様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　３１　日現在

機関番号：　１５３０１
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３０３８４

研究課題名（和文）論文文書解析システムの実装と新しい電子図書館サービスに関する研究

研究課題名（英文）Study on Implementation of a Scholarly Document Analysis System and New Digital Library Services

研究代表者

太田　学（OHTA, Manabu）

岡山大学・自然科学研究科・教授

研究者番号：１０３２６０１９

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：本研究では、学術論文から書誌情報、図表、専門用語など様々なメタデータを抽出する論文文書解析システムを提案した。抽出したメタデータは、データベースと付き合わせて同定を行うことで、情報の関連付けなどに利用される。また、抽出したメタデータを利用した新しい電子図書館サービスについて検討し、タブレット端末による学術論文のオンライン閲覧の利便性を向上させるインタフェースを提案した。

研究成果の概要（英文）：The purpose of this study is to implement a document analysis system which extracts various metadata such as bibliographic information, figures, tables, and technical terms from scholarly papers. The extracted metadata are associated with relevant information by identifying the entities by referring to databases. This study also examined new digital library services using the extracted metadata and proposed a user interface to support browsing scholarly papers on tablets.

研究分野：　情報工学

キーワード：　電子図書館　文書解析　情報抽出　メタデータ　CRF　ウェブ　電子書籍　閲覧支援

１．研究開始当初の背景

社会の隅々まで文書の電子化が浸透したこともあり、Kindle のような電子書籍閲覧端末が急速に普及し、大学等でも機関リポジトリの構築が進むなど、インターネットアクセス可能な情報アーカイブが分散的かつ組織的に整備されるようになった。このような情報アーカイブへ効率よくアクセスするには、書誌情報等のメタデータの整備が不可欠であるため、書誌情報を含む様々なメタデータを文書から自動抽出する技術は、知的資産としての情報アーカイブ実現に必須の技術といえる。しかし研究開始当初、良質のメタデータが付与された電子文書を低コストで作成する技術はまだ成熟していなかった。そこで本研究では、このようなメタデータを低コストかつ高精度に自動抽出する文書解析システムを提案した。これは、電子図書館のような大規模な論文データベースの構築に利用できる他、大学が整備している機関リポジトリや小規模な学会等で電子文書のメタデータ付与を半自動的に行う場合にも活用できる。

学術論文の場合、重要なメタデータの一つに、論文題目、著者名、雑誌名、発行年等の要素から構成される書誌情報がある。書誌情報は、論文タイトルページに記載されているだけでなく、論文末尾の参考文献欄に集約されている。よって参考文献の書誌情報を正確に抽出することは、例えば文献を同定して当該文献へのリンクを生成するための重要な前処理となる。このような書誌情報抽出には、論文をレイアウト解析して末尾の参考文献欄の領域を抽出する技術や、参考文献文字列をパージングする技術が必要となる。研究開始当初、機械学習による先端的な技法を適用することで、これらの性能向上が図られていたが、実際のメタデータ作成業務に使用できるレベルには遠かった。さらに、電子図書館の利用者のために、どのようなメタデータを抽出してどのように活用すべきか十分には検討されていなかった。

２．研究の目的

本研究は、文献同定や論文閲覧支援のために、電子文書から様々なメタデータを抽出するための論文文書解析システムの実装を目指した。とりわけ学術論文の電子文書からメタデータとして参考文献情報を抽出し、書誌要素へ分解する。その基盤技術として、条件付き確率場（CRF）などの機械学習手法を適用し、高精度な自動抽出を低コストで実現する。さらに論文全体から、実験に関わる情報や専門用語等もメタデータとして抽出し、それを利用して電子書籍閲覧端末などによる文書のオンライン閲覧の利便性を向上させる。また、評価実験のために、実験に関わる図表や段落等の実験情報をマークアップした論文コーパスの整備を合わせて行う。

３．研究の方法

本研究は、学術論文からメタデータを抽出するための論文文書解析システムの開発と、抽出したメタデータを活用したタブレット端末における論文閲覧支援サービスの提案に分けられる。論文文書解析システムの開発では主に、学術論文の参考文献文字列から書誌情報を高精度かつ低コストで抽出する方法と、論文全体から実験に関わる図表などの実験情報を自動抽出する方法の実現に取り組んだ。よって、本研究の課題は大きく以下の三つにまとめられる。

(1) 参考文献文字列からの書誌情報抽出

一般に学術論文の末尾に記載される参考文献リストには、引用文献等が集約されており、その書誌情報を抽出、同定して、当該文献とのリンク生成等ができれば大変有用である。本研究では CRF により、論文中の参考文献文字列をトークン列に変換し、そのトークンの書誌要素を高精度に推定する方法を提案する。とりわけ、CRF の学習コストを低く抑えながら、高品質な書誌情報を獲得する方法について検討する。

(2) 論文からの実験情報抽出

本研究では、論文全体から実験に関わる図表や段落等の実験情報を抽出するとともに、抽出した表のデータをグラフに自動変換するなどの可視化について検討する。またその重要かつ有意義な副産物として、実験情報のアノテーションを含む評価用論文コーパスを整備する。

(3) 論文閲覧支援

学術論文から抽出したメタデータは、関連するウェブコンテンツとのリンク生成に利用可能で、そのようなリンクはオンラインでの論文閲覧支援となる。本研究では、学術論文から抽出した専門用語等を利用した論文閲覧支援インタフェースを提案し、それをタブレット端末である iPad で利用する論文ブラウザのプロトタイプを実装する。

４．研究成果

(1) 参考文献文字列からの書誌情報抽出

学術論文の参考文献欄に記載された参考文献文字列から、CRF によりその書誌情報を抽出する方法を提案した。提案手法は、参考文献文字列をまずトークン列に変換（トークン化）し、次に各トークンに書誌要素ラベルを付与することで書誌情報を抽出する。

例えば、

M. Ohta, R. Inoue, and A. Takasu, Empirical evaluation of active sampling for CRF-based analysis of pages," in Proc. of IEEE IRI 2010, 2010, pp.13-18.

という参考文献文字列をパージングして、

<Author>M. Ohta</Author>
<DC>, </DC>

<Author>R. Inoue</Author>
<DC>, </DC>
<DAND>and </DAND>
<Author>A. Takasu</Author>
<DC>, </DC>
<DS>“</DS>
<Title>Empirical evaluation of active sampling for CRF-based analysis of pages</Title>
<DE>,” </DE>
<Conference>in Proc. of IEEE IRI 2010</Conference>
<DC>, </DC>
<Year>2010</Year>
<DC>, </DC>
<DPP>pp.</DPP>
<Page>13-18</Page>
<D>.</D>
のように著者名や論文題目といった重要な書誌要素ラベルを付与することを目的とする。ここで D から始まるラベルは書誌要素を区切るデリミタに付与するラベルである。

実験では、以下の 3 種類の学術論文誌の論文の参考文献文字列を利用して、書誌情報抽出精度を評価した。

①情報処理学会論文誌（IPSJ）：4,574 件
②電子情報通信学会英文論文誌（IEICE-E）：4,497 件
③電子情報通信学会和文論文誌（IEICE-J）：4,787 件

まず参考文献文字列のトークン化において、個々の書誌要素に対応する文字列を過不足なく一つのトークンとして抽出することができるかどうかを評価した。実験では、情報処理学会論文誌（IPSJ）で 83%、電子情報通信学会英文論文誌（IEICE-E）で 90%、電子情報通信学会和文論文誌（IEICE-J）で 93%の参考文献文字列を過不足なくトークン列に分割できた。

次に、トークン化と書誌要素ラベル付与を行った後の、各学術論文誌における書誌情報抽出精度を表 1 にまとめる。なお、CRF の学習に用いた、書誌要素ラベル付きの参考文献文字列は、いずれの論文誌でも約 4,000 件である。表 1 に示すように、これらの論文誌では、90～94%の参考文献文字列から正しく全ての書誌情報を抽出できることを確認した。

表 1 参考文献書誌情報抽出精度

論文誌	IPSJ	IEICE-E	IEICE-J
抽出精度	90%	93%	94%

本研究ではさらに、トークン列に書誌要素ラベルを付与する CRF の学習データが少ない場合に、能動サンプリングと擬似学習データ、転移学習を利用して抽出精度を高める方法を提案した。能動サンプリングは、書誌情報抽出が困難なサンプルを優先的に学習することで、少量の学習データで高精度な抽出を実現する。既存のラベル付きデータから自動生成する擬似学習データを利用して、人が作成するコストの高い学習データを削減する。本研究の転移学習は、他雑誌のデータで学習した書誌情報抽出器（CRF）の書誌要素推定結果を、対象雑誌の書誌情報抽出に利用することを指す。抽出する書誌情報や参考文献文字列に表れる特徴には雑誌の種類によらない共通点があるため、これらの情報を間接的に対象雑誌の CRF 抽出器で利用することで書誌情報抽出精度の向上を図る。

実験の結果、能動サンプリングが少量データでの学習に極めて有効であること、擬似学習データを追加したり、また、他雑誌のデータで学習した CRF が推定した結果を利用したりすることで、少量学習データによる書誌情報抽出精度が向上することを確認した。少量学習データにおいて、さらに効果的に書誌情報抽出精度を向上させるには、各雑誌の参考文献文字列の書式などの類似点を精査する必要がある。

これらの知見から、多様な学術雑誌を扱う電子図書館でも、同じ体裁をもつ学術雑誌ごとに、能動サンプリングによって比較的少量の学習データから参考文献書誌情報を整備できることが分かる。そして、いくつかの雑誌の参考文献書誌情報が一定量整備されれば、それを学習データとしてそれらの雑誌の高精度な書誌情報抽出器が得られる。その高精度な抽出器に未整備の学術雑誌の参考文献書誌要素推定を手伝わせれば、その学術雑誌の参考文献書誌情報抽出の省力化になる。

(2) 論文からの実験情報抽出

本研究では、実験データや結果、評価指標など実験に関する情報が記載された図表や段落を実験情報と呼び、論文全体を解析してそれらを自動抽出する方法を検討した。これは、複数の論文から抽出した実験に関する情報を効果的に集約して論文閲覧者に提示すれば、有用な論文閲覧支援になるからである。

本研究ではルール等を用いて、まず図表等の論文構成要素を抽出し、その後論文構成要素を実験情報の図、表、段落等に分類することで、実験情報を抽出した。その結果、論文構成要素の抽出実験では、平均で F 値 0.94 となった。また、抽出した論文構成要素の分類による実験情報の抽出実験では、平均で F 値 0.78 となることを確認した。さらに抽出した実験情報の表を自動でグラフに変換する方法について検討し、実際に複数の表を棒グラフに自動変換した。評価実験により、グラフ化された実験情報の視認性の向上を確認するとともに、問題点なども把握した。

また、国立情報学研究所（NII）が主催する情報検索ワークショップ NTCIR9 の会議録論文約 100 件などに対し、図表、脚注、参考文献等を人手でマークアップした論文コーパスを整備した。本研究で検討した実験情報抽出は、主にこの論文コーパスにより評価実験を行った。

(3) 論文閲覧支援

本研究では、オンライン論文閲覧支援のため、学術論文から抽出した専門用語等を含む重要語と著者キーワードを word2vec により関連付けて、著者の意図に沿って重要語を組織化することを提案した。さらに、手掛かり語により重要語を Data や Method 等のカテゴリに分類するなどの論文閲覧支援を合わせて提案した。これらの機能を iPad で利用できるように実装した学術論文ブラウザのプロトタイプの画面を図１に示す。

図１では、左側の窓の左の列に著者キーワードの一覧が表示され、その右の列に選択された著者キーワードの関連重要語が表示されている。すなわち、ここでは著者キーワード一覧から“word2vec”が選択され、その右に関連重要語が 7 語表示されている。また、図１の右の“CATEGORY”のタブを選択すると、論文の全文から抽出した重要語が、Data や Method 等のカテゴリごとに色分けされて表示される。またその上の五つのタブは論文の節を表しており、いずれかを選択すると、その節に出てくる重要語を知ることができる。この論文ブラウザを使うことでユーザは、自身の興味に応じて重要語や著者キーワードに柔軟にアクセスできる。

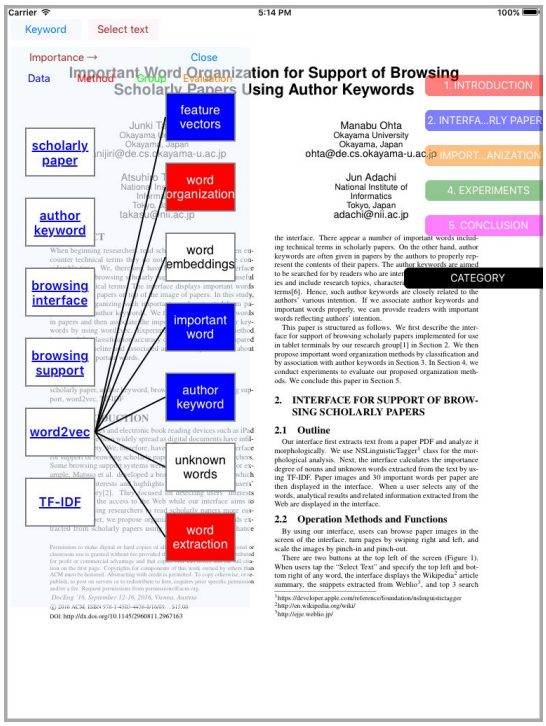

図 1 学術論文ブラウザのプロトタイプ

さらに、著者キーワードと関連付けた重要語が適切であるかどうか評価実験を行った。その結果、提案した方法では、著者キーワード 1 語に平均で 2.5 語の関連のある重要語と 2.2 語の関連のない重要語を関連付けることを確認した。よって、著者キーワードと重要語の関連付けについてはさらに改良が必要と考えている。また、学術論文ブラウザのプロトタイプを利用したユーザスタディによって、提案した論文閲覧支援インタフェースを多面的に評価することも今後必要である。

５．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計６件）

① Daiki Matsuoka, Manabu Ohta, Atsuhiro Takasu, and Jun Adachi, Examination of effective features for CRF-based bibliography extraction from reference strings, Proc. of 11th International Conference on Digital Information Management (ICDIM 2016), 査読有, 2016, pp. 259-264.
DOI: 10.1109/ICDIM.2016.7829774

② Junki Tanijiri, Manabu Ohta, Atsuhiro Takasu, and Jun Adachi, Important word organization for support of browsing scholarly papers using author keywords, Proc. of 16th ACM Symposium on Document Engineering (DocEng 2016), 査読有, 2016, pp. 135-138.
DOI: 10.1145/2960811.2967163

③ 川上尚慶, 太田学, 高須淳宏, 安達淳, 少量学習データによる参考文献書誌情報抽出精度の向上, 情報処理学会論文誌：データベース, 査読有, Vol. 8, No. 2, 2015, pp. 18-29.
https://ipsj.ixsq.nii.ac.jp/ej/index.php?action=pages_view_main&active_action=repository_view_main_item_snippet&index_id=1022&pn=1&count=20&order=7&lang=japanese&page_id=13&block_id=8

④ Naomichi Kawakami, Manabu Ohta, Atsuhiro Takasu, and Jun Adachi, Cost evaluation of CRF-based bibliography extraction from reference strings, Proc. of 16th International Conference on Asia-Pacific Digital Libraries (ICADL 2014), 査読有, LNCS 8839, 2014, pp. 268-278.
DOI: 10.1007/978-3-319-12823-8_28

⑤ Manabu Ohta, Daiki Arauchi, Atsuhiro Takasu, and Jun Adachi, Empirical evaluation of CRF-based bibliography extraction from reference strings, Proc. of 11th IAPR International Workshop on Document Analysis Systems (DAS 2014), 査読有, 2014, pp. 287-292.
DOI: 10.1109/DAS.2014.64

⑥ Atsuhiro Takasu and Manabu Ohta, Rule management for information extraction from title pages of academic papers, Proc. of Third International Conference on Pattern Recognition Applications and Methods (ICPRAM 2014), 査読有, 2014, pp.

438-444.
DOI: 10.5220/0004827204380444

〔学会発表〕（計２０件）
① 上田和也，新妻弘崇，太田学，学術論文の実験情報分類の評価，電子情報通信学会 2017 年総合大会，情報・システムソサイエティ特別企画学生ポスターセッション，2017.3.22，名城大学（愛知県名古屋市）．
② 浪越大貴，太田学，高須淳宏，安達淳，参考文献書誌情報抽出における確信度による CRF 学習データの削減，第 9 回データ工学と情報マネジメントに関するフォーラム（DEIM2017），2017.3.8，高山グリーンホテル（岐阜県高山市）．
③ 吉次優，太田学，高須淳宏，機械学習による学術論文の引用意図分類の一手法，電子情報通信学会研究会（データ工学と食メディア），2016.12.1，国立情報学研究所（東京都千代田区）．
④ 松岡大樹，太田学，高須淳宏，安達淳，CRF による参考文献書誌情報抽出のための辞書素性の拡充，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM2016），2016.3.2，ヒルトン福岡シーホーク（福岡県福岡市）．
⑤ 谷尻淳喜，太田学，高須淳宏，安達淳，著者キーワードを利用した学術論文閲覧支援の一手法，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM2016），2016.3.1，ヒルトン福岡シーホーク（福岡県福岡市）．
⑥ 内田裕太，太田学，高須淳宏，安達淳，学術論文からの参考文献文字列抽出の一手法，第 8 回データ工学と情報マネジメントに関するフォーラム（DEIM2016），2016.2.29，ヒルトン福岡シーホーク（福岡県福岡市）．
⑦ 吉次優，太田学，高須淳宏，手掛かり語による学術論文の引用意図分類の一手法，情報処理学会第 162 回 DBS 研究発表会，2015.11.26，芝浦工業大学（東京都江東区）．
⑧ 平井久貴，新妻弘崇，太田学，高須淳宏，学術論文からの実験情報抽出とその可視化，情報処理学会第 162 回 DBS 研究発表会，2015.11.26，芝浦工業大学（東京都江東区）．
⑨ 松岡大樹，太田学，高須淳宏，安達淳，CRF による参考文献書誌情報抽出のための有効な素性の検討と拡充，情報処理学会第 162 回 DBS 研究発表会，2015.11.26，芝浦工業大学（東京都江東区）．
⑩ 赤澤琢朗，太田学，高須淳宏，安達淳，CRF による様々な種類の学術論文からの参考文献文字列の自動抽出，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（DEIM2015），2015.3.4，ホテル華の湯（福島県郡山市）．
⑪ 樫本達矢，太田学，高須淳宏，学術論文からの構成要素抽出手法の改良，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（DEIM2015），2015.3.3，ホテル華の湯（福島県郡山市）．
⑫ 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（DEIM2015），2015.3.2，ホテル華の湯（福島県郡山市）．
⑬ 平井久貴，新妻弘崇，太田学，高須淳宏，学術論文からの実験情報抽出の一手法，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（DEIM2015），2015.3.2，ホテル華の湯（福島県郡山市）．
⑭ 川上尚慶，太田学，高須淳宏，安達淳，少量学習データによる参考文献書誌情報抽出，第 7 回 Web とデータベースに関するフォーラム（WebDB Forum 2014），2014.11.20，芝浦工業大学（東京都江東区）．
⑮ 前野明子，太田学，高須淳宏，学術論文閲覧支援インタフェースのための頭字語の活用，情報処理学会第 160 回 DBS・第 131 回 OS・第 35 回 EMB 合同研究発表会，2014.11.18，芝浦工業大学（東京都江東区）．
⑯ 平井久貴，新妻弘崇，太田学，CRF による学術論文からの実験情報抽出の一手法，電子情報通信学会 2014 年総合大会，情報・システムソサイエティ特別企画学生ポスターセッション，2014.3.20，新潟大学（新潟県新潟市）．
⑰ 川上尚慶，太田学，高須淳宏，安達淳，CRF による参考文献書誌情報抽出のための学習コストの削減，第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM2014），2014.3.4，ウェスティンホテル淡路（兵庫県淡路市）．
⑱ 石本茜，太田学，高須淳宏，安達淳，CRF による学術論文からの参考文献文字列の抽出，第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM2014），2014.3.4，ウェスティンホテル淡路（兵庫県淡路市）．
⑲ 樫本達矢，太田学，高須淳宏，学術論文からの構成要素抽出の一手法，第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM2014），2014.3.4，ウェスティンホテル淡路（兵庫県淡路市）．
⑳ 前野明子，太田学，高須淳宏，学術論文閲覧支援インタフェースの試作，第 6 回データ工学と情報マネジメントに関するフォーラム（DEIM2014），2014.3.3，ウェスティンホテル淡路（兵庫県淡路市）．

〔図書〕（計１件）
① Atsuhiro Takasu and Manabu Ohta, Springer, Pattern Recognition Applications and Methods (Chapter: Utilization of multiple sequence analyzers for bibliographic information extraction), 2015, pp. 222-236.

〔その他〕
受賞

① 第 7 回 Web とデータベースに関するフォーラム（WebDB Forum 2014）学生奨励賞，少量学習データによる参考文献書誌情報抽出，川上尚慶，2014.11.20.
② 第 6 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 2014）学生プレゼンテーション賞，CRF による参考文献書誌情報抽出のための学習コストの削減，川上尚慶，2014.3.4.

６．研究組織

(1)研究代表者

太田　学（OHTA Manabu）
岡山大学・自然科学研究科・教授
研究者番号：10326019

(2)研究分担者

なし

(3)連携研究者

なし

(4)研究協力者

なし