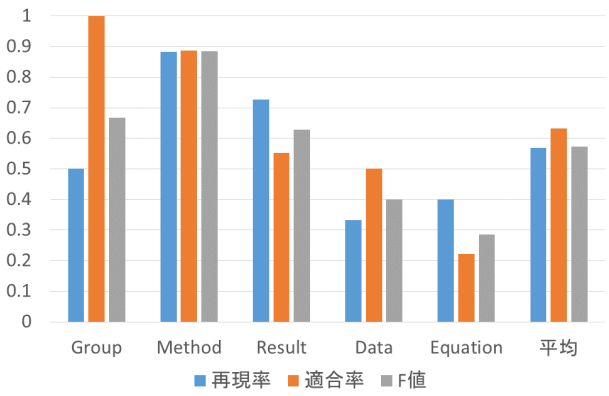

図 1　石井らの手法による引用意図分類結果
Fig. 1　Result of citation intention classification by Ishii's method

4.2　引用意図分類方法

4.1 節の方法で収集した手掛かり語を用いて，引用箇所の引用意図を分類する．引用箇所に含まれる各単語と表 2 の手掛かり語を比較し，一致した数に基づいて引用箇所の

表 2　各方法で選出した手掛かり語の件数
Table 2　Number of clue words obtained by each method

	手掛かり語の選出方法				
	(1)	(2)	(3)	(4)	(5)
Group	78	13	14	86	87
Method	823	417	422	965	970
Result	67	22	24	80	82
Data	7	8	8	8	8
Equation	8	2	2	9	9
合計	983	462	470	1,148	1,156

引用意図を定める．引用意図ごとに手掛かり語の数に差があるため，その一致した数をクラスごとの手掛かり語の異なり語数で割り，その値（分類スコア）が最大となるクラスに分類する．

例えば，NTCIR-9 の論文中に“In [2] we can see the task mentioned in more detail.”という引用箇所がある．表 1 の手掛かり語を例に説明すると，Group の手掛かり語である“task”，Method の手掛かり語である“detail”がこの文に含まれる．手掛かり語の出現回数はそれぞれ 1 回ずつのため，これらをクラスごとの手掛かり語の異なり語数で割ると（Group，Method）＝（1/9，1/54）となり，この引用箇所は Group に分類される．ただし，実際の分類は，表 1 に示した石井らの手掛かり語ではなく，表 2 にまとめた提案手法で収集した手掛かり語を用いて行う．

5.　引用意図の分類実験

4.2 節の方法で引用意図の分類実験を行った．4.1 節で示した手掛かり語の抽出に用いた NTCIR-9 の論文中の 236 文の引用箇所の引用意図を分類した結果の正解率を図 2 に示す．また，正解率の定義を式 (1) に表す．

$$\text{正解率} = \frac{\text{正しい引用意図に分類された引用箇所の数}}{\text{引用箇所の総数}} \quad (1)$$

ただし，図 2 の「石井らの分類方法」は，3 節で説明した石井らの方法による分類結果である．

図 2 に示された通り，手掛かり語の選出方法 5 種類のうち (1)，(4)，(5) の 3 種類で石井らの結果を上回った．

次に，引用意図のクラスごとの分類結果（再現率，適合率，F 値）を図 3〜図 7 に示す．

図 3 の Group では，石井らの分類方法では再現率が低く，適合率が高いのに対し，選出方法 (1)〜(5) では再現率が高く，適合率が低かった．これは，石井らの用いた Group の手掛かり語が，他の引用意図の手掛かり語に比べて特徴的なもの，つまり他の引用意図を持つ引用箇所の文にあまり出現しない単語が多かったことが理由の 1 つと考えられる．選出方法 (1)〜(5) で用いた手掛かり語は，石井らの 9 件に比べて件数が多くなり，引用意図 Group との関連が必ずしも大きくない語も含まれているため適合率が下がった．

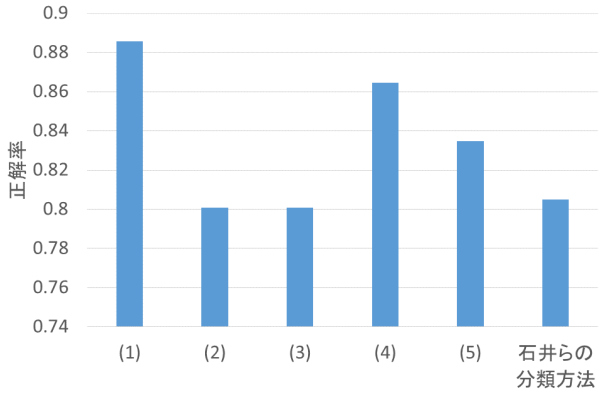

図 2　引用意図分類の正解率
Fig. 2　Accuracy of citation intention classification

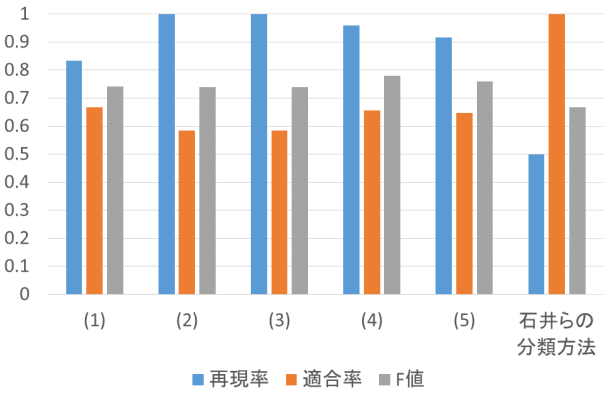

図 3　Group の引用意図分類結果
Fig. 3　Result of citation intention classification of “Group”

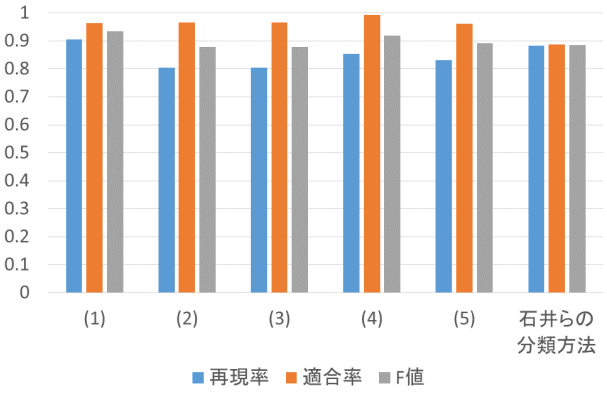

図 4　Method の引用意図分類結果
Fig. 4　Result of citation intention classification of “Method”

図 4 では，選出方法 (1)〜(5) のいずれの方法においても Method の F 値が高い．Method は 5 種類の引用意図の中でも最も引用箇所の件数が多く，手掛かり語も多かった．これにより分類スコア算出が正しく行われやすく，精度が

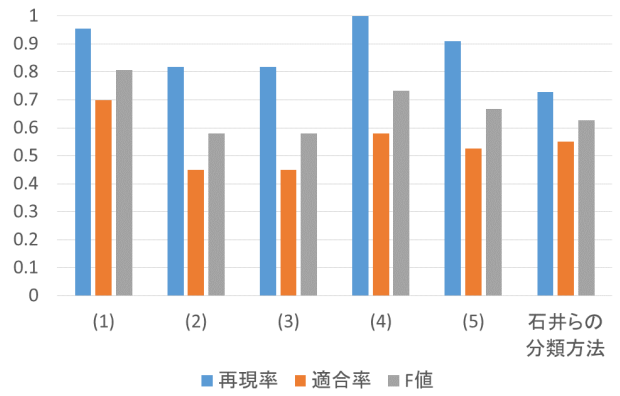

図 5　Result の引用意図分類結果
Fig. 5　Result of citation intention classification of “Result”

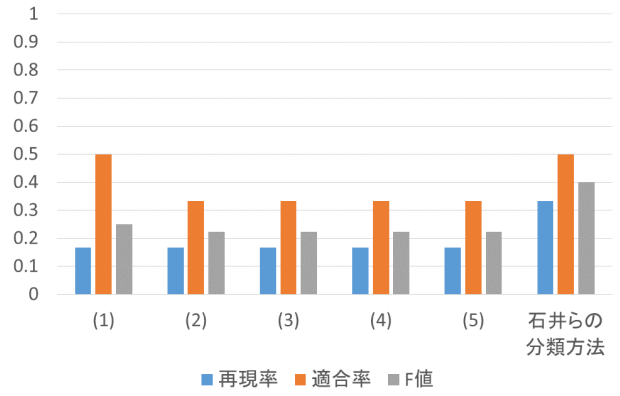

図 6　Data の引用意図分類結果
Fig. 6　Result of citation intention classification of “Data”

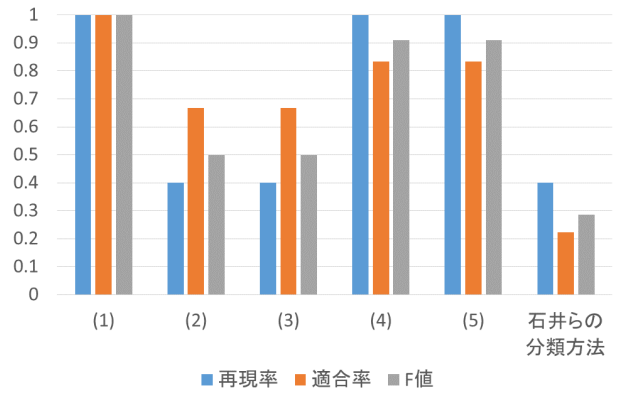

図 7　Equation の引用意図分類結果
Fig. 7　Result of citation intention classification of “Equation”

高くなった．

図 6 では，選出方法 (1)〜(5) のいずれの方法においても Data の F 値が低い．Data は石井らの方法よりは手掛かり語が多いが，出現回数の少ない手掛かり語が多かったため

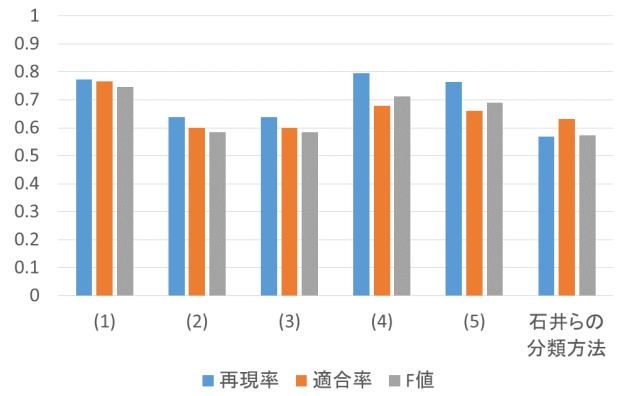

図 8　引用意図分類結果の比較
Fig. 8　Comparison of result of citation intention classification

表 3　分類した引用意図クラスと件数
Table 3　Classified citation intention classes and their numbers

		分類クラス				
		Group	Method	Result	Data	Equation
正しいクラス	Group	114	3	3	0	0
	Method	56	761	65	9	4
	Result	6	3	101	0	0
	Data	5	6	10	9	0
	Equation	0	6	0	0	19

分類スコアが正しく算出されず，精度が低くなった．

図 5，図 7 より，図 2 において正解率の高かった (1)，(4)，(5) と比べて (2)，(3) は，Result，Equation の F 値が低くなっている．これは，Result，Equation の引用箇所に含まれる単語には出現が低頻度のものが多いためで，(2)，(3) のように 1 度しか出現しなかった単語を手掛かり語から除外すると，分類できない引用箇所が増えた．

図 8 に選出方法ごとの分類結果の平均を示す．図 8 より，先行研究である石井らの分類と比較して，F 値の平均では全ての選出方法において上回ったが，図 5 の Result や図 6 の Data において石井らの分類方法より悪かった (2)，(3) は，図 8 でも石井らの分類方法による結果と大きな差は見られなかった．また，選出方法 (1)〜(5) のうち，最も精度が高かったのは，1 クラスにしか現れない単語をそのクラスの手掛かりとした選出手法 (1) となった．

表 3 に，分類した引用意図のクラスとその件数を示す．表 3 を見ると，Method を Result や Group と誤っているものが多かった．これは，Method の引用箇所が最多であることの他に，分類スコアが手掛かり語の異なり語数で割ったものであることが影響している．そのため，手掛かり語の多い Method では他の 4 クラスに比べて 1 つの手掛かり語の重みが相対的に軽くなってしまい，他クラスの手掛かり語が含まれている場合に誤分類が起きやすい．

6. まとめ

本稿では，論文中の引用を表す文から手掛かり語を選出し，これらを用いて引用箇所の引用意図を分類する手法を提案した．まず引用箇所から手掛かり語の候補となる単語を抽出し，“Group”，“Method”，“Result”，“Data”，“Equation”の 5 つの引用意図ごとに出現回数をカウントして，1 クラスにしか現れない候補をそのクラスの手掛かり語とした．そして，分類対象の引用箇所に含まれる手掛かり語の数を，各クラスの手掛かり語の異なり語数で割って分類スコアとし，それが最大となるクラスに引用箇所を分類した．評価のため，NTCIR-9 の論文中の引用箇所 236 文を正しく分類できるかどうか実験し，先行研究である石井らの分類方法による結果と比較した．その結果，Data 以外の 4 種類のクラスにおいて石井らの手法より分類精度が良かった．Data の結果が悪かったのは，手掛かり語の件数が少なかったことが主な理由と考えられる．

今後の課題としては，Data の分類精度の向上を図りたい．また，他の論文誌などでも実験を行い，手法の有用性について検討する必要がある．

謝辞

本研究の一部は，科学研究費補助金基盤研究 (C)(課題番号 25330384，15H02789) および国立情報学研究所公募型共同研究の援助による．ここに記して深謝する．

参考文献

[1] 石井仁子，太田学，高須淳宏，“引用意図を利用した学術論文閲覧支援のための適切な被引用箇所の特定”，第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM Forum 2015)，F3-5，2015.
[2] 島広幸，建石由佳，“章立てに注目した論文の研究内容による自動分類”，言語処理学会第 16 回年次大会発表論文集，pp.341-344，2010.
[3] MeCab : http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html
[4] 榊剛史，松尾豊，石塚満，“制約付きクラスタリングを用いた論文分類”，The 20th Annual Conference of the Japanese Society for Artificial Intelligence，1A1-1，2006.
[5] 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.
[6] 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.2640-2649，2001.
[7] NTCIR-9 (the 9th NII Test Collection for IR Systems) : http://research.nii.ac.jp/ntcir/ntcir-9
[8] Xpdf : http://www.foolabs.com/xpdf
[9] TreeTagger : http://www.cis.uni-muenchen.de/schmid/tools/TreeTagger