科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　２８　日現在

機関番号：　１５３０１
研究種目：　若手研究(B)
研究期間：　2011～2013
課題番号：　２３７００１１９

研究課題名（和文）文書画像とウェブを活用した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 New Digital Library Services Using Document Images and the Web

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）　　3,400,000円、（間接経費）　　1,020,000円

研究成果の概要（和文）：学術論文を蓄積する電子図書館では、論文中の書誌情報などを自動抽出する技術が求められる。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や、参考文献文字列をトークン列に変換して得られる各トークンが、いずれの書誌要素に該当するか、条件付確率場(CRF)により推定して抽出する方法を提案した。また論文から抽出した専門用語を利用した関連論文推薦サービスを提案し、電子書籍閲覧端末による学術論文閲覧支援方法を検討した。

研究成果の概要（英文）：Digital libraries which store academic papers require some techniques for automatic extraction of bibliographic and other information from the papers. This study proposed a method of automatically extracting bibliographic information from OCRed document images of title pages of academic papers and from reference strings listed at the end of papers by using a conditional random field (CRF). This study also proposed a recommendation service of related papers using the technical terms extracted from academic papers and investigated ways to support online-browsing of academic papers using tablet PCs.

研究分野：　情報工学

科研費の分科・細目：　情報学・メディア情報学・データベース

キーワード：　電子図書館　文書画像　ウェブ　情報抽出　CRF　電子書籍　閲覧支援　論文推薦

１．研究開始当初の背景

iPad や Kindle のような電子書籍閲覧端末が急速に社会に普及する一方で、Google Books のような書籍の大規模な電子化が進んでいる。ここで求められているのは、単なる電子化ではなく、世の中にある紙ベースの文書をまるごと電子化して、世界中のどこからでもアクセス可能な情報アーカイブを作る技術である。電子図書館はこの情報アーカイブを代表するものであり、電子図書館では電子化した書籍の書誌情報は不可欠である。よって、書誌情報を含む様々な情報を文書から自動抽出する技術は、このような情報アーカイブ実現のための核となる技術といえる。

学術論文の文書画像を利用する電子図書館では、図書のすべての紙面をスキャンしてユーザの閲覧に供するとともに、検索用途に用いるため、論文題目、著者名、キーワード、発行年などの書誌情報をデータベースに入力する必要がある。これらの書誌情報は、通常論文のタイトルページや参考文献に記載されているが、学術論文からの自動抽出技術は確立されておらず、本研究開始当初、膨大な人的コストをかけて作成しているのが実情であった。また多くの電子図書館がウェブ経由でアクセスできるにも関わらず、組織外部のウェブとのリンクが十分とは言い難かった。

２．研究の目的

本研究は、光学式文字読取装置(OCR)や既存のレイアウト解析技術などを用いて処理した様々な種類の論文誌の学術論文から、OCR の認識誤りやレイアウト解析の誤りに頑健で、高精度かつ汎用的に書誌情報を自動抽出できるシステムの開発を目的とした。また、論文から抽出した情報を適切なウェブコンテンツと関連付ける方法を提案して、論文閲覧支援の観点から新しい電子図書館サービスの創出を目指した。

３．研究の方法

本研究は大きく、学術論文からの情報抽出法の開発と、抽出した情報を活用した論文閲覧支援サービスの提案に分けられる。本研究では、情報抽出をさらに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、本研究で扱う課題は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、論文誌毎に決まったレイアウトで書かれている。そこで本研究では、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方法を提案した。

(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。そこで本研究では、論文中の参考文献文字列をトークン列に変換して得られる各トークンの書誌要素を、CRF により推定して抽出する方法を提案した。

(3) 論文閲覧支援方法の提案

学術論文から抽出した情報は、当該論文と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。そこで本研究では、学術論文から抽出した専門用語を利用した論文閲覧支援サービスの提案と、その機能をもつオンライン学術論文ブラウザのプロトタイプを実装した。さらに、急速に普及する電子書籍閲覧端末による学術論文閲覧支援の方法についても検討した。

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方法を提案した。具体的には以下の手順で抽出する。

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。

以下の 3 種類の学術論文誌の論文データを利用して、まず書誌要素抽出精度を評価した。

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

なおこのデータ作成に用いられた OCR の文字認識精度は、アブストラクトの部分で 99%、参考文献の部分で 97%であった。論文タイトルページからそこ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、書誌要素抽出精度を表１に示す。表にある通り、300 件の論文を学習データとして用いると、94～96%程度の論文タイトルページから正しく全

ての書誌要素を抽出できることが分かる。しかし学習データ件数が 100 件では、80～92%の論文タイトルページからしか全ての書誌要素を正しく抽出できていない。

表 1 書誌要素抽出精度（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%	-

次に、CRF による抽出誤りを人手で修正するためのコストの削減について検討した。本研究では、CRF による書誌要素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判断して、検出する方法を検討した。表 2 に、この確信度が低い論文を人手で何件確認すれば、CRF による自動抽出と合わせて最終的な書誌情報の精度として 99%が実現できるか調査した実験結果を示す。実験に使用したデータは、表 1 に示した書誌要素抽出実験と同じものである。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（学習に用いる論文数）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を示している。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。ただし、学習データ件数が少なく、表 1 に示す CRF による書誌要素抽出精度が低い場合は、99%という精度は達成するには、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も多い。本研究は、全論文の 10%程度を人が確認するという後処理コストで、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。

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%	-

今後の課題としては、一般に学術論文のレイアウトは論文誌ごとに異なるので、汎用性を持たせるために複数の抽出器を用意し、これらを効率的に連携させる方法の検討などがある。

(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 によりその書誌要素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要素を抽出する。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（タグ）を付与することを目的とする。ここで<D*>は書誌要素を区切るデリミタに付与するラベル（タグ）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、書誌要素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要素抽出精度を表 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 書誌要素抽出精度（参考文献文字列）

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

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要素を抽出できることを確認した。

また、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 による書誌要素抽出後に、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。

(3) 論文閲覧支援方法の提案

本研究では、オンライン論文閲覧支援の一つの方法として、学術論文から抽出した専門用語を利用して、その関連論文を推薦する手法を提案した。具体的には、閲覧論文の論文題目とアブストラクトから抽出した各専門用語で検索される論文集合と、それらの論文の論文題目とアブストラクトに出現する専門用語集合からなる二部グラフを生成し、HITS アルゴリズムを利用してこの二部グラフのリンク解析を行い、推薦する関連論文を得た。この機能を実装したオンライン学術論文ブラウザのプロトタイプを図 1 に示す。

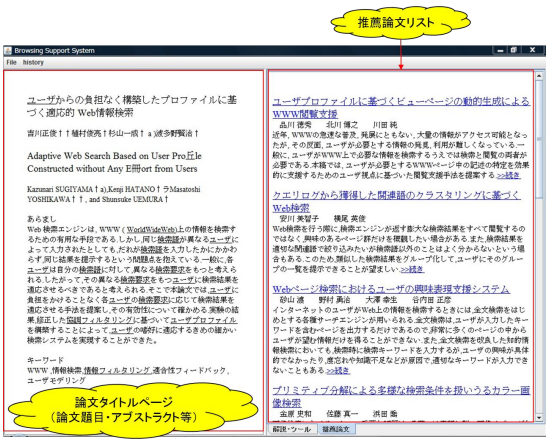

図 1 学術論文ブラウザのプロトタイプ

図 1 では、左側に論文タイトルページの文書画像を OCR で処理して得たテキストから抽出した論文題目などの主要な書誌要素、右側にランク付けられた関連論文のリンクリストが表示されている。要するにこれは、左側の論文を読んでいるユーザに、右側で関連論文を推薦するブラウザである。また、提案手法による関連論文の推薦精度を、情報検索で一般的なベクトル空間モデルに基づく論文推薦のそれと比較することで、適切な論文を推薦していることを確認した。

また、実験データや実験結果、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、論文全文を解析して自動抽出する方法を検討した。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ユーザに提示できれば、有用な論文閲覧支援になると考えている。さらに本研究では、タブレット PC 等の電子書籍閲覧端末に適した学術論文の閲覧方法についても検討した。

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９件）

① 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

② 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/

③ 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

④ Manabu Ohta, Ryohei Inoue, and Atsuhiro Takasu, Empirical evaluation of CRF-based bibliography extraction from research papers, Proc. of IADIS International Conference Information Systems 2012 (IS 2012), 査読有, 2012, pp. 18-26.
http://www.is-conf.org/

⑤ Manabu Ohta, Toshihiro Hachiki, and Atsuhiro Takasu, Related paper recommendation to support online-browsing of research papers, Proc. of Fourth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2011), 査読有, 2011, pp. 130-136.
DOI: 10.1109/ICADIWT.2011.6041413

〔学会発表〕（計１２件）

① 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, 兵庫.

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, 兵庫.

③ 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，兵庫．
④ 井上諒平，太田学，高須淳宏，CRFによる論文文書画像の書誌要素推定における自動誤り検出，第4回Webとデータベースに関するフォーラム (WebDB Forum) 2011，2011.11.5，東京．
⑤ 荒内大貴，太田学，高須淳宏，安達淳，CRFによる参考文献文字列からの書誌要素抽出の一手法，第4回Webとデータベースに関するフォーラム (WebDB Forum) 2011，2011.11.5，東京．

〔その他〕
受賞
① Outstanding Paper Award at IADIS International Conference Information Systems 2012 (IS 2012), Empirical evaluation of CRF-based bibliography extraction from research papers, Manabu Ohta, Ryohei Inoue, and Atsuhiro Takasu, 2012.3.12.
② 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
太田　学（OHTA MANABU）
岡山大学・大学院自然科学研究科・教授
研究者番号：10326019

(2)研究分担者
なし

(3)連携研究者
なし