脚注表示機能を備えた論文閲覧システム Sidenoter

阿辺川 武　　相澤 彰子
国立情報学研究所
{abekawa, aizawa}@nii.ac.jp

1　はじめに

近年，学術文献の電子化は大きく進み，投稿から出版まで紙を経由することなく流通することが一般的になってきた．一部の海外学術出版社では独自の XML フォーマットを定義し，1 つの XML ファイルから紙の印刷物や PDF，EPUB といった電子フォーマットへ変換する 1 ソースマルチユースな出版工程を実現している．日本でも科学技術振興機構 J-STAGE 3 において，本文 XML[1] の入稿を推奨するようになるなど着実に XML 化が進んでいる．論文が XML フォーマットで保存されれば，XHTML に変換し Web ブラウザで自由な文字サイズで閲覧でき，PMC(PubMed Central) の PubReader[2] のようなデスクトップにもモバイルにも対応した専用ソフトで快適に閲覧できるようにもなる．

しかし，多くの学術出版では，電子化といっても紙に印刷可能な PDF ファイルのみを生成し，公開時には PDF ファイルをそのまま配布するにとどまっている．もともと PDF は印刷することを考慮して策定されたフォーマットのため，どのような環境でも同一のレイアウトを実現できるが，一方で画面の小さな端末に対して，カラム数を変更したり，文字サイズを変更するなどカスタマイズして表示することができない．また，電子化される以前に発表された論文についても，画像としてスキャンし，PDF に変換して電子的に配布するため，やはり PDF フォーマットの論文が数多く流通しているのが実情である．

我々は現在，論文閲覧システムを開発しているが，このような現状では PDF で配布される論文も対象とせざるを得ず，レイアウトの再現性重視のファイル形式をいかに扱うかが課題となる．幸い本文テキスト自体は，電子的に制作された PDF からは比較的容易に抽出でき，スキャンした画像からも OCR を用いて 100%の精度ではないが抽出できる．本システムでは，論文のレイアウト重視の制約を逆に利用し，論文本文から得られる補足的な情報をページレイアウト上にそのまま表示する手法を開発した．

言語処理学会では，20 周年記念事業の一貫として 2013 年 7 月に過去の年次大会の予稿集を学会 Web ページ上で公開 [3] しており，本システムは，学術文献の閲覧システムのケーススタディとして，この言語処理学会の年次大会の予稿集を閲覧するものとして作成された [4]．特に年次大会のような学会の会議に参加し，聴講中の予稿集の閲覧を支援するシステムになることをめざしている．

本稿では，我々が開発している文献閲覧システム Sidenoter が備える機能および言語処理学会年次大会予稿集のデータ処理について説明する．

2　システムの持つ機能

本システムの基本構造は，PDF で配布される論文を画像に変換し，Web ブラウザ上で閲覧する仕組みからなる．図 1 は，そのスクリーンショットである．

2.1　閲覧機能

- 連続ページめくり
  紙の予稿集の冊子と同様にページ順に予稿を閲覧できる．従来は 1 つの予稿につき 1 つの PDF ファイルであるので，複数の予稿を連続して閲覧することは不可能であった．本システムでは検索時にヒットした予稿に対しても検索結果順に連続して閲覧できるため，検索結果と PDF を行きつ戻りつする手間を軽減できる．

- 拡大縮小表示
  Web ブラウザの画面サイズやモバイルデバイスの画面解像度に応じて，適切な解像度の文書画像が

[1] https://www.jstage.jst.go.jp/pub/html/AY04S230_ja.html
[2] http://www.ncbi.nlm.nih.gov/pmc/about/pubreader/
[3] http://www.anlp.jp/guide/nenji.html
[4] http://kmcs.nii.ac.jp/nlp_annual/ で公開中

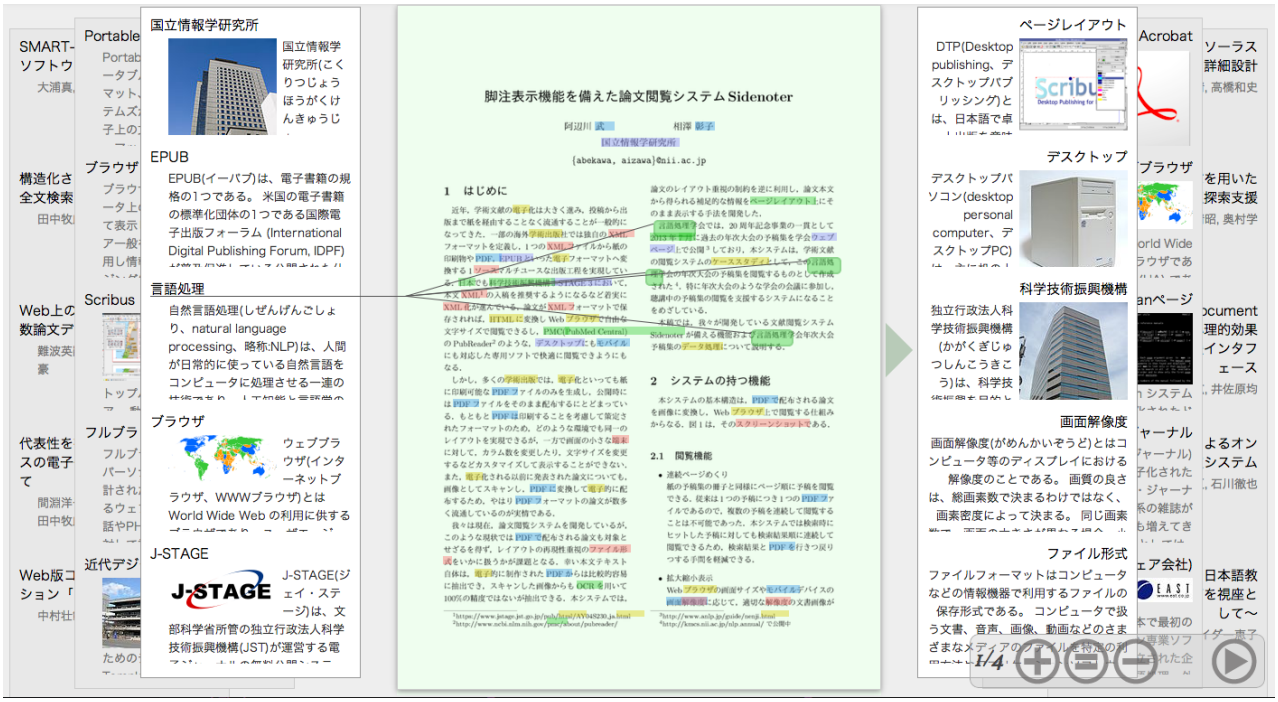

図 1: システムのスクリーンショット

表示できる．また画像の拡大縮小表示の他，2〜4 ページまでの割付表示が可能であり，4 ページ表示で次々と文献を移動させていけば，紙の予稿集においてページをパラパラめくる感覚に近い．

- ブックマーク機能
  論文を次々と閲覧していく際，後で読みたい論文をブックマークして整理しておくことができる．

- 背景色変更
  紙に印刷する際は白背景に黒い文字であることが一般的であるが，バックライトが視覚に入るディスプレイで，長時間，白と黒の組み合わせといったコントラストの高い画面を見続けると目が疲れてしまう．そのため背景色を薄い黄や青といった色に変更し，コントラスト値を下げる機能を実装した．

2.2 検索機能

既存の論文検索サイトにある検索機能と同様で，全文検索で任意のキーワードを含む文献が検索できるほか，タイトル，著者，セッション名などの属性を指定した検索ができる．そして連想検索エンジン GETA[3] を使用した関連文献検索も可能である．

2.3 脚注表示機能

本システムは，現在表示している論文のページの本文を解析し，ページの補足情報をページの左右の脚注部 (Sidenote) に表示する機能を有している．現在表示できる補足情報には次の 2 種類がある．

- **本文中のキーワードに関する情報**
  辞書や百科事典のような見出し語集合とその説明項目が存在するとき，ページ本文中から見出し語を抽出し，説明部分を脚注部に表示する [1]．本システムでは，Wikify[2] や Amazon Kindle の X-Ray[5] の技術と同様に Wikipedia をリソースとして用いており，本文中に Wikipedia のタイトル文字列が出現したとき，その説明文と画像を表示している．本文中でマッチしたキーワードは，文書画像の上にオーバーレイでハイライト表示する．現状，キーワードの語義曖昧性解消や表示する説明のランキングなどは力を入れておらず，今後の課題となっている．

- **ページの一部と関連する情報**
  表示しているページの本文の全部もしくは指定した一部に対して，関連する情報を検索し，ヒットした項目を脚注部に列挙する．本システムでは検

[5] http://www.amazon.com/gp/help/customer/display.html/?nodeId=200729910

索アルゴリズムには前述の GETA を，検索対象のリソースとして Wikipedia と言語処理学会の予稿集を用いている．

筆者の知る限り，論文に対し外部のリソースを本文に結びつける形で併記して表示するシステムは，いまのところ存在しない．

3　データ処理

図 2 に本システムでのデータ処理についてのフローを掲載する．

3.1　テキスト情報の抽出

言語処理学会年次大会の予稿集は PDF で公開されているが，1995 年から 2004 年までが紙の紙面をスキャンしたデータであり，2005 年より PDF 入稿となっている．そのためスキャンデータに関しては，OCR ソフトウェアを用いてテキストを認識し，透明テキスト付き PDF に変換する必要がある．市販の OCR ソフトウェアはそれなりの認識率で日本語・英語を認識するが，2 段組においてカラム間の間隔が極端に狭い場合や，変則的に図表が挿入されている場合にレイアウト認識を誤ることがあり，言語処理としてテキストを使用するには致命的な誤りにつながる．そこで，我々が作成した論文 PDF の OCR ガイドライン [6] にしたがい，OCR ソフトウェアの修正機能を使って人手でレイアウトの誤認識を修正した．

2005 年以降の PDF に関しても，使用しているテキスト抽出ツールではうまく抽出できない PDF があり，そのような PDF については一度画像に変換し，OCR であらためてテキスト認識を実行した．

PDF からテキストを抽出するにあたり，本システムで要求する機能を実現するためには，日本語のような分かち書きのない言語の文字は文字単位で，分かち書きのある言語では単語単位でページ内座標を得る必要がある．しかし，フリーで使用できるプログラムを各種検討したが，条件を満たす使い勝手のよいツールが見つからなかったため，Poppler パッケージ [7] に含まれる pdftotext に独自にパッチをあて対応している．

[6] http://researchmap.jp/muvdvo4s1-385/
[7] http://poppler.freedesktop.org/

3.2　画像変換

本システムで使用している画像形式は，言語処理分野の文献は主に文字主体の内容構成であること，そして背景色を変更する機能に対応するため，PNG 形式を採用している．PDF から PNG への画像変換には，フォント情報を持つ PDF からダイレクトに透過 PNG ファイルを作成できる Poppler パッケージ中の pdftocairo を使用している．スキャン画像から作成された PDF については，背景色 (白色) を透明色に指定した PNG に変換する．高解像度の画像ではファイルサイズが大きくなるため，白黒のページについては 6 色に，カラーのページでは 64 色に減色している．

4　言語処理結果を文書にシームレスに表示する機能

自然言語処理では解析した結果を，元の文書に注釈として追加する処理が多く，近年では文書の保存形式として XML フォーマットを採用し，アノテーションタグを XML 文書に追記する形式となっている．しかし，XML ファイルで表現できても，紙の紙面や PDF のような人間にとってが可読性の高いレイアウトに変換した時，注釈を適切な形で表示する仕組みは今のところ存在しない．

2 節では，文書から得られた情報を左右の脚注部に表示する機能について説明したが，元文書を画像として表示しているので，注釈情報を脚注部だけでなく，文書の上に直接オーバーレイして表示することが可能である．文書上に表示するにあたり，注釈情報と対応するテキストの位置情報が必要になるが，我々は，注釈情報に対応する部分が PDF 中のどのテキストと対応するかを求める仕組みを開発した．

図 3 は，論文とその発表スライドについて，人手で論文中の段落と発表スライドの各ページを対応付け，それを本システムで表示させた例である．本機能により，従来，XML 文書中のアノテーションタグや，テキストに対する簡易な形での付加情報としてしか付与できなかった言語処理の解析結果を，可読性の高いレイアウトで可視化することが可能になる．

5　おわりに

本稿では，脚注部に本文を補足する情報を表示する論文閲覧システムについて解説し，表示するために必要なデータ処理について説明した．そして従来，どち

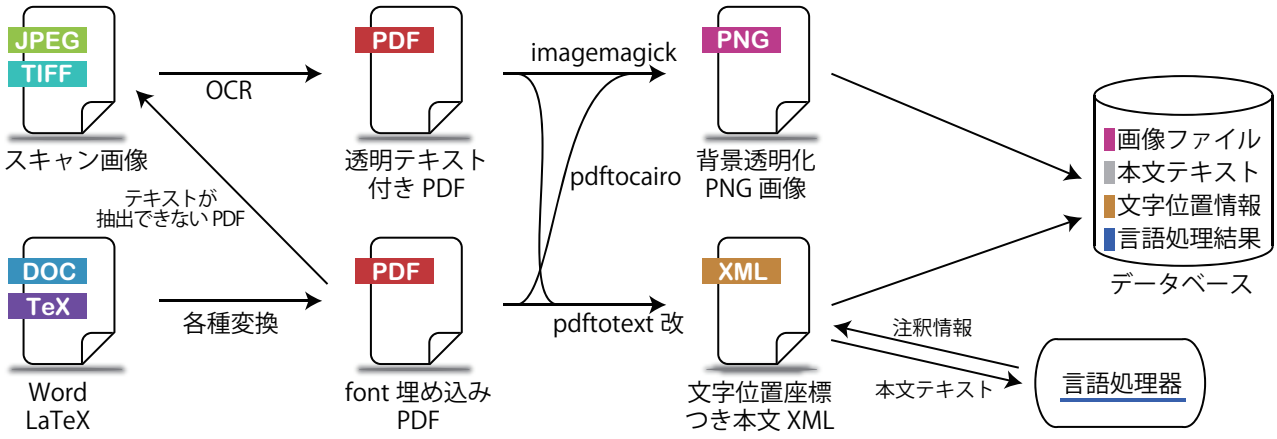

図 2: データ処理の流れ

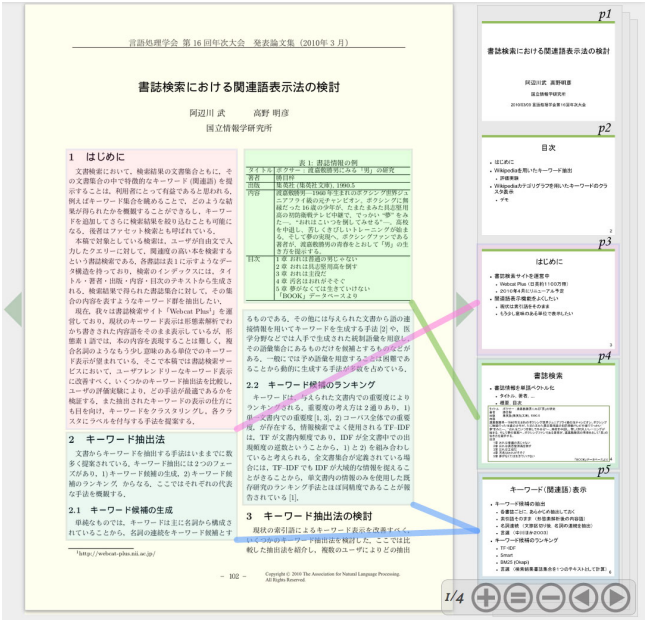

図 3: スライドと論文対応の例

らかというと機械処理の目的のために言語処理技術を用いて解析された注釈情報を，文書レイアウトに反映させ，人間が理解しやすい形で表示する機能について紹介した．

今後の開発予定として，言語横断の関連文献 (ACL Anthology[8] など) の検索表示，SNS に対応したコメント追記機能などを考えている．

電子化にともない XML ファイルとして投稿された論文を，本システムではどのようにして扱うのか，いったん PDF を経由していままで同様に処理するのか，XHTML や EPUB のようなリフロー型のフォーマットで直接ブラウザ上で表現するのか，これも今後の課題である．

参考文献

[1] 阿辺川武, 間下亜紀子. 文章中のコンテクストに適合した関連動画の検索. 情報処理学会研究報告エンタテインメントコンピューティング 2013-EC-27(19), 2013.

[2] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *The 18th ACM Conference on Information and Knowledge Management*, pp. 233–242, 2007.

[3] 西岡真吾. 汎用連想計算エンジン GETA. コンピュータソフトウェア, Vol. 26, No. 4, pp. 87–106, 2009.

[8] http://aclweb.org/anthology/