様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １８ 日現在

機関番号：６２６１５
研究種目：基盤研究(B)
研究期間：2012～2014
課題番号：２４３０００９７

研究課題名（和文）機械学習による統合的書誌メタ情報編集システムの実装

研究課題名（英文）Implementation of an Integrated System for Editing Meta Data Using Machine Learning Techniques

研究代表者
安達　淳（ADACHI, Jun）

国立情報学研究所・コンテンツ科学研究系・教授

研究者番号：８０１４３５５１

交付決定額（研究期間全体）：（直接経費）　　13,000,000円

研究成果の概要（和文）：本研究は、統合的なメタ情報編集の環境として、電子文書の中からメタ情報を抽出し、その編集とともに他のデータベースと付き合わせて同定を精度良く行うためのシステムの実装を目指した。そのため、学術論文をレイアウト解析し、メタ情報として書誌および引用文献情報を高精度かつ低コストに抽出する方法を提案した。また書誌情報抽出法の評価などのために、学術論文の書誌メタ情報アノテーション付き参考文献文字列コーパスを作成した。

研究成果の概要（英文）：The purpose of this study is to implement a system which extracts meta data from digital documents to edit and identify the meta data accurately by matching against other databases as an integrated environment for editing such meta data. We proposed accurate and inexpensive methods to extract bibliographic and citation information as meta data from research papers of which the layout was analyzed. We also created several reference string corpora of research papers by annotating the reference strings with bibliographic information for evaluating bibliographic information extraction methods.

研究分野：情報工学

キーワード：書誌パージング　メタデータ　コーパス　CRF　機関リポジトリ　テキストマイニング

１．研究開始当初の背景

電子書籍閲覧端末が急速に普及する背景には、社会の隅々まで文書の電子化が浸透したことがある。大学等では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。一方、電子文書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とその組織化が不可欠である。しかしその手間がかかることから、本研究の開始当初、良質のメタ情報が付与された電子文書の作成技術はまだ定着していなかった。従って、書誌情報を含む様々なメタ情報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のための核となる技術といえる。またこのようなメタ情報抽出システムは、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情報付与処理を半自動的に行う場合にも活用できる。

学術論文の場合、メタ情報には、論文題目、著者名、雑誌名、発行年などの要素から構成される書誌情報が該当する。また、情報の組織化には、個々の引用文献に対する当該論文へのリンクのようなメタ情報も重要である。これらのメタ情報の作成を自動化するには、文字列のパージングや書誌レコードの同定技術が必須である。またそれ以前に、論文の PDF ファイルのページのレイアウトを解析し、タイトルページや末尾の参考文献欄に記載されている書誌情報領域を抽出する必要がある。本研究開始当初、機械学習によるデータ同定や抽出に関す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、実際のデータ作成業務に使用できるレベルには遠かった。特に、機械学習を用いて、雑誌ごとのレイアウトの相違にも頑健で高精度の抽出手法を実用化するのは難しく、個々の雑誌のレイアウトに即したテンプレートなどを作成するレベルに止まっていた。

２．研究の目的

本研究は、実用的な統合的メタ情報編集環境システム（Metadata Workbench、以下 MWB）として、電子文書の中からメタ情報を抽出し、その編集とともに他のデータベースと付き合わせて同定を行うためのシステムの実装を目指した。特に学術論文のような電子文書からメタ情報として書誌および引用文献情報を抽出し、書誌要素への分解などを行う。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、高精度かつ低コストな自動抽出を実現する。システム的には抽出誤りに対する実用的な対処が可能な柔軟なシステムを実装することを特徴とする。また、システムの汎用性を担保するために、学習と評価のためのデータベースとして参考文献文字列コーパスの整備を合わせて行う。

３．研究の方法

本研究では、学術論文から書誌情報を高精度かつ低コストで抽出する方法の実現と、それをモジュール化し、レイアウト解析や文献レコード同定など他のモジュールと有機的に結合した MWB の開発に主に取り組んだ。MWB の開発では、システム評価のテストベッドに必要な参考文献文字列コーパスを合わせて整備した。なお書誌情報抽出は、論文のタイトルページからの書誌情報抽出と、参考文献文字列からの書誌情報抽出に分けて研究をすすめた。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) 論文のタイトルページからの書誌情報抽出

一般に、学術論文のタイトルページには、論文題目や著者名などの書誌情報が、論文誌ごとに決まったレイアウトで書かれている。本研究では、論文タイトルページの文書画像を OCR で解析して得られる各テキスト行の書誌要素を、CRF により推定して抽出する。そこで、このとき避けられない抽出誤りを確信度によって検出し、人手で確認する後処理のコストを削減する方法について検討した。

(2) 論文の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は、関連文献が集約されており、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して、後処理を含む全体の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を提案した。

(3) MWB の開発

実現した書誌情報抽出アルゴリズムや既存のソフトウェアを利用して、学術論文の電子文書からユーザが必要な情報を自由に抽出して編集できる統合的なソフトウェア環境を実装した。またその重要かつ有意義な副産物として、書誌メタ情報のアノテーションを含む評価用参考文献文字列コーパスを整備した。

４．研究成果

(1) 論文のタイトルページからの書誌情報抽出

本研究では、学術論文のタイトルページの文書画像を、OCR でレイアウト解析および文字認識して得られる XML ファイルを CRF への入力とし、CRF によって各行に書誌要素ラベルを付与することでその書誌要素を抽出する。具体的には以下のような手順となる。

①論文タイトルページの各行を、長さ、幅、隣接する行との距離などを特徴とする特徴ベクトルで表す。

②チェーンモデルのCRFのモデルを書誌要素ラベル付きの行の列より求める。
③書誌要素ラベルのついてない論文の行の列に、②で求めたCRFにより書誌要素ラベルを付与する。
④同一ラベルをもつ連続する行をまとめることによって、複数行にわたるタイトルや著者等の領域を抽出する。

実験では以下の 3 種類の学術論文誌の論文データを利用して、書誌要素抽出精度を評価した。なおこのデータ作成に用いられたOCRの文字認識精度は97～99%であった。
①情報処理学会論文誌(IPSJ)：479 件
②電子情報通信学会英文論文誌(IEICE-E)：473 件
③電子情報通信学会和文論文誌(IEICE-J)：174 件

論文タイトルページにある全ての書誌要素を正しく抽出できる論文の割合を表す、書誌情報抽出精度を表1に示す。表1から、300件の論文を学習データとして用いると 94～96%、学習データ件数が 100 件では 80～92%の抽出精度であることが分かる。

表 1 書誌情報抽出精度（タイトルページ）

学習データ件数	20	100	300
IPSJ	83%	92%	94%
IEICE-E	70%	90%	96%
IEICE-J	66%	80%	-

本研究では、CRF による抽出誤りを人手で修正するための後処理のコスト削減を図った。具体的には、CRF による書誌情報抽出結果に確信度を定義し、この確信度が低い論文は抽出誤りを含む可能性が高いと見なして検出する方法を検討した。CRF による書誌情報の自動抽出後に、最終的な書誌情報の抽出精度として 99%を実現するために、人が確認すべき確信度の低い論文の全論文に対する割合を表2に示す。

表 2 後処理コスト（データの割合）

学習データ件数	20	100	300
IPSJ	45%	18%	10%
IEICE-E	80%	52%	11%
IEICE-J	98%	52%	-

表 2 から、例えば情報処理学会論文誌(IPSJ)では、学習データ件数を300とすると、CRF による自動抽出後に確信度の低い 10%の論文を人が確認し誤りがあれば訂正するという後処理によって、99%の書誌要素抽出精度が実現できることが分かる。また学習データ件数が300の場合、電子情報通信学会英文論文誌(IEICE-E)についてもほぼ同様の結果となった。しかし、学習データ件数が少なく、表1に示すCRFによる書誌情報抽出精度が低い場合は、99%という精度を達成するために、半数以上の論文を人が事後に確認しなければならないことがあることが分かる。本研究は、雑誌ごとに300件程度の学習データがあれば、全論文の10%程度を人が自動抽出後に確認すれば、99%という高い精度が実現できることを実験によって示した点に意義がある。

(2) 論文の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。ただしこれらの参考文献文字列は文書画像から抽出したものではないので、OCRの文字認識誤りは含まれない。
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 3 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 3 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 3 書誌情報抽出精度（参考文献文字列）

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

電子情報通信学会和文論文誌において、論文タイトルページと同様に確信度に基づく後処理コストを評価したところ、CRF による書誌情報抽出後に、およそ 1/4 の参考文献文字列を人が確認すれば、99%の精度を実現できることが分かった。この確信度による抽出誤りの検出精度をさらに高められれば、学術論文の参考文献欄のための実用的な書誌情報抽出・編集システムが実現できる。

本研究ではさらに、書誌要素ラベル付与を行う CRF の学習データが少なくても、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、確信度の低い、つまり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抽出を実現する。擬似学習データは自動生成するため、これが有効であれば、人が作成するコストの高い学習データを削減できる。本研究で言う転移学習は、他雑誌用の書誌情報抽出器（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 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、他雑誌の学習データで学習した 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、多様な学術論文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。高精度の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ため、これは MWB の運用上好ましい。

(3) MWB の開発

本研究では、実用的な成果物として MWB を開発した。開発した MWB は、学術論文のレイアウト解析、参考文献文字列のパージング、サポートベクトルマシン（SVM）による文献レコード同定などのモジュールを有機的に結合したもので、文字列編集のソフトウェアと連携して動作する。さらに外部発表向けに図１に示すような MWB のデモシステムを実装し、情報処理学会など一部学会の電子図書館の持つ実データのメタ情報の編集を試みた。図１では、入力された論文 PDF のタイトルページと参考文献欄を抽出、解析して、論文題目などを自動抽出している。参考文献欄については、文字列パージングによる書誌情報抽出のみならず、学術情報データベースを利用して文献同定を行い、同定候補を表示する機能を実現している。

図 1 MWB デモシステムの画面

一方、書誌情報抽出性能の評価などのために、書誌メタ情報のアノテーション付きの参考文献文字列コーパスを作成した。整備したコーパスは、電子情報通信学会、情報処理学会、日本物理学会、日本アレルギー学会、IEEE の和文または英文の論文誌などの論文の引用文献情報であり、合計で約 2 万 6 千件に上る。学術的観点からこのコーパスは、研究コミュニティなどにおいて共有することにより、他の研究グループが同一コーパスに基づき性能を評価し、競争的に性能の向上を図ることを期待している。そのため、今後性能のより高い手法が開発されれば、それを MWB に移植し、随時システムの性能向上を行えるという点でも意義がある。

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８件）

① 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

② 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

③ 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

④ 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

⑤ Manabu Ohta, Ryohei Inoue, and Atsuhiro Takasu, Empirical evaluation of CRF-based bibliography extraction from research papers, IADIS International Journal on Computer Science and Information Systems, 査読有, Vol. 7, No. 2, 2012, pp. 18-31.
http://www.iadisportal.org/ijcsis/papers/2012150102.pdf

⑥ Manabu Ohta, Daiki Arauchi, Atsuhiro Takasu, and Jun Adachi, Error detection of CRF-based bibliography extraction from reference strings, Proc. of 14th International Conference on Asia-Pacific Digital Libraries (ICADL 2012), 査読有, LNCS 7634, 2012, pp. 229-238.
DOI: 10.1007/978-3-642-34752-8_29

⑦ 太田学，井上諒平，高須淳宏，CRFによる学術論文タイトルページからの書誌情報抽出における誤り検出，日本データベース学会論文誌，査読有，Vol. 11，No. 2，2012，pp. 37-42.
http://dbsj.org/journal/dbsj_journal/dbsj_journal_vol_11_no_2_37_42/

⑧ Manabu Ohta and Atsuhiro Takasu, A document analysis system for linking cross-document entities, Proc. of Fourth International Conference on Creative Content Technologies (CONTENT 2012), 査読有, 2012, pp. 14-20.
http://www.thinkmind.org/index.php?view=article&articleid=content_2012_1_30_60066

〔学会発表〕（計１４件）

① 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）.

② 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）.

③ 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）.

④ 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）.

⑤ 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）.

⑥ 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）.

⑦ 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）.

⑧ 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）.

⑨ 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）.

⑩ 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）.
⑪ 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）.
⑫ 樫本達矢，荒内大貴，太田学，データ工学に関する学術論文からの実験情報抽出の試み，電子情報通信学会 2013 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，2013.3.21，岐阜大学（岐阜県岐阜市）.
⑬ 川上尚慶，荒内大貴，太田学，高須淳宏，安達淳，文献種類別に分類した参考文献文字列からの書誌情報抽出の一手法，第 5 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 2013），2013.3.5，ホテル華の湯（福島県郡山市）.
⑭ 荒内大貴，太田学，高須淳宏，安達淳，CRF による和英文の参考文献文字列からの自動書誌要素抽出，情報処理学会第 156 回データベースシステム研究発表会，2012.12.12，キャンパスプラザ京都（京都府京都市）.

〔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.
③ Best Paper Award at Fourth International Conference on Creative Content Technologies (CONTENT 2012), A document analysis system for linking cross-document entities, Manabu Ohta and Atsuhiro Takasu, 2012.7.27.

６．研究組織
(1)研究代表者
安達　淳（ADACHI JUN）
国立情報学研究所・コンテンツ科学研究系・教授（副所長）
研究者番号：80143551

(2)研究分担者
太田　学（OHTA MANABU）
岡山大学・自然科学研究科・教授
研究者番号：10326019

(3)連携研究者
高須　淳宏（TAKASU ATSUHIRO）
国立情報学研究所・コンテンツ科学研究系・教授
研究者番号：90216648