

内部構造解析機能と脚注表示機能を備えた論文閲覧システム Paper Browsing System with Structure Analysis and Displaying Annotation on Side-note Windows

阿辺川 武* 相澤 彰子
Takeshi Abekawa Akiko Aizawa

国立情報学研究所
National Institute of Informatics

Abstract: In this paper, we introduce our on-going efforts to construct a scientific paper browsing system to assist users to read and understand advanced technical content. The paper features on two major functions that are prerequisite for such systems: document structure analysis for image, PDF, and XML formatted articles, and automatic link detection that help users access richer information from diverse external sources. We also present technical details of our current implementation to generate and display the linked external data in side-note windows with a target paper image.

1 はじめに

近年、学術文献の電子化は大きく進み、論文は投稿から出版、読者の手元まで印刷媒体を経由することなく電子フォーマットで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、あるいは PDF, XHTML, 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と呼ばれる出版工程が実現されている。日本でも科学技術振興機構 J-STAGE 3 において、本文 XML¹ 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 フォーマットになれば、デバイスの特性に応じて自由にレイアウトや文字サイズを変更できるリフロー型と呼ばれる方法で論文を表示できるようになる。例えば XHTML に変換すれば Web ブラウザで自由な文字サイズで閲覧でき、PMC(PubMed Central)の PubReader² のようなデスクトップにもモバイルにも対応した専用アプリケーションで快適に閲覧できるようになる。

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、出版時に PDF ファイルをそのまま配布するに留まっているのが現状である。PDF は画面上でも印刷した紙でも同一のページレイアウトを維持することを目的して策定され

たフォーマットのため、紙面レイアウトは常に不変である。そのため画面の小さなデバイスに対して、コラム数を変更したり、文字サイズを変更するなどの表示のカスタマイズが不可能である。電子化が進む以前に発表された論文についても同様に、紙面をスキャンして画像に変換後、PDF にパッケージングするため、やはりレイアウト固定型の 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、Web ブラウザ上で動作する論文閲覧システム SideNoter を開発している。XML 形式の論文のみを対象にできれば、表示レイアウトの自由度の高さから、閲覧に適したインターフェースを開発することができる。しかし、上述の現状では PDF で配布される論文が大半を占めるため、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、論文自体は画像で表示し、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重ねて表示する手法を採用した。

現在、言語処理学会年次大会過去 20 年分の予稿集を用いて、開発中のシステムの使い勝手を検証している³。本システムは、PC につながったディスプレイで論文を閲覧するだけでなく、年次大会のような学会の会議に参加し、モバイルデバイスで聴講中の予稿集を閲覧する支援になることもめざしている。

*連絡先：国立情報学研究所
〒101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: abekawa@nii.ac.jp

¹https://www.jstage.jst.go.jp/pub/html/AY04S230_ja.html

²<http://www.ncbi.nlm.nih.gov/pmc/about/pubreader/>

³<http://kmcs.nii.ac.jp/nlp-annual/> で公開中

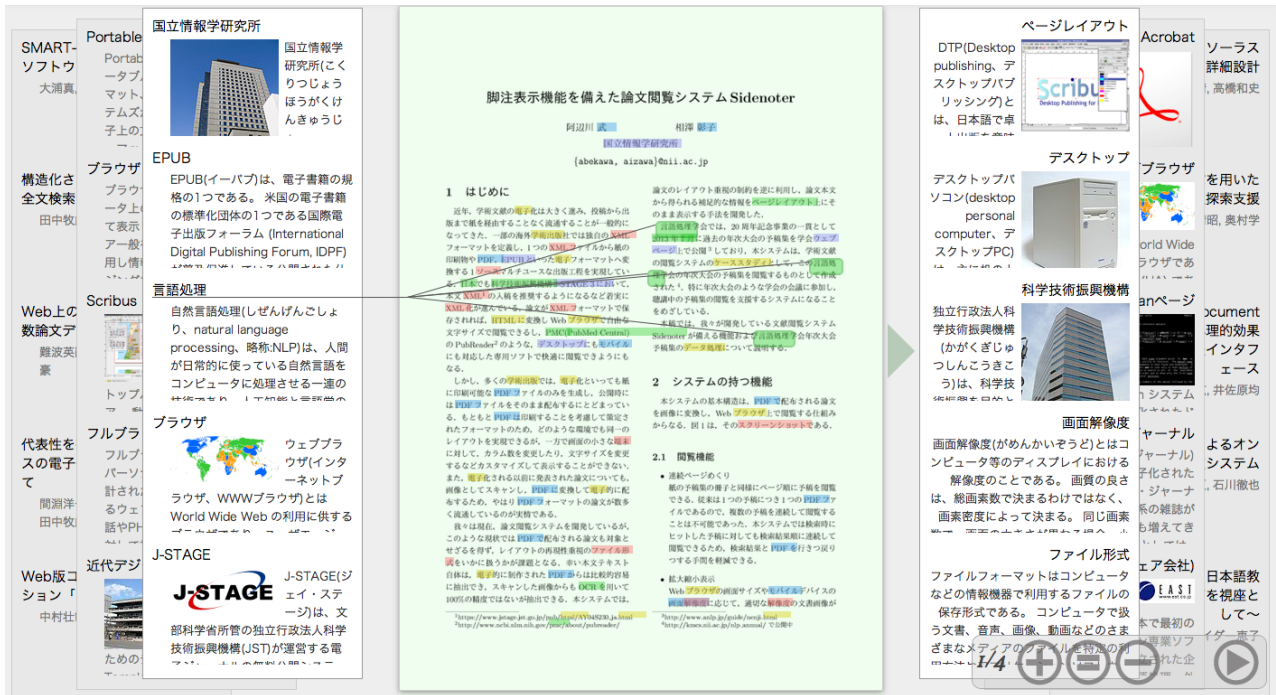


図 1: システムのスクリーンショット

以下、本稿では、我々が開発している文献閲覧システム SideNoter の基本設計およびデータ処理方法を説明し、次にシステムが備える機能について紹介し、最後に外部で付与された注釈を論文レイアウト上に表示する仕組みについて説明する。

2 関連研究

現在，電子的に流通している論文フォーマットの大半は PDF 形式であり，PDF 形式の論文を表示できるソフトウェアを論文閲覧システムとみなせば，EndNote⁴や Mendeley⁵などの引用文献管理ツールもその 1 つと言える。

書物を読むとき、気になった事柄や、記述に対する疑問を紙面の余白に書き込みながら読書することを Active Reading と呼び、深い読書のためには重要な機能である [1]. この概念を取り入れたシステムとして、タブレット上にペンを使って書き込みがおこなえるデバイス XLibris[4] が 1998 年に発表されている. 論文の閲覧に特化したシステムとして、画面上で本文テキストを自由に切り取り再構成できる LiquidText[5] や、本文テキストの行間を引き伸ばしてそのスペースに書き込みをおこなう TextTearing[2] などがある.

一方、内容理解を目的とする論文読解支援システムとして、鉢木らは、論文と Web を連携させ、論文のタイトルページから専門用語を抽出し、その用語を解説するページを論文横に表示するシステムを提案している [10]。石戸谷らは論文と映像の部分要素に対するアノテーションの獲得と蓄積をおこなうための仕組みを開発し、論文と映像を自由に切り替えて閲覧するシステムを提案した [8]。

本システムの特徴は、以上のような論文閲覧と論文読解に必要とされる機能を部分的に取り入れ、さらにサイドノート部分に様々な情報が提示でき、それらを論文レイアウト上で本文と結び付けられるところにある。

3 基本設計

最初に本システムのスクリーンショットを図 1 に掲載する。システムは PC もしくはモバイルデバイスの Web ブラウザ上で動作し、マウス、キーボード、タッチ機能で操作する。論文本文は、3.1.2 節で説明する画像変換から得られた画像そのものを画面中央に表示する。ハイライト部分や特定の領域を画像の上にオーバーレイする形で塗りつぶすことができる。そして本文画像の両サイドには、論文読解を支援する各種リソースを掲載するスペースがあり、必要に応じて、中央の論文本文と線分で接続することが可能である。

⁴<http://endnote.com/>

⁵<http://www.mendeley.com/>

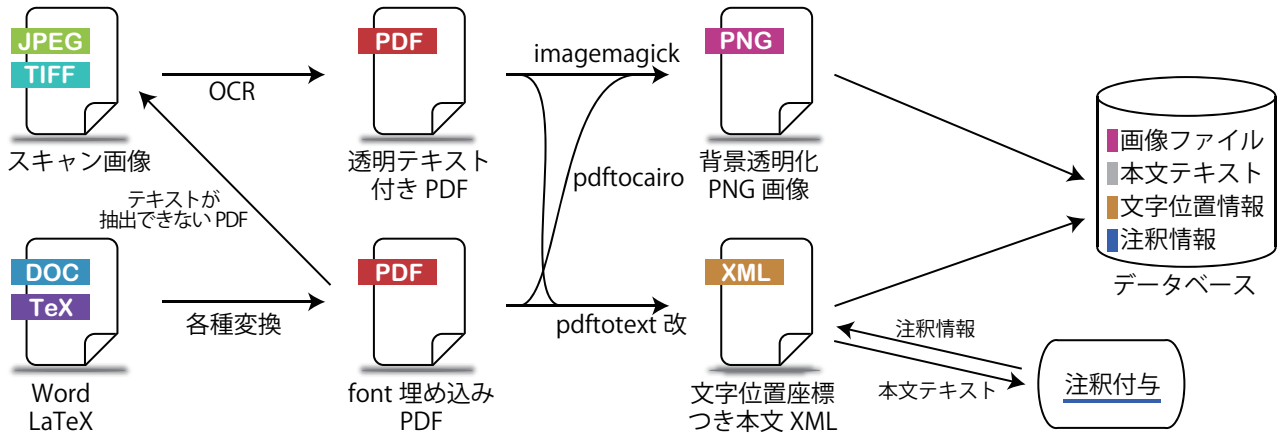


図 2: データ処理の流れ

3.1 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.

3.1.1 テキスト情報の抽出

現在，本システムで対応する論文の形式は，紙に印刷された論文および PDF で流通する論文である．紙の論文に対しては，スキャナーを使用してデジタル画像データに変換後，OCR ソフトウェアを用いて文字認識を実行し，透明テキスト付き PDF に変換する．一方， $\text{\LaTeX}$ や Microsoft Word から作成された PDF 形式の論文はほとんどが PDF 内にテキスト情報が保存されているが，一部の PDF ではフォントのグリフ情報と既存の文字コードの対応表が存在しないものもある．そのような PDF では正しくテキスト情報が抽出できないため，一度画像データに変換した後，OCR を施しテキスト情報を得る．

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、オープンソースソフトウェアライセンスで提供され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⁶に含まれる `pdftotext` に独自にパッチをあて対応している。

3.1.2 画像変換

本システムの本文表示で使用している画像形式は、本文ページの背景色を変更する機能に対応するため、透

過 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、`imagemagick`⁷ を用いて背景色 (白色) を透明色に指定した PNG に変換する。

通信トラフィックを減少させるため、Web ブラウザの表示ウィンドウの大きさに合わせて解像度を変えた画像を用意し、白黒のページについては 6 色に、カラーのページでは 64 色に減色した画像を表示に用いている。

4 論文閲覧機能

本節では、論文閲覧に必要であると考える機能を考え、システムの持つ機能を従来の形式(紙、PDF、XML・XHTML・EPUB))と比較しながら説明する。

- 本文検索

ユーザが指定した任意のキーワードで本文テキストの検索ができる。紙に印刷された論文ではもちろんできないが、PDF や XML 形式では可能である。

- 専門用語の Web 検索

5 節で説明する自動用語認識により、特定の専門用語ではクリックすると、その用語の説明ページを表示するか、あるいは Web で検索するかを選択できる。PDF や XML 形式では、ユーザが用語を自ら選択しコピー後、別ウィンドウでテキストボックスにクエリーとして貼り付け検索を実行しなければならない。

- 連続ページめくり

紙の論文では、通常、冊子体で出版され、複数の

⁶<http://poppler.freedesktop.org/>

⁷<http://imagemagick.org/>

論文をページ順に順番に閲覧できる。一方でPDFやXML形式などの電子媒体では、閲覧単位が1論文なので、次の論文を閲覧する際には、一度リストページに戻る必要がある。本システムでは複数の論文が掲載ページや検索結果など順序付きリストとして定義されているとき、カーソルキーあるいは画面上のボタンを押すことで連続して次々と論文を閲覧できる仕組みを持つ。これにより、一覧ページと論文とを行きつ戻りつする手間を削減できる。

- 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、適切な解像度の本文画像を表示とともに、画像の拡大表示の他、2～4ページまでの割付表示が可能である。紙の論文では文字が小さく、もっぱら電子媒体で論文を読む人には必須の機能である。

- コントラスト変更

紙に印刷された論文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コンピューターディスプレイで、白と黒の組み合わせのハイコントラストを持つ画面を長時間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。

- 柔軟なレイアウト

モバイルデバイスでは画面が小さいという制約があるが、紙やPDFのような固定レイアウトのフォーマットでは、2カラムを1カラムで表示するといった画面の大きさに合わせた柔軟なレイアウトの変更ができない。一方でXMLのようなリフロー形式では自由にレイアウトの変更ができる。本システムでは画像ベースの表示のため残念ながらレイアウトの変更はできない⁸。

- 別ページ図表の閲覧

論文中の図表はその大きさから、図表の説明文と同一ページに配置できない場合があり、説明文があるページと図表があるページを交互に見ながら論文を読む状況が発生する。リフロー形式では要素の配置を変更すれば同一画面に表示できるが、レイアウト固定フォーマットでは難しい。本システムでは、PDF形式から図表を個別に切り出す処理を開発することで、説明文とともに図表を表示する機能を搭載する予定である。

表 1: 論文フォーマットごとの機能特性

	紙	PDF	XML	本システム
本文検索	×	○	○	○
専門用語の Web 検索	手入力	Copy	Copy	Click
連続ページめくり	○	×	×	○
拡大縮小表示	×	○	○	○
コントラスト調整	×	×	○	○
柔軟なレイアウト	×	×	○	×
別ページ図表の閲覧	×	×	○	予定
参考文献リンク	×	×	×	○
書き込み	○	○	○	○
書き込みの集約	×	×	×	予定

- 参考文献リンク

本システムでは、論文末尾の参考文献部分から参考文献のリストを取得し、i-Linkage システム [9] を用いて書誌データベースと同定処理をおこなう。インターネット上で公開されている文献の場合にはそのリンクを付与し、なければ、論文は CiNii⁹ への、書籍は Webcat Plus¹⁰ の該当書誌へのリンクを生成する。

- 書き込み

紙の論文では自由に書き込みや下線を引くことができ、PDF 形式にも注釈機能がある。本システムにおいても、本文画面上でマウスを用いて領域選択をおこなうと、その領域内のテキスト部分がハイライトされ、コメントの記述やハイライトの保存ができる。

- 書き込みの集約表示

紙や PDF 形式への書き込みは紙面そのものあるいは論文ファイル内に保存されるため、書き込みの内容と論文を切り離すことが出来ない。そのため、関連文献すべての書き込みを後からまとめて参照することが難しい。本システムでは、書き込みは論文ファイルと分離して保存されるので、ユーザによる書き込みの集約表示が可能である(実装予定)。

最後に、本節で説明した本システムの各機能について従来の論文フォーマットとの比較を表 1 に掲載する。

5 論文読解支援機能

本システムは、論文の表示・操作・書き込みといった閲覧に必要な機能とともに、現在表示している論文

⁸k2pdfopt: <http://www.willus.com/k2pdfopt/> のようなツールを用いれば、固定型のフォーマットでもデバイスの幅に応じたレイアウトの変更ができる

⁹<http://ci.nii.ac.jp/>

¹⁰<http://webcatplus.nii.ac.jp/>

のページの本文を解析し、ページの補足情報をページ画像左右の脚注部 (Side-note) に表示する論文読解支援機能を有している。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というリソースが存在するとき、論文本文中から見出し語を自動抽出し、説明部分を脚注部に表示する。本システムでは、Wikify[3] や Amazon Kindle の X-Ray¹¹ 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、論文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の精度は高くなく、今後の課題となっている。

キーワードの説明として、辞書の語釈文のようなあらかじめ静的に定義された文章のほかに、キーワードをインターネットの検索エンジンに渡し、その検索結果を表示する動的な情報提示も可能である。本システムではキーワードを動画検索サイトで検索し、上位1位の結果を表示することができる[6]。最近では大学の講義などがインターネット上で数多く公開されているため、キーワードを解説する講義映像がヒットすることが増えてきた。

- 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全部あるいは指定する部分文章に対して、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。本システムでは検索アルゴリズムには連想検索エンジン GETA[7] を用いており、検索対象は文書集合として定義できるものならなんでもよく、現在公開されているシステムでは Wikipedia 全ページと、言語処理学会の予稿集を用意している。

これらの情報は、日々刻々と内容が変化するため、Web ページに表示する際に、本文テキストからリアルタイムでキーワード抽出、説明部分の検索、生成がおこなわれる。

6 注釈情報をシームレスに表示する機能

ユーザによる論文本文に対する書き込みやハイライト付与は、本文ページがレイアウトされた状態での注

¹¹<http://www.amazon.com/gp/help/customer/display.html/?nodeId=200729910>

釈付与であり、該当する本文テキストや座標との対応付けを保存しておけばよい。一方、人手による大量の注釈付与作業や、機械的な注釈付与を考えた場合、表示レイアウトそのままの扱いは難しいため、通常、論文のソースファイルあるいは、テキストフォーマットに変換したデータに対して注釈付与をおこなう。

近年ではソース文書の保存形式として XML フォーマットを採用し、注釈タグを XML 文書に挿入する形式が主流となっているが、付与された注釈情報を PDF のような人間にとって可読性の高いレイアウトに適切な形で表示する仕組みは今のところ存在しなかった。

そこで、我々は注釈情報をレイアウトされた論文画面上に表示するにあたり、XML 中の注釈情報が PDF 中の論文本文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。あらかじめ PDF から座標情報付きのテキストを抽出しておき、XML 中の注釈を指定するタグが囲むテキストと照合し、注釈情報を表すテキストに対し、ページ番号と座標を記録するものである。

本システムでは、論文本文を画像として表示し、本文テキストの座標情報を内部で有しているため、注釈情報を論文画像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。

本機能を利用した例として、論文本文とその発表スライドの対応を示す。最初に論文 PDF から本文のみの XML ファイルを作成し、発表スライドの各ページに対応する論文のテキスト部分を XML タグで囲んだ。次に XML ファイルと PDF ファイルを本システム処理し、表示させた例が図3である。本機能により、XML 文書中の注釈情報および、係り受け解析や名詞句のチャンキングなどテキストに対する様々な解析結果を、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。

7 おわりに

本稿では、紙に印刷された論文および PDF 形式で流通する論文を対象に、効率的な論文閲覧を実現する機能を有し、論文の本文情報から得られる様々な情報を脚注部に表示する論文閲覧システムを説明した。さらに、論文から抽出したテキストに対し付与した注釈情報を、整形された論文レイアウト上に同時に表示する仕組みを開発した。

現状のシステムでは、論文の内部構造解析として参考文献部分の抽出をおこなっているが、今後、図表領域、およびタイトルやセクションなどの構成要素の認識をおこなう予定である。論文のレイアウト画像上に情報を表示できる本システムは、情報提示法の自由度が高く、これらの認識した要素をどのような形で表示すれば、より効果的に論文閲覧、論文理解になるかを考えていきたい。

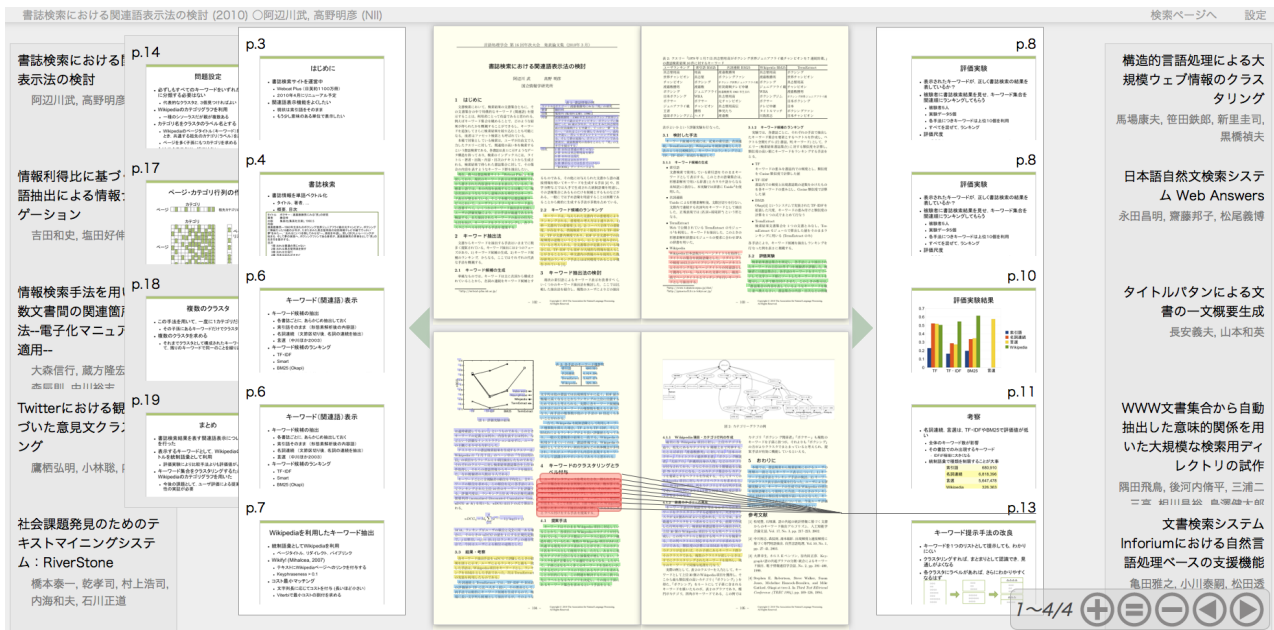


図 3: スライドと論文対応の例

また、大量の論文に対し内部構造解析をおこない、それぞれ認識された要素を有機的に結びつけ、本システムで表示したとき、どのような世界が見えてくるのでしょうか？ プロトタイプシステムが完成した際には追って報告したい。

参考文献

- [1] Mortimer Jerome Adler. *How to Read a Book*. Simon and Schuster, 1940. 邦題:本を読む本. 外山滋比古, 榎未知子訳. 講談社学術文庫. 1987.
- [2] Francois Guimbretire Dongwook Yoon, Nicholas Chen. Texttearing: opening white space for digital ink annotation. In *the 26th annual ACM symposium on User interface software and technology*, pages 107–112, 2013.
- [3] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *The 18th ACM Conference on Information and Knowledge Management*, pages 233–242, 2007.
- [4] Morgan N. Price, Bill N. Schilit, and Gene Golovchinsky. Xlibris: the active reading machine. In *CHI '98 Cconference Summary on Human Factors in Computing Systems*, pages 22–23, 1998.
- [5] Craig S. Tashman and W. Keith Edwards. Liquidtext: A flexible, multitouch environment to support active reading. In *CHI '11 Conference on Human Factors in Computing Systems*, pages 3285–3204, 2011.
- [6] 阿辺川武 and 間下亜紀子. 文章中のコンテキストに適合した関連動画の検索. In *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.
- [7] 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, 26(4):87–106, 2009.
- [8] 石戸谷顕太郎, 山本圭介, 大平茂輝, and 長尾確. 映像と論文へのアノテーションに基づく論文読解支援システム. 映像情報メディア学会誌, 66(11):J461–J470, 2012.
- [9] 相澤彰子, 高久雅生, and 大山敬三. 大規模データベースを利用したリンケージシステムの提案と実装. *DBSJ Letters*, 6(4):17–20, 2008.
- [10] 鉢木稔浩, 太田学, and 高須淳宏. Web 資源を利用した学術論文閲覧支援システム. In *情報処理学会研究報告データベース・システム研究会報告 2009-DBS-149(14)*, pages 1–6, 2009.