論　文

構造と構成要素に基づく学術論文の自動判定*

石田栄美*1
安形輝*2
宮田洋輔*3
池内淳*4
上田修一*5

ウェブ上に存在するPDFファイル群から学術論文を自動的に判定する手法の開発を目的とした。まず，学術論文の構成要素と構造がどのように発達してきたかを調査した。英語と日本語の学術論文1,172件に対して，それらが顕れているかを調査した。その結果，論文は共通した構成要素を有しており，見出しを持つ論文のうち40%近くがIMRAD形式またはそれに近い構造を採っていた。次に，これらの結果をもとに，学術論文を自動的に判定するためのルールを構築した。ウェブ上から無作為に収集した英語と日本語のPDFファイル集合を用いて判定性能を実験したところ，ランダムフォレストによる判定器を用いた場合，F値は英語集合では0.74，日本語集合では0.53であった。これらの結果から，本研究で用いたアプローチにより構築した判定ルールにより，ウェブ上に存在するPDFファイル群から学術論文を自動的に判定できる可能性が示唆された。

目　次

* 2013年1月4日受付　2014年4月5日受理
*1 いした　えみ　九州大学
*2 あがた　てる　亜細亜大学
*3 みやた　ようすけ　帝京大学
*4 いけうち　あつし　筑波大学
*5 うえだ　しゅういち　立教大学

1. はじめに

一般に文書は，その目的や役割に応じて特有の形式と構成を持っており，その構成要素や記載順序は社会習慣や法令・規則によって規定されている。例えば，法令文，判決文，事務文書あるいはビジネス文書には，決まった形式がある。また，特許出願書については，「特許法施行規則」があり，規格には『JIS Z 8301 規格票の様式及び作成方法』[1] があるように，形式に関して詳細な規定が用意されている場合もある。

多くの場合，文書は時を経るに従って，その文書を作成し利用する人々の間に一定の合意が生まれ，慣習となり，取捨選択がなされ，特有の形式や構成が定着してきたと考えられる。

現在では，多くの文書はデジタル・ファイルとして扱われるようになってきた。そこで，文書の持つ形式や構成を分析し，特徴を抽出し，大量の文書ファイルの中から特定の種類の文書を判定することが研究課題として取りあげられるようになってきた[2]。

ここでは，文書形態の一種として，学術論文を取りあげる。どのような分野においても学術的な研究成果は，一定の手続きを経て公表される。学術論文は，学問分野の専門化，学術雑誌の普及等とともに特有の形式を持つようになってきたと考えられる。

本稿では，ウェブ上に存在する文書の中から，こうした学術論文特有の形式や構造，特徴をもとに学術論文を自動的に判定する方法について検討する。具体的には，クローリングにより収集したPDFファイル群から学術論文PDFファイルを自動判定する方法を考案し，評価する。なお，特定の専門分野や言語に依存しない判定方法の構築を目的としていることから，テキスト中の語の出現数の多寡等の情報を用いず，学術論文の形式や構造にもとづいた学術論文判定ルールを構築した。さらに，判定ルールの有効性，汎用性を検証するため，日本語と英語で書かれた学術論文を対象に，判定実験を行った。

筆者らは，既に，日本語学術論文のPDFファイルの自動判定の意義や方法について検討した[3]。前報では，電子ジャーナルや論文のセルフアーカイビングが普及し，学術論文のPDFファイルの自動判定が重要な課題となっていることを指摘したが，その後，言語や分野を問わず，その傾向は，さらに一般化してきているといえる。そのため，日本語の論文に限ることなく，学術研究における国際共通言語でもある英語の論文に対しても，学術論文のPDFファイルの自動判定を行うことが必要であろう。また，その背景として，こうした学術論文を取り巻く環境の変化により，学術論文の要素と構成，形式や文体は，国際的に，また分野を超えて，さらに共通性が強まりつつあると考えられる。

そこで，まず，2章において学術論文の形式と構成要素の歴史と現状を概観する。3章では，2章で得られた論文の形式と構成要素に関する実態調査を行い，学術論文の自動判定ルールに用いることができる特徴を確認する。4章では，判定ルールを構築し，実験用集合を用いて，その有効性を検証する。

2. 学術論文の要素と構成

八杉龍一は，学術論文を“それぞれの学問分野で専門の研究者によって書かれるもので，その著者が自分の研究でえた結果を報告し自分の意見をのべたものであり，それによってその学問分野に新知見をもたらすもの”[4] と定義している。学術論文の著者は専門分野を持つ研究者であり，論文はその研究成果を報告し，その中には新しい知見が含まれている。学術論文は，研究成果を説明するだけでなく，そのオリジナリティを主張するという目的にそった形式や構成が選ばれることになる。

研究者は，論文を投稿する学術雑誌を選択するが，論文の形式や構成は，各学術雑誌の投稿規程や執筆規程で定められており，投稿する著者はこれらの規程に従って執筆することになる。このように，学術雑誌によって規程は異なるが，背後には共通の規則があると考えることができる。例えば，図の題名を「第1図」とするか，「図1」と

するかは，個々の雑誌によって異なるが，図の題名は図の下に，表の題名は表の上に記載するのは共通の規則である。

2.1 学術論文の発展

ザイマン（Ziman, John）は，科学のコミュニケーションのもっとも重要な媒体は，学術定期刊行物における「原論文（primary paper）」であり，17世紀後期のまったく新しい発明だったと述べている[5]。それまでの科学情報の伝達手段は，手紙（letter）であり本だった。科学論文は，数ページと極端に短く，新しい原理を組み立てるものではなく，引用文献によって，これまでの研究で固められた足場から，おそるおそる一歩を踏み出すものであり，これは近代科学が，激しい競争下にあるにもかかわらず，協力的であることを示している。また，短期間で公表され，論争を引き起こすと同時に優先権を確立する働きをしている。もう一つ，科学論文は，公共の文書となり，検索と引用のための組織化がなされる[5]。

1665年の *Philosophical Transactions* 等の学術雑誌の創刊とともに，現在のような論文形式が始まったわけではなく，かなり長く手紙を中心とする時期が続いた。1740年から1859年の *Philosophical Transactions* 中の論稿に占める手紙の割合を調査した飯田崇文は，1760年代には，手紙が5割を占めていたが，その後の一世紀の間に徐々に減少し，1850年代には，ほぼなくなっていると報告している[6]。

19世紀の後半には，学術雑誌と学術論文の数が増大し，ヴィッカリー（Vickery, Brian C.）によれば，19世紀中に200万点の科学技術論文が発表された[7]。ハーモン（Harmon, Joseph）は，この19世紀には実験計画法の大幅な進歩があり，統計的手法や実験結果と観測結果の説明等が論文に導入されたため，専門研究者による雑誌編集において，新しい形式が求められるようになったと述べている[8]。

その結果，20世紀になると項目別構造（topical structure）と呼ばれる，論題，抄録，序論，方法あるいは実験の詳細，結果，考察，結論，謝辞，引用文献という形式に変わった。ハーモンによれば，これは，実験に関する報告の中で論理的にまた効率的に必要な情報を組織化するに論理的で効率的な方法だった[8]。トレリース（Trelease, Sam F.）らによる英語で書かれた最初の「論文の書き方」（1927）でもこの構造が推奨されている[9]。

日本では，1929年から1934年にかけて，三点の「論文の書き方」が出版されている。最も古い久保猪之吉著『医学論文の書き方　後篇　第1』（1929）[10]は，論文の構成として，“1. 表題及目次, 2. 緒言, 3. 歴史, 4. 材料及方法, 5. 結果成績, 6. 総括結論, 7. 謝辞, 8. 文献, 9. 抄録”をあげている。「3. 歴史」とは“是迄の同問題に関する歴史的記述を分析によつて”行うものであるとされ，すなわち先行研究の記述であるので，全体の構成は，ほぼ現代に近い形であるといえる。久保のあげた論文の構成は，ハーモンのいう「項目別構造」，すなわち，トレリースらの「論文の書き方」によく似ているが，影響関係は明らかではない。

2.2 IMRAD形式の普及

一方，学術論文の構成について，IMRAD形式，すなわち，序論（introduction），方法（method），結果（result），それに検討と結論（discussion）からなる論文の構成の発展をもとにした説明がある。デイ（Day, Robert A.）は1989年に，パスツール（Pasteur, Louis）の論文がIMRAD形式であったことを明らかにしている[11]。また，デイとガステル（Gastel, Barbara）は，論文の書き方の解説書の中で科学論文の歴史について言及し，ハーモンとは異なった説明をしている[12]。17世紀半ばに誕生した学術雑誌の記事は，叙述的だった。“「私は最初これを観た。それからこれを見た」とか「まず私はこれを行い，次にあれを行った」”と時間軸に沿った記述が多かった。19世紀になると科学は発展をとげ，その中で方法が重要となってきた。微生物学研究のための純粋培養法を発展させたパスツールは，論文に自分の実験方法を詳細に記述した。その後，実験の再現性が研究の基本として認識されるようになった，そして，パスツール流の方法を中心とした論文の構成が力を持つようになった。これが，現在のIMRADと呼ばれる形へと発展してきた[12]。

デイらはIMRADの論理性は質問の形式で示すことができるとして，次のように述べている。

「どのような疑問（問題）について研究しているのか？」という質問の答えが「序論」です。「その研究をどのように進めたのか？」という質問の答えが「方法」となります。また，「何が見出されたか？」に対する答えが「結果」であり，「それらの結果は何を意味しているのか？」という質問に対する答えが「討論」ということになります。

そして，IMRADは著者の指針となるだけでなく，編集者や査読者，論文の読者に共通の基盤を与えるといっている[12]。

ソラッチ（Sollaci, Luciana B.）らは，1935年から1985年までの*the British Medical Journal, JAMA, The Lancet, the New England Journal of Medicine*の医学のコアジャーナル四誌に掲載された論文におけるIMRAD方式の状況を調査した[13]。1935年には，どの雑誌にもIMRAD形式の論文はなかったが，1960年代から急速に増加した。*The New England Journal of Medicine*では1975年に100%となり，1985年には他の三誌においても全論文がIMRAD形式となったことが明らかになっている。ただし，詳しくみれば，*the British Medical Journal*では，部分的にIMRAD形式である論文は1935年でも2割存在し，1960年代までは増加しているので，医学分野におけるIMRAD形式への移行は，長期にわたり緩やかに進んだとみられる。

現在では「項目別構造（topical structure）」という語は使われず，IMRADという用語が一般に使われるようになってきた。「項目別構造（topical structure）」は，論文の本体だけではなく，テキストに附随する論題，著者や引用文献等の要素を含んでいる。一方，IMRADは，本体部分のみの構造である。

IMRADは医学生物分野を中心に発展し，現在では，自然科学から社会科学までの広い分野で共通に用いられていると考えられる。原著が1977年に刊行されたエコ（Eco, Umberto）著の『論文作法』は，人文科学を対象とした論文の書き方解説書であり，1991年に邦訳されて以来，刷数を重ねているが，この中には，論文の構成に関しての記述はない[14]。人文科学では，論文の構成についての規範への意識は薄かった。

澤田昭夫は同じ1977年に，自然科学分野の論文構成について理解した上で，分野別の違いを強調しない立場に立って，

すべての学問分野に共通しているのは，さまざまなデータ，事実を眺め，不可解な現象についての問を考え，その問に答えるための仮説をたて，その仮説を事実とつきあわせてそれが現実をよく説明するかどうか検証して答を出すというプロセスです。自然科学の論文はだいたいどれも，「問を示す序」，「実験の方法」，「実験の結果・評価・論議，結論」というふうにでき上っています。人文，社会科学では結果を再現するような実験はできませんが，事実によって仮説を検証する手続きは，自然科学の実験による検証の過程に相当するもので，全体の手続きは根本的には違っていません。

と述べている[15]。

さらに，澤田は，社会科学と人文科学では，現象の発生と成立とを順を追って説明する「時間的アプローチ」と論理的，分析的に説明する「論理的アプローチ」とがあるとし，5W1 Hの5Wが議論の整理の手がかりとなると述べている[16]。これは，前述のデイのIMRADの説明と通じるところがある。このように人文科学でも論文の構成についての意識がたかまり，また類型の提案もなされるようになってきた。

学術論文の本論にあたる部分の構成は，自然科学分野では，序論，方法，結果，結論という構成をとるということは，数十年前から定着している。しかし，これをIMRADと呼ぶようになったのは，最近のことである。国内の社会科学や人文科学の研究者のほとんどは，IMRADという語を知らなくても，自然科学の論文では，序論，方法，結果，結論という構成を用いていることは徐々に理解されてきている。

さらに，学術論文を構成する要素として，本論の他に，論文名，著者名，所属，抄録，キーワード，謝辞，引用文献があることも論文の書き方の指南書で繰り返して述べられてきている。たとえば，1986年に制定された科学技術情報流通技術基準

の「学術論文の構成とその要素（SIST08）」（2010年版は「学術論文の執筆と構成」[17]）では，学術論文の構成要素とその記載要領についての基準が示されている。

また，学術論文が電子ジャーナルに掲載されるようになったために起きた変化も見逃すことができない。たとえば，論文の識別子であるDOI（Digital Object Identifier System）が掲載されることが多くなっているが，これは，学術論文の自動判定に利用できる。そして，言語による影響を受けやすいが，論文中で用いられる特有の表現もPDFファイル群から学術論文のPDFファイルを区別するのに重要な手がかりといえる。

以上のように，学術論文には形式や基本となる構造，構成が存在し，論文執筆の指南書等でもそれらに従うことが推奨されている。実際に，論文の形式，構造，要素が共通していれば，学術論文を自動的に判定する際の重要な手掛かりとなる。

3. 学術論文の構造と構成要素に関する調査

学術論文におけるIMRAD形式の割合，構成する要素を調査した。本研究では，ウェブ上に存在する学術論文PDFファイルに対して自動判定することを目的としているため，調査対象は，ウェブ上に存在する学術論文とした。

3.1 論文の構造・構成に関する既往調査

論文の構造や構成要素を把握しようとする試みはいくつかの研究例がある。医学分野では，ソラッチらが，1935年から1985年までに*the British Medical Journal*，*JAMA*，*Lancet*，*New England Journal of Medicine*に掲載された1,297件の原著論文を対象に調査を行った[13]。1985年には医学分野ではIMRAD形式を採る論文が100％になっており，1955年から75年の20年間で，それまでの4倍にまでIMRAD形式が普及したことを明らかにした。物理学分野では，バザーマン（Bazerman, Chrales）は，1893年から1980年までに*Physical Review*に掲載された学術論文の，様々な特徴について調査している[18]。*Physical Review*掲載論文では，1950年以降，論文が見出しによって構造化されていき，時代が新しくなるにつれ，見出しとして論文固有の名前を用いるのではなく，「Experiment」のようなより抽象化された見出しを持つようになっていったことを指摘している。実験心理学分野では，倉田敬子と坂上貴之が，*Journal of Experimental Analysis of Behavior*，*Journal of Experimental Psychology: Animal Behavior Processes*，*Journal of Experimental Psychology: Human Learning and Memory*，*Memory and Cognition*の1975年と1998年の実験を行った全論文を対象とした調査を行った[19]。序論や考察の割合が増え，結果をそのまま提示するものから，議論や解釈の重視へと変化してきていることを示した。リン（Lin, Ling）らは，2007年に掲載された，工学，応用化学，社会科学，人文学の39分野の433論文を対象に調査した[20]。多くの実証論文がIMRAD形式に従っているが，「序論 — 文献レビュー — 方法 — 結果と考察 — 結論（ILM[RD]）」という構造を採る論文が最も多かったとしている。

以上のように，分野を限定した調査はいくつか見られ，分野による特徴は明らかになっている。本研究では，分野を限定しない調査を行う。

3.2. 論文の構造・構成要素の調査

3.2.1 調査対象

分野や学術雑誌の格付けに依存しない広範な学術論文集合を得るために，国内で出版された雑誌については「CiNii本文収録刊行物ディレクトリ」に掲載されたタイトルから無作為に抽出した。ディレクトリには英語の雑誌も含まれているが，無作為に抽出したものの中には含まれていなかった。これらの集合を，以下では，「日本語集合」と呼ぶ。国外の雑誌については*Web of Science*，*Journal Citation Reports*，*Arts & Humanities Citation Index*の収録誌リストから無作為抽出して選定した。選定した雑誌には，英語以外の雑誌タイトルやタイトルは英語であっても本文は他の言語である雑誌も含まれていた。しかし，これらの雑誌から以下に述べる方法で選ばれた論文は，すべて英語で書かれた学術論文であったので，これを「英語集合」と呼ぶ。

標本に含まれた雑誌タイトルに対して電子ファイルが提供されているかを，データベース，検索エンジン等を用いて手作業で確認した。PDFファ

イルが入手できる場合は，比較的新しく，また入手したファイルの出版時期を揃えるため，2010年第1号までの中から最新号を選択し，当該号の雑誌の中で2番目に掲載された論文を調査対象としてダウンロードした。2番目に掲載された論文としたのは，最初の論文は巻頭言や概説である可能性が高いためである。本調査では，ウェブ上に存在し，PDFファイルで入手できるもののみを対象としたが，できるだけ多くの学術論文を収集するため，有料のものも含めた。入手したPDFファイルの中に書誌や詩等，学術論文の形式をとらないものがあった場合，雑誌自体を調査対象から除外した。その結果，日本語集合では490ファイル，英語集合では682ファイルの計1,172ファイルを取得した。

3.2.2 調査項目

2章で概観した学術論文の構成要素をもとに，PDFファイルに，「論文名」，「著者名」，「所属」，「抄録」，「キーワード」，「引用文献」，「掲載雑誌名」，「ISSN」，「DOI」の構成要素の有無，構造を表す「見出し」の有無を確認した。また，「書字方向(縦書・横書)」も確認した。

論文が「見出し」を持っている場合は，その見出しの文字列を記録した。記録された見出しが，表1に示した語によって明示的に構造が表現されている場合のみ，構造に関するラベルを付与した。たとえば，「考察」という見出しであれば，「結論(Discussion：D)」という構造が明示的にあったとみなすが，考察にあたる章が「本提案手法の有効性」という見出しでは，明示的に構造があるとみなさない。ラベルは，IMRADを構成する「序論（Introduction：I)」「方法（Method：M)」「結果（Result：R)」「結論（Discussion：D)」に加え，論文の構造調査で対象となっている，「文献レビュー(Literature review：L)」と「結語(Conclusion：C)」を追加した6ラベルとした。一つの論文の中で同じ役割のラベルが連続して付与された場合は，それらを一つにまとめた。たとえば，「はじめに」，「試料」，「実験方法」，「結果」，「考察」という5つの見出しで構成される論文は，IMMRDというラベルが付与されるが，連続したMをまとめ，IMRDとした。

表1 学術論文の構造を明示する語の例

	英語	日本語
I	Introduction Background	はじめに 序論
M	Material and Method Methodology	試料と方法 実験
R	Results Finding	実験結果 結果
D	Discussion Implication	考察 議論
L	Literature Review Related Research	文献レビュー 先行研究
C	Conclusion Summary	おわりに 結論

3.3. 調査結果

3.3.1 構成要素

論文の構成要素について集計したところ，論題名（英語100%，日本語100%)，著者名（英語99.9%，日本語100%)，所属（英語97.2%，日本語100%)，掲載誌名（英語98.5%，日本語97.8%）という学術論文を識別するための基本的な要素については，ほぼすべての学術論文で明示されており，また日本語集合と英語集合との間に大きな差はなかった。その他の要素に関して対象集合別に，表2に示した。抄録（英語89.0%，日本語89.2%)，引用文献（英語98.4%，日本語98.8%）に関しては，論題名や著者名と比べると低いが，いずれも高い割合で明示されており，調査対象とした集合の間で大きな違いはなかった。キーワードは，抄録や引用文献に比べては低いが，日本語で75.5%，英語で63.2%の論文に含まれていた。書字方向は全ての論文が横書きであった。

学術論文の構造化の指標とした見出しの有無に関しては，日本語集合のうち見出しがないものが0.8%に対し，英語集合は7.3%であり，英語集合の方が見出しのない論文の割合が高かった。見出しを持たなかった論文50件のうち，44件(88.0%)が *Arts & Humanities Citation Index* から抽出されたものであったことからもわかるように，英語集合の学術論文が，文学や宗教学のような人文学の論文を多く含んでいたことが，その要因と考えられる。

ISSNとDOIでは，言語間で大きな違いがあった。ISSNは日本語で出版された学術雑誌ではど

表2 学術論文を構成する要素の比率

	日本語（N=490）		英語（N=682）		計（N=1,172）	
抄録	437	89.2%	607	89.0%	1,044	89.1%
キーワード	370	75.5%	431	63.2%	801	68.3%
引用文献	482	98.4%	678	98.8%	1,156	98.6%
見出し	486	99.2%	632	92.7%	1,118	95.4%
ISSN	0	0%	293	43.0%	293	25.0%
DOI	1	0.2%	508	74.5%	509	43.4%

れにも含まれていなかったが，英語集合では4割以上の論文に含まれていた。DOIに関しても，日本語集合にはほとんど含まれていなかった。これは，2012年3月15日に「ジャパンリンクセンター」が認定されるまで日本にはDOIの登録機関がなく付与手続きが容易ではなかったためと考えられる。

3.3.2 構造

本調査では，レビュー論文は調査対象から除外したため，調査対象となったのは，見出しを持っていた1,118件のうちの1,087論文である。調査の結果，35パターンの構造が検出された。総計で上位10構造を表3に示した。

Introduction以外の明示的な構造を持たない論文が最も多く，380件と全体の35%を占めていた。次にIMRD，IMRDCが多かった。IMRAD形式の中でも，最後にConclusionを持つ事例も多かった（187件，17%）。また，IMRAD形式を採っていてもIntroductionの見出しを持たず，今回の調査ではMRDとなる事例もあった。完全なIMRAD形式ではないが，Conclusionが付加されているもの，Introductionを持たないもの等を派生的な形式も含めると，ウェブ上に存在する学術論文PDFファイルは，40%程度がIMRAD形式を採っていることがわかる。

1985年には，医学分野の論文の100%がIMRAD形式を採っていたソラッチらの調査[13]と比較すると，医学系以外の分野も含まれる本調査の結果は，IMRAD形式に近い形式を含めても40%程度と，その比率は低かった。ただし，本調査でも，生物科学や医学分野の論文に関してはIMRAD形式やIMRADにConclusion等の構造が一つ加わるIMRD[C]や，IMRADのうちの一つの構造が存在しない形式等，IMRADに近い形式を採る論文は多かった（生物科学60.9%，医学68.3%）。39の分野のインパクトファクターが高い英文誌を対象としたリンらの調査[20]では，ILM[RD]CやIM[RD]CのようなIMRADに近い形式が80%以上を占めていた。この調査よりも本調査対象の論文ではIMRAD形式の割合が低かったが，その理由として，本調査対象は，様々な分野が含まれるだけでなく，インパクトファクターが高い雑誌に収録された論文に限定せず，ウェブから入手できる論文を対象としたことが考えられる。

表3 学術論文の上位10構造

構造	件数	比率（%）
I	380	35.0%
IMRD	191	17.6%
IMRDC	187	17.2%
構造なし	136	12.5%
IMRD	45	4.1%
IM	35	3.2%
IMRDC	19	1.7%
IMRDCL	16	1.5%
IMRDC	14	1.3%
IMR	10	0.9%
その他	54	5.0%
計	1,087	100%

3.4 学術論文の構成要素と構造

本調査では，2章で得られた論文の構成要素や構造が，ウェブ上に存在する学術雑誌に，どの程度，顕れているかを調査した。その結果，調査対象の論文は，論題名，著者名，所属，掲載誌名，抄録，引用文献，キーワード等多くの共通した構成要素を持っていることが明らかになった。構造

については，明示的にIMRAD形式，IMRADに近い形式を採っている論文が全体の40%程度であった。表3において，構造を持たない論文を除いた，IMRAD形式やそれに準ずる構造を持つものの割合は87.5%である。以上の結果は，インパクトファクターが低い雑誌を含めた場合にも学術雑誌に掲載された論文では，上に示したある種の共通する構成要素，構造を有していることを示したといえる。また，これらのことは，どのような言語で書かれた論文かを問わず，英語と日本語で共通の特性である。

他の文書が混在する文書群から学術論文を自動的に判定することを考えた場合，これらの共通する構成要素や構造の情報を利用することは有効であるといえる。すべての構成要素や構造を有していなかったとしても，部分的に有していれば，学術論文の可能性が高いと考えられる。また，本調査は，学術雑誌に掲載された論文のみを対象としているが，同様の形式を用いている大学紀要の論文，学会発表の予稿等にも適用できる。次章では，本調査から得られた学術論文が共通に有している構成要素や構造に関する情報を利用した学術論文を自動的に判定するルールの構築とその検証を行う。

4. 学術論文の自動判定

学術論文に特有の構成要素や構造を用いて学術論文を自動判定するためのルールを構築した。さらに，そのルールの有効性を検証するために，大規模なPDF集合を構築し，学術論文の自動判定実験を行った。

なお，本論文における「学術論文」は，一定の体裁を持ち研究成果を伝達するために書かれている論文とし，学術雑誌に掲載された論文だけでなく，大学紀要の論文，会議録，学会発表の予稿を含めている。

4.1. テキストのジャンル分類に関する既往研究

学術論文を自動的に判定する手法については，テキストのジャンルやタイプを判定する研究手法とアプローチが類似している。

アルガモン（Argamon, Shlomo）らは，実験科学と歴史学のそれぞれ6分野，計12の学術雑誌を対象に，それらを識別することを試みている。機械学習手法を用い，theやas等の機能語の頻度と学術雑誌論文を分析して得られた拡張やコメント，モダリティを表現するand, moreover, in general等を特徴素として用いている[21]。スタマタトス（Stamatatos, Efstathios）らは，テキスト中に出現する単語の頻度情報をもとに，*The Wall Street Journal*紙のコーパスを用いて，Review & outlook（Editorial）やLetters to the editor等のジャンルにテキストを分類する実験を行っている[22]。その際，頻出語を特徴素として用いているが，句読点の位置も重要な情報であることを指摘している。クマリ（Kumari, Pranitha K.）らは，ジャンルに関連する語を抽出するため，語幹処理手法を組み合わせた手法を用いて抽出した特徴素を用いて，ブログ，オンラインショップ，リスト等で構成されている7genreコーパスに対してランダムフォレスト（Random forest classifier）を用いてジャンル分類を行っている[23]。カナリス（Kanaris, Ioannis）らは，7genreコーパスと芸術，リンク，ヘルプ，オンラインショップ等8ジャンルで構成されているKI-04という2つのテストコレクションを用いて，分類実験を行っている[24]。この際に用いた特徴素は，文字列ベースのNグラム（Character n-gram）とページに含まれるHTMLタグである。分類器には，サポートベクターマシン（SVM：Support Vector Machine）を用いている。また，リム（Lim, Chul Su）らは，ウェブページを個人／公共／企業のホームページ，リンク集，研究レポート，FAQ等11種類のジャンルに分類することを試みている[25]。その際，特徴素として，(1) URLの深さ，ファイルの名前，ドメイン名，faq, paper, researchのようなURLに含まれる語等URLに関する情報，(2) HTMLタグ，(3) 文字数，語数，文数，一文における語の数，9つのPOS（part-of-speech）タグの数等テキストに出現する語とPOSタグの情報，(4) 内容語と機能語の数，(5) 平叙文や命令文の数等，構造的な情報を特徴素として分類を行っている。シェファード（Shepherd, Michael）らは，テキストを内容（Content），形式（Form），機能（Functionality）の3属性に分け，それぞれの特徴を用いて分類を

行っている[26]。内容には，メタタグや一般的な語の使用回数等が含まれ，形式には画像数，CSSの定義，ページのドメイン・サブドメインの情報，語数等が含まれる。機能は，リンク数，外部リンクや内部リンクの割合等の情報である。これらの情報を用いて，個人／企業／組織のページ等のジャンルにウェブページを分類している。ターパ（Thapa, Chaman）らは，テキスト中に出現する語からTF-IDFの値が高い語を選択し，POSタグの情報，固有表現，リンク構造，サブドメイン数，パスの平均の長さ等のURLの情報を用いて，公共／私的／非営利団体／営利団体等のホームページの分類を試みている[27]。

テキストのジャンルやタイプの判定を行う研究では，既存のジャンルやタイプに分類されたテキスト集合を用いて，テキストに出現する語やテキストの属性情報を特徴素として分類器を学習させている。分類器そのものの開発をしている例もあるが，分類器に入力する特徴素の選定に重点をおいている研究が多い。本研究においても，同様に，特徴素の選定に重点をおく。

また，これらで用いる特徴素は，テキスト中に出現する数百から数千の語及びその頻度情報が用いられることもあるが，主題を対象にした分類とは異なり，ジャンルやタイプの特徴を表した特徴素を用いている。本研究では，前報[3]で用いた判定ルールと同様，学術論文の特徴を表す語を特徴素として用いた。学術論文は，3章の調査で示したように，分野や言語を問わず学術論文としての共通の構成要素や構造が存在する。これらの特徴はテキスト中に多く出現するものではないが，いくかの構成要素や構造が出現していれば，学術論文として成り立つ特徴である。学術論文というタイプを判定するには，これらの特徴を特徴素として取り入れることが有効であると判断した。また，通常は，特徴素の出現傾向等，判定器が学習するため大量の学習用データが必要になるが，学術論文の特徴を表現する特徴素が出現するかしないかで判断するため，大量の学習用データが必要ではないという利点もある。

さらに，ウェブページを対象にした分類の場合は，語に加えて，HTMLタグやURL，リンク構造，ファイルの属性情報も特徴素として用いている例が多かった[26][27][28]。本研究では，PDFファイルが対象のため，HTMLタグの情報を用いることはできないが，ファイルのURLのドメイン名，URLに用いられている単語，ファイルの属性情報は，既存の研究と同様，特徴素とした。

なお，本研究では，これらの特徴素群を学術論文を特徴づけるものとして考えており，学術論文か否かを判定する規準となるものであることから，ここでは「判定ルール」と呼ぶ。

4.2. 判定ルールの構築

前報[3]で用いた判定ルールは，「ファイルサイズ」，「ページ数」，「ページの向き」，「ファイルのURLのドメイン」，「文体」，「出現キーワード」等から構成されていた。本稿では，前章で示した論文の構造と構成要素を反映した新しい判定ルールを構築した。これには，ファイルの属性，論文の基本構造，構成要素，論文／非論文中の表現を基本的なカテゴリとしている。また，言語に依存しない汎用的なルールを目指した。構築した判定ルールを表4に示す。

ファイルの属性としては，「ファイルの構造」，「ファイルのドメイン」，「URLに含まれる語」を特徴素として用いた。ファイルの構造は，「ファイルサイズ」，「ページ数」，「レイアウト」，暗号化の有無である。ファイルサイズ，ページ数，レイアウトは，後述する論文と非論文で構成される実験用集合を構築した際に調査し，差異があったものを用いた。ドメインとは，そのファイルのURLのドメイン情報を用いている。

論文の基本構造は，先に示したIMRAD形式に従い，序論，方法，結果，考察，結論とした。構成要素は，抄録，キーワード，引用文献，謝辞，図表，附録とした。判定ルールに適用する場合は，これらが見出し語として出現するか否かを基準としている。見出し語の判定は，語の前に半角空白があり，語の後に改行があるものとした。また，同じ役割を示すものでも，実際の論文の中で使用される語の表現が異なる場合がある。たとえば，「抄録」という構造は，英語論文では，「abstract」「summary」等の語で表現されており，日本語では「抄録」「要約」「要旨」「概要」の他に「abstract」「summary」で表現されている場合もある。その

表4　判定ルール

		英語	日本語
ファイルの属性	構造	ファイルサイズ	
		ページ数	
		レイアウト（縦型・横型）	
		暗号化の有無	
	domain	.edu	.ac.jp
		.com	.com|.co.jp
		.gov	.go.jp
	URL中の語	paper|article|research	paper|article|research
基本構造	序論	introduction|background	はじめに | 序論 | 緒言 | 目的 | 背景
	方法	method|methodlogy	手法 | 方法 | 分析
	結果	finding|result	（実験 | 調査）結果
	考察	discussion|implication	考察 | 議論
	レビュー	literature review|related research	（先行 | 既往）（文献 | 研究 | 調査）
	結論	conclusion|summary	（結論 | 結語 | まとめ | おわりに）
構成要素	抄録	abstract|summary|synopsis	抄録 | 要約 | 要旨 | 概要 |abstract|summary
	キーワード	keyword|key word	keyword|key word| キーワード
	引用文献	bibliography|reference	（（引用 | 参考 | 参照）文献 |reference| 参考資料 | 注 | 註）
	謝辞	acknowledgement	謝辞 |acknowledgement
	図表	(figure| fig.), table	図 | 表
	附録	appendix	appendix| 付録 | 附録
	所属	university|institute	大学 | 研究所 | † | ‡ |University
	欄外誌名	journal|bulletin	紀要 | 学会誌 | 論文誌
論文中の表現	研究	research	研究
	論文	paper|article	研究報告 | 論文
	コード	doi |issn	doi:|doi |issn
	本論文	this (article|paper|research)	本（論文 | 研究）
	被験者	subject	被験者
	研究法	survey|experiment|analysis|theory| hypothesis	実験 | 調査 | アンケート | 理論 | 仮説
	査読者	referee|reviewer	査読者
非論文	報道資料	press release	プレスリリース | 報道資料

ため，本研究では，英語論文の場合，判定ルールでは「abstract」「summary」「synopsis」のいずれかの語が出現していれば「抄録」という構造が出現したとみなした。

論文中に用いられる表現，非論文中で用いられる表現は，論文の特有の表現として本文中で用いられる可能性が高い語を選定した。選定に関しては，前報[3]で用いた表現をもととし，著者らが行った論文の判定作業等から得られた語等をもとにした。結果として，研究，論文，コード，本論文，被験者，研究法，査読者等について言及するときに用いられる表現を選定した。DOI，ISSNのコードは，論文の本文ではなく，フッターやヘッダーに記述されていることが多いが，先の実態調査においても有効性が認められたため含めた。「欄外誌名」とは，コード同様，欄外に記述されている事が多い論文誌名のことであり，「紀要」「学会誌」「論文誌」等が含まれる。

学術論文の判定は，これらの判定ルールを判定器に特徴素として入力し，判定器が学術論文か否かを判定する。ファイルサイズとページ数は，数値をそのまま判定ルールの特徴素として用いる。URL中の語は，URLに含まれる語にpaperやresearch等があれば1となる。レイアウトは縦長であれば1，横長でなければ0とするダミー変数を用いた。基本構造，構成要素，論文／非論文中の表現は，該当する語が出現していれば1，出現していなければ0とした。

4.3 実験用PDFファイル集合の構築

実験では，英語と日本語のそれぞれ20,000件のPDFファイル集合を，(1) PDFファイルのURLの収集，(2) URLに基づいたファイルのダウンロード，(3) 人手による学術論文の判定により，構築した。

4.3.1 PDFファイルのURLの収集

分野に偏りがない実験集合を構築するため，まず，検索エンジンのAPIを用いて，PDFファイルのURLを収集した。APIには，米国のサーチエンジンYahoo!のYahoo! Search BOSS[28]を用いた。Yahoo!を選択した理由は，(1) APIが公開されていること，(2)「PDF」というファイルタイプを指定した検索が可能なこと，(3) 単位時間あたりの検索数に制限が少ないこと，(4) 検索結果の取得数の制限が少ないことである。

PDFファイルは，分野に偏りがなく，できるだけ広範囲から収集するため，検索エンジンAPIで用いる検索語には，分野に偏りがない一般名詞を用いた。英語ファイルの収集には，WordNet3.0[29]に付属されている名詞句辞書（index.noun）に収録された117,797語句を用いた。複数語から構成される句も，検索語としてそのまま用いた。1検索式あたりの検索結果のうち，上位500件までのURLを収集した。検索結果が500件に満たないものは，全検索結果のURLを収集した。APIによる検索とURLの収集は2010年7月に実施した。収集したURL集合には重複が含まれていたため，重複を除いた結果，22,591,139件の重複のないURLを得た。なお，検索エンジンAPIの動作確認，検索結果の適切さを検証するために行った事前調査において，言語フィルタを用いることで，大半が英語で記述されているにも関わらず，ごく一部に他言語での表現が含まれたものが排除されたため，言語の指定は行わなかった。

日本語PDFファイルの収集には，日本語WordNet[30]とIPAdic[31]の両方に登録されている名詞27,384語を検索語として用いた。2010年12月に，上記の検索語を用い，ファイルタイプを「PDF」に限定し，言語の指定を「日本語」とし，URLを収集した。検索結果の上位1,000件までを取得し，重複を除去した結果，6,602,504の重複のないURLを得た。

次に，それぞれに収集したURL集合から無作為に3万件のURLを抽出し，英語ファイルは2010年8月に，日本語ファイルは2011年1月に，URLに基づきダウンロードした。日本語PDFファイル集合に関しては，URLのサブドメインが「.cn」「.tw」「.hk」「.kr」「.sg」のものは削除した。これらのサブドメインを持つファイルは中国語で書かれていたためである。

ダウンロードしたファイル集合に対し，Apache PDFBox1.2.1[32]を用いて，テキスト抽出を行った。一部分でもテキスト抽出ができたものは，英語ファイルでは27,848件，日本語ファイルでは27,158件であった。さらに，これらのファイル群からそれぞれ20,000件を無作為に抽出し，英語PDFファイル，日本語PDFファイルの実験用集合とした。以下では，それぞれ，英語PDF集合，日本語PDF集合と呼ぶ。

4.3.2 人手による学術論文の判定

英語と日本語PDF集合それぞれに対し，以下の判定規準をもとに，5名の判定者が，本論文で定めた学術論文であるか否かの確認を行った。

(1) 論文の形態をとっている
(2) タイトル，著者名，所属機関が明記されている
(3) 引用，参考文献がある
(4) 1論文が1ファイルで構成されている
(5) 2ページ以上である

判定者間での基準の統一をはかるために，いくつかのサンプルを用いたトレーニングセッションを行い，判定作業中も，判定者による判断が分かれたものに対しては合議し，基準の統一をはかった。また，一回目の判定で学術論文と判定された

表5　PDF集合の基本統計

	英語			日本語		
	論文	非論文	計	論文	非論文	計
ファイル数	2,011	17,989	20,000	587	19,413	20,000
平均文字数	50,152	48,220	48,414	37,821	24,666	25,052
平均ページ数	13.1	17.9	17.4	11.1	10.0	10.0
縦長レイアウトの比率	99.9%	93.7%	94.3%	99.5%	92.2%	92.4%

表6　英語PDF集合のドメイン分布

論文			非論文		
ドメイン	件数	比率（%）	ドメイン	件数	比率（%）
.edu	502	25.0%	.com	5,411	30.1%
.org	360	17.9%	.org	3,658	20.3%
.com	209	10.4%	.edu	1,967	10.9%
.uk	101	5.0%	.uk	1,070	6.0%
.gov	37	1.8%	.gov	939	5.2%
その他	802	39.9%	その他	4,936	27.5%
計	2,011	100%	計	17,981	100%

ものについて，他の判定者が改めて判定を行った。英語PDF集合に関しては，抄録や標題は英語であるが，本文が英語以外の言語で書かれた学術論文については，学術論文集合から除外した。

4.3.3　実験集合の特性

英語と日本語のPDF集合における論文と非論文のファイル数，ファイルサイズ，ページ数，縦型の割合を表5に示す。英語PDF集合の論文の割合は2,011件（10.1%）であり全体の1割程度しか論文が含まれていなかったが，日本語PDF集合における論文の割合は587件(2.9%)であり，それに比べると低い。日本語PDF集合の構築では，英語PDF集合の構築と比べて，用いた検索語が異なることや検索結果の上位500位までを取得したしたという違いはあるが，英語PDF集合に比べて，日本語PDF集合に占める論文の割合は低い。

縦長レイアウトの割合は，PDFファイルの1ページ目の縦の長さと横の長さを比較したときに縦が長いものの割合を示している。英語PDF集合も日本語PDF集合も，論文における縦長レイアウトの割合は99.9%と，ほとんどの論文が縦長レイアウトであった。その他，両言語とも文字数は論文のほうが大きいが，ページ数に関しては，英語については非論文のほうが長く，日本語では論文のほうが長かった。

表6は英語PDF集合中のURLのトップドメインの上位5位までのファイル数とその比率を論文と非論文別に示したものである。論文のドメインは.eduが多く，全体の1/4を占めている。次いで.org，.comの順になっており，この上位3ドメインで論文ファイルの53.3%を占めている。一方，非論文のドメインは.comが最も多く，次いで.org，.eduである。これらの3ドメインは英語PDF集合全体の中でも主要なドメインであるが，論文ではeduドメイン，非論文ではcomドメインが主のドメインであるといえる。表7に日本語PDF集合中のURLのセカンドレベルドメインの上位5位のファイル数と割合を論文と非論文別に示した（.comのみトップレベルドメイン）。論文のドメインの54.5%は.ac.jpであった。非論文ファイルのドメインは，.comの割合が最も高く14.3%であったが，それ以外には顕著な差はみられなかった。

論文，非論文の間でそれぞれファイルの属性やドメインの分布に違いがみられ，これらも論文を判定する一つの手掛かりとすることができるため，判定ルールに含めた。

表7 日本語PDF集合のドメイン分布

論文			非論文		
ドメイン	件数	比率（%）	ドメイン	件数	比率（%）
.ac.jp	320	54.5%	.com	2,775	14.3%
.go.jp	47	8.0%	.co.jp	2,434	12.5%
.or.jp	40	6.8%	.ac.jp	1,913	9.9%
.co.jp	28	4.8%	.or.jp	1,505	7.8%
.com	12	2.0%	.go.jp	1,299	6.7%
その他	140	23.9%	その他	9,487	48.9%
計	587	100%	計	19,413	100%

4.4. 実験に用いた判定器と評価尺度

4.4.1. 判定器

本研究では，複数の判定器を用いた分類実験を行い，その判定性能を比較した。具体的には，SVM，アダブースト（AdaBoost），ランダムフォレスト，ナイーブベイズ，決定木（C4.5）を用いた。実験にはWeka 3.5.6[33]を用いた。

(1) SVM

SVMは，ヴァプニク（Vapnik, Vladimir N.）によって提案された2クラス分類器の一種である[34]。高い汎化性能を持ち，カーネル関数を用いることにより非常に高次元のデータを扱うことができる点が特徴であり，投入する属性数が多くなりがちなテキスト分類において，多くの応用事例があり，もっとも性能が高いとされる。なお，本実験で用いたカーネルは，多項式カーネル（polynomial kernel）である。

(2) アダブースト（AdaBoost）

ブースティング（Boosting）法は，精度がそれほど高くない複数の弱学習器の重み付けを学習することで性能を高める手法である。アダブーストは初期のブースティング法を改良したもので，単語の有無による弱学習器をアダブーストによって組み合わせた分類器が最近傍法（k-NN法）やナイーブベイズ法による分類器よりも高い判定性能を示している[35]。本実験で組み合わせた弱分類器は決定株（decision stump），繰り返し数は100回である。

(3) ランダムフォレスト

集団学習（ensemble learning）であり，アダブーストと同様に精度がそれほど高くない複数の弱学習器の組み合わせ方，重み付けを学習することで性能を高める手法である[36]。本実験で，弱分類器として用いた決定木の数は100，各決定木が無作為に選択した素性数は5であった。

(4) ナイーブベイズ

ベイズ確率に基づく分類器である。この分類器を論文判定に応用した先行研究では，精度は低いが再現率が高いという他の分類器とは異なる結果を示した。

(5) 決定木（C4.5）[37]

決定木（decision tree）は可読性の高い分類器であり，近年ではアダブースト等の集団学習の弱学習器として使われることが多い。ここではWekaに実装されているC4.5（モジュール名はJ48）を用いた。

4.4.2. 評価尺度

評価尺度として，精度（P），再現率（R），F値（F）を用いた。F値はパラメータαの値によって，精度と再現率の重みを変えることができる。ここでは精度と再現率が同じ重みを持つもっとも一般的な$\alpha = 0.5$とした場合のF値を用いた。

本実験では，10点交差検定における各評価尺度の値を求め，それらを平均した値を算出した(macro-averaging)。

$$P = \frac{\text{システムが判定した正解件数}}{\text{システムが論文と判定した件数}}$$

$$R = \frac{\text{システムが判定した正解件数}}{\text{全論文件数}}$$

$$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}}$$

表8 PDF集合を対象にした自動判定結果

	英語PDF集合			日本語PDF集合		
判定器	精度	再現率	F値	精度	再現率	F値
アダブースト	0.72	0.53	0.61	0.67	0.35	0.46
決定木	0.76	0.65	0.70	0.65	0.38	0.48
ナイーブベイズ	0.41	0.84	0.55	0.30	0.78	0.43
ランダムフォレスト	0.79	0.70	0.74	0.68	0.44	0.53
SVM	0.73	0.54	0.62	0.72	0.10	0.18

表9 日本語PDF集合（論文割合補正後）の判定結果

判定器	精度	再現率	F値
アダブースト	0.80	0.60	0.69
決定木	0.73	0.64	0.69
ナイーブベイズ	0.60	0.81	0.69
ランダムフォレスト	0.77	0.65	0.71
SVM	0.79	0.60	0.69

4.5 実験結果

4.5.1 英語と日本語PDF集合を用いた結果

20,000件の英語PDF集合と日本語PDF集合を用いて，判定ルールの有効性を実験により検証した。判定ルールに基づき，先に示した5判定器を用いて，学術論文の自動判定を行った。実験は，10交差検定を用いた。それぞれの実験集合における結果を表8に示した。英語PDF集合に対する判定では，精度が最も高いのはランダムフォレストによる0.79である。再現率が最も高いのはナイーブベイズによる0.84である。F値でみるとランダムフォレストを用いた結果が0.74と最も高い。日本語PDF集合に対する判定では，英語に比べていずれの評価尺度でも値が低くなっているが，もっとも高い精度はSVMによる0.72，再現率はナイーブベイズの0.78，F値はランダムフォレストの0.53であった。

前報[3]の日本語PDF集合を対象に行ったルールベースでの判定実験では，最も高い性能は分類器投票（Vote）を用いたものであり，F値は0.49であった（精度0.44，再現率0.54）。これに比べると，すべての評価尺度で判定性能の向上がみられる。特に，英語PDF集合に関しては，7割を超える確率で学術論文を判定することができ，大きな改善が見られた。

4.5.2 補正した日本語PDF集合を用いた結果

日本語PDF集合に対する判定性能は，いずれの判定器においても，英語PDF集合より低くかった。この原因はいくつか考えられるが，英語と日本語の集合では，学術論文の割合が，それぞれ10.1%と2.9%であり，日本語PDF集合に占める論文の割合が低い。本実験では，ウェブ上に存在するPDFファイルの集合を，実験集合でも再現するため，論文比率の調整は行わなかったが，本節では，日本語PDF集合を英語PDF集合と同様の割合，つまり，論文10%，非論文90%の割合にした集合を作成し，再実験を行った。これを補正済日本語PDF集合と呼ぶ。非論文は，日本語PDF集合からランダムに選択した。

補正済日本語PDF集合に対する実験結果を表9に示す。英語PDF集合には及ばないものの，ランダムフォレストにおいてF値が0.71であり，補正前の日本語PDF集合に対するF値に比べて大きく向上した。日本語PDF集合を用いた結果（表8）と比較しても，すべての判定器において精度，再現率，F値ともに向上した。この結果から，日本語PDF集合の性能が低い要因のひとつとして，日本語PDF集合に占める論文の割合が大きく影響していることが考えられる。また，英語PDF集合と補正済日本語集合で同程度の結果

が得られたことから，この判定ルールは言語に依存しないことを示唆している。

4.6 判定ルールに関する考察

学術論文の構造や構成に基づいて構築した判定ルールを用いて，実際にウェブ上で公開されたPDFファイルからの論文の自動判定実験を行った結果，日本語PDF集合に関しては前報より大きく性能が向上し，英語PDF集合では，ランダムフォレストを用いたとき，精度，再現率ともに7割を超え，さらに高い性能が示された。また，英語PDF集合と論文比率を同様にした補正済日本語PDF集合を用いて実験したところ，英語集合に対して行った場合の性能とほぼ同水準の結果を得ることができた。この結果は，本研究で構築した判定ルールは，言語に依存しないルールであることを示している。

本研究は，判定ルールの構築目的であるため，判定ルールの精緻化に重点を置いた。しかし，判定性能の向上のためには，判定ルールに起因しない誤判定を分析することも考えられる。たとえば，PDFファイルからテキスト抽出を失敗している例である。具体的には，半角英数字の部分は抽出されているが日本語の部分は文字化けしている例，縦書きのものは一文字ずつ改行されている例等があった。このようなテキスト抽出の失敗は，PDFファイルの性質やテキストを抽出する際に用いるソフトウェアに依存する部分が大きく，これらは今後の課題として検討していく。

5. 結　論

本研究では，これまでに学術論文の形式に関して言及された文献や論文の書き方等の指南書から，学術論文の構成要素と構造を整理した。その結果，学術論文の構成要素として，標題，著者，所属，抄録，引用文献等があり，構造としてはIMRAD形式が普及していることが示された。その上で，これらの構成要素や構造について，ウェブ上に存在する学術論文を対象に調査したところ，英語で書かれた論文でも日本語で書かれた論文でも，多くの論文がこれらの要素，構造を持っていることが明らかになった。さらに，これらの構成要素，構造をもとにして学術論文を自動的に判定するルールを構築した。その際，分野に依存せず，かつまた，言語に依存しない判定ルール群を指向した。構築した判定ルールを，日本語と英語のそれぞれ20,000件のPDFファイル集合に対して適用したところ，英語PDF集合に関しては7割を超える性能で学術論文を判定できることが示された。

以上の結果から，学術論文の特徴を示す要素を見つけることができれば，少ない判定ルールでも，学術論文を自動的に判定することができる可能性を示すことができた。機械学習で一般的に行われている判定は，テキスト中の語を用いて統計的分析を行い，それらの特徴からテキストを自動的に判別している。本研究でのアプローチはこれとは異なるが，実験結果により，本アプローチの有効性が示されたといえる。つまり，学術論文であることを的確に表す特徴はテキスト中に出現する語の頻度だけで決まるわけではなく，テキストの形式や構造といった特徴を捉え，それらから学術論文を構成するための判定ルールを構築するというアプローチも有効であることが示された。また，これらの判定ルールは，英語と日本語だけでなく，他の言語に対しても比較的容易に適用することができる。

さらに，本研究のアプローチは，学術論文だけではなく，特定の形式に則ったかたちで記述されているテキストの判別にも適用できる可能性がある。

謝　辞

本研究はJSPS科研費21300095, 22500220の助成を受けたものです。

引用文献

1) 日本工業規格．JIS. Z. 8301.：2008.「規格票の様式及び作成方法」http://kikakurui.com/z8/Z8301-2011-01.html,（参照 2012-12-25）

2) Argamon, Shlomo., Whitelaw, Casey., Chase, Paul., Hota, Sobhan Raj., Garg, Navendu., Levitan,Shlomo. "Stylistic text classification using functional lexical features," *Journal of the American Society of Information Science*. Vol. 58, No. 6, 2007, p. 802-822.

3) 安形輝，池内淳，石田栄美，野末道子，久野高志，

上田修一「日本語学術論文PDFファイルの自動判定」『Library and Information Science』No. 56, 2006, p. 43-63.
4) 八杉龍一『論文・レポートの書き方』明治書院, 1971. 200p.
5) Ziman, John. M.『社会における科学（上）』松井巻之助訳, 草思社, 1981, 206p.
6) 飯田崇文「「学術論文の社会学」試論―「書簡」から「論文」へ：Philosophical Transactions（1740～1859）」『早稲田大学大学院文学研究科紀要第1分冊』No. 46, 2000, p. 59-65
7) Vickery, Brian. C.『歴史のなかの科学コミュニケーション』村主朋英訳, 勁草書房, 2002, 268p.
8) Harmon, Joseph. E. "The Structure of Scientific and Engineering papers: A Historical Perspective," *IEEE Transactions on Professional Communication*. Vol. 32, No. 3, 1989, p. 132-138.
9) Trelease, Sam. F. and Yule, Emma. S. *Preparation of Scientific and Technical Papers*. Williams & Wilkens, 1927, p. 117.
10) 久保猪之吉『医学論文の書き方　後篇　第1』久保猪之吉, 1929, [69]p.
11) Day, Robert. A. "The Origins of the Scientific Paper: The IMRAD Format," *American Medical Writers Association Journal*. Vol. 4, No. 2, 1989, p. 16-18.
12) Day, Robert. A. and Gastel, Barbara.『世界に通じる科学英語論文の書き方』美宅成樹訳, 丸善, 2010, 321p.
13) Sollaci, Luciana B. and Pereira, Mauricio G. "The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey," *Journal of Medical Library Association*. Vol. 92, No. 3, 2004, p. 364-367.
14) Eco, Umberto.『論文作法 調査・研究・執筆の技術と手順』谷口勇訳, 而立書房, 1991, 274p.
15) 澤田昭夫『論文の書き方』講談社, 1977, 259p.
16) 澤田昭夫『論文のレトリック』講談社, 1986, 330p.
17) 科学技術情報流通技術基準（SIST）SIST08：2010.「学術論文の執筆と構成」http://sist-jst.jp/pdf/SIST08_2010.pdf,（参照 2012-12-25）
18) Bazerman, Charles. "Modern evolution of the experimental report in physics: Spectroscopic articles in Physical Review, 1893-1980.," *Social Studies of Science*. Vol. 14, No. 2, 1984, p.163-196.
19) 倉田敬子『学術情報流通とオープン・アクセス』勁草書房, 2007, 196p.
20) Lin, Ling. "Evans, Stephen. Structural patterns in empirical research articles: A cross-disciplinary study," *English for Specific Purposes*. Vol. 31, No. 3, 2012, p. 150-160.
21) Argamon, Shlomo., Dodick, Jeff. and Chase, Paul. "Language use reflects scientific methodology: A corpus-based study of peer-reviewed journal articles." *Scientometrics*. Vol. 75, No. 2, 2008, p. 203-238.
22) Stamatatos, Efstathios., Fakotakis, Nikos D., and Kokkinakis, George K. "Text Genre Detection Using Common Word Frequencies," *Proceedings of the 18th conference on Computational linguistics (COLING '00)*, Vol. 2, 2000, p. 808-814.
23) Kumari, Pranitha K. and Reddy, A. Venugopal. "Performance Improvement of Web Page Genre Classification," *International Journal of Computer Applications*. Vol. 53, No. 10, 2012, p. 24-27.
24) Kanaris, Ioannis. and Stamatatos, Efstathios. "Learning to recognize webpage genres," *Information Processing and Management*. Vol. 45, No. 5, 2009, p. 499-512.
25) Lim, Chul Su., Leeb, Kong Joo. and Kima, Gil Chang. "Multiple sets of features for automatic genre classification of web documents." *Information Processing & Management*. Vol. 41, No. 5, 2005, p. 1263-1276.
26) Shepherd, Michael., Watters, Carolyn. and Kennedy, Alistair. "Cybergenre: automatic identification of home pages on the web," *Journal of Web Engineering*. Vol. 3, No. 3, 2004, p. 236-251.
27) Thapa, Chaman., Zajane, Osmar., Rafiei, Davood., and Sharma, Arya M. "Classifying websites into non-topical categories," *Proceedings of the 14th international conference on Data Warehousing and Knowledge Discovery (DaWaK '12)*, 2012, p. 364-377.
28) Yahoo! Developer Network. Yahoo! BOSS Search Services (Build your Own Search Service). http://developer.yahoo.com/boss /search/（参照 2012-12-28）.
29) Princeton University. "WordNet; A lexical database for English" http://wordnet. princeton.edu/（参照 2012-12-28）. WordNetは, 英語の概念辞書である。各語に品詞情報が付与されており, 語の概念関係も示されている。ダウンロードして利用することが可能である。
30) 独立行政法人情報通信研究機構. "日本語WordNet," http://nlpwww.nict.go.jp/ wn-ja/（参照 2012-12-28）上記のWordnetに着想を得た, 日本語版WordNetである。ライセンスを保持すれば利用可能であるが, 現時点で, 収録されている語数は少ない。
31) Sourceforge.jp. "IPAdic legacy," http:// sourceforge.jp/projects/ipadic/（参照 2012-12-28）. 形態素解析器Chasen用辞書であり, 形態素と品詞等の情報を持つ語彙表である。
32) The Apache Software Fundation. "Apache PDFBox — Java PDF Library," http:// pdfbox.apache.org/（参照 2012-12-28）. PDFファイルを操作するライ

ブラリであり，PDFからテキストを抽出することができる。
33) The University of Waikato. "Weka; Waikato Environment for Knowledge Analysis" http://www.cs.waikato.ac.nz/ml/weka/ (参照 2012-12-28).
34) Vladimir N. Vapnik. The nature of statistical learning theory, 2nd ed. New York, Springer, 2000, xix, 314p.
35) Schapire, Roberte. E. and Singer, Yoram. "BoosTexter: A Boosting-based System for Text Categorization," *Machine Learning*, Vol. 39, No. 2/3, 2000, p. 135-168.
36) Breiman, Leo. "Random Forests," *Machine Learning*, Vol. 45, No. 1, 2001, p. 5-23. DOI: 10.1023/A:1010933404324
37) Quinlan, J. Ross. C4.5: Programs for Machine Learning. San Francisco, CA, Morgan Kaufmann Publishers, 1993.

Automatic Detection of Scientific Papers Based on Their Structure and Elements

Emi ISHITA
Kyushu University
Teru AGATA
Asia University
Yosuke MIYATA
Teikyo University
Atsushi IKEUCHI
University of Tsukuba
Shuichi UEDA
Rikkyo University

In this paper, we develop rules for the automatic detection of scientific papers from PDF files on the Web. We inspected the structure and elements of scientific papers and observed that scientific papers typically have certain basic elements and an IMRAD format. We examined 1,172 scientific papers on the Web. The results indicate that the papers share common elements such as title, authors, keyword, and references and 40% of the papers, which have an explicit structure, have an IMRAD or a similar format. We develop rules for automatic detection of scientific papers using information based on their structure and elements obtained from the inspection process. The rules are evaluated using English and Japanese PDF collections, which were compiled by random sampling from the Web and consisted of 20,000 files each. Random forest classifier is performed and an F-value of 0.74 is obtained for English PDF files and 0.53 for Japanese PDF files. These results indicate that the rules developed using the approach given in this study can detect scientific papers from PDF files on the Web.