人工知能学会 インタラクティブ
情報アクセスと可視化マイニング研究会(第10回)
SIG-AM-10-01

文書構造に基づく対話的情報アクセスにむけて

Towards Interactive Information Access based on Document Structures

加藤 恒昭[1*]　岩月 憲一[1]　山口 和紀[1]
Tsuneaki Kato[1]　Kenichi Iwatsuki[1]　Kazunori Yamaguchi[1]

[1] 東京大学 大学院 総合文化研究科
[1] The University of Tokyo Graduate School of Arts and Sciences

Abstract: A framework is examined, in which the users interactively access documents, like scientific papers, with a physical structure appearing in the layout and a logical structure based on their contents. It supports effective and flexible use of the documents by allowing the users to retrieve relevant logical units through specification of their contents and/or roles in the document, and to browse those units and their contexts by strolling across both logical and physical structures. The whole framework and a method of document analysis that reconstructs the logical structure of a document and constructs its representation are mainly discussed in this paper.

1　はじめに

一般に文書は，章立てのような意味内容に基づく論理構造と，印刷・表示される場合のレイアウトに対応する物理構造を持つ．本稿では，これらの構造を利用することで，様々な検索意図に対応しうる情報アクセス環境が構築できることを述べる．まず，情報アクセスにおいて，文書全体でなく，文書の構造を用いてその部分にアクセスできることの必要性を述べ，そのような構造が対話的な情報アクセスにおいても重要であることを指摘する（2 節）．続けて，文書構造に基づく情報アクセスによってどのような検索意図に応えられるかを掘り下げ，そのために必要な文書表現を検討する（3 節）．その後，そのような文書表現を得るための文書の論理構造抽出について，方針と現状を報告する（4 節）．最後に関連研究について言及し（5 節），今後の方針を述べて全体をまとめる（6 節）．

以下，学術論文や学会発表予稿集，特に言語処理学会 20 周年記念で公開された年次大会予稿集[1]を，構造を持つ文書の例として議論を進めるが，その議論は，意味内容に基づく論理構造と，それと結びついたレイアウト等の物理構造を持つ情報源に自然に拡張できる．例えば，Wikipedia のようなマルチメディア事典，コマ割りという論理構造かつ物理構造を持つコミック等についても，同じようなニーズが存在し，同じ枠組みで捉えることができると考えている．

*連絡先：東京大学大学院総合文化研究科言語情報科学
〒 153-8902 東京都目黒区駒場 3-8-1
E-mail: kato@boz.c.u-tokyo.ac.jp

[1] http://www.anlp.jp/resource/annual_meeting.html

2　情報アクセスと文書構造

一般に文書として流通している情報は，情報アクセスの単位として必ずしも適当なものでなく，文書の構成要素に直接アクセスできることが必要である．例えば，学術論文や学会発表予稿集は研究活動を進めるにあたっての重要な情報であり，様々な検索意図に基づいた情報アクセスが行われる．それらに答えるために必ずしも文書全体が必要なわけではない．ある評価指標の定義が知りたいのであればひとつの式がその回答になるであろうし，その評価指標を利用するための評価実験の概要が知りたければ，論文の一節だけを提示すればよい．その評価指標がどの程度一般的なものであるかを知りたいのであれば，それを用いている論文の数だけでも参考になる．この例のような文書の一部に関心があるという場合に限らず，そこで述べられている研究そのものに興味関心がある場合でも，利用者は論文を最初から丁寧に通読していくわけではない [16]．梗概や導入だけを読んで，その価値を，読み進めるに値するかを判断することも多い．であればまずはその部分だけを提示するのが適切であろう．

文書全体ではなくそこに含まれる特定の情報が利用者のニーズを満たすということは，パッセージ検索 [4, 6, 12] や質問応答 [15] の動機となっている．ただ，初期のパッセージ検索の動機は文書の適合性を測る場合にそれ全体の特徴ではなく，その部分に注目した方がよいというものであるし，質問応答は文書全体の主題と無関係にそこに含まれる情報を利用しようというものであった．そこでは，文書の構成要素が文書とは独

立に扱われていて，構成要素が文書という構造の中である役割を持っており，それに基づいてアクセスされるという視点は弱い．上述の評価や梗概の例のように，文書の構成要素はそれ自身の特徴だけでなく，文書という構造の中での役割に基づいて利用できることが求められる．あわせて，これらの取り組みでは，対話的な情報アクセスの観点が欠けている．

学術論文を含め，様々な情報の活用は対話的・探索的に行われる．複数の検索結果を斜め読み的に閲覧して，必要な情報を見定めるという，既に述べたような利用に加えて，ある評価指標の定義からその利用方法への関心の拡大．関心を持った文書からそこで引用されている文書への推移等，Bates のいう Berrypicking[2] での推移，Ellis のモデルにおける Chaining[5] のような推移に対応しなければならない．文書間の推移については，例えば文書を引用関係で結び付けたハイパーテキスト構造を閲覧の対象とすること等が試みられているが，文書内に閉じた閲覧やブラウジングにおいても，それぞれの情報の文脈を提示することや概要から詳細への焦点の推移が重要になる．最初の例に戻れば，評価指標の式からそれを含んだ評価実験の記述への推移や，その逆の推移が自然に行えることが望ましい．その点でも，文書を単位とせず，文書の構造を意識することが必要である．そして，そのような文脈や構造を利用者に自然に提示するものとして，論文誌，予稿集に掲載されていてレイアウト，物理構造が有益であることが期待される．このような形式は文書閲覧の形式として馴染みがあることに加えて，一般にはテキスト検索の対象とならない図表類を情報として含んでおり，対話的な検索を通じてそれらの情報を提供する機会を与えることになる．

このような着眼に基づいて，1) 文書を意味内容に基づく論理構造を持つものと捉え，情報アクセスの単位をその構造の構成要素とするような情報アクセス環境の実現を検討する．論文等の場合，文書の論理構造はいわゆる章立てに対応し，あわせて，タイトルや著者情報，参考文献などが論理構造の構成要素（論理要素）となる．ここで，単に文書を小さな単位に分割・分解するのではなく，それぞれがどのような文脈にあったか，どのような構造の一部であったか，を維持し，検索意図との照合やその後のインタラクションに利用する．2) このような情報アクセスを対話的プロセスの一部とするために，文書が論理構造のみでなく，レイアウトのような物理構造を持ち，図表等の視覚情報を含むことを活かした閲覧やブラウジング等のインタラクションを検討する．レイアウト等の物理構造は論理構造と一定の関係を持つが，必ずしも同じものではない．検索が論理構造に基づいて行われるので，このようなインタラクションはあわせてこの論理構造を意識し，物理構造と論理構造を行き来できなければならない．

3 検索意図との照合

前節で述べた様々な検索意図について分類し，それに応えるためにどのような情報が必要かを検討する．

検索意図は，まず，文書（この場合は研究論文）そのものを必要するものとその部分（構成要素）で応えられるものとに分類される．研究論文はすべて何らかの研究について論じていると看做せるので，その研究を特徴付ける概念が，文書の主題となる．したがって，文書そのものへの検索意図は研究に関する記述を求めていると考えられるが，その研究の指定の仕方は大きく以下の３つに分けられる．

1. 主題に基づくもの
   例:「WordNet についての研究」
2. その他の情報によるもの
   例:「知識源として WordNet を用いている研究」
3. メタ情報（書誌情報）によるもの
   例:「2014 年以降に発表された研究」

知識源や評価尺度として何を利用しているか，どのような文献を参照しているか等は必ずしも主題として研究を特徴づけるものではないので，1. と 2. は区別される．著者や著者が所属する組織等文書そのものから得ることができるメタ情報もあるが，情報とメタ情報の違いとして 2. と 3. が区別される．2. の検索意図に応えるためには，文書の主題を反映する文書表現だけでなく，特定の役割や部分における特徴を蓄積する必要がある．典型的な例は参照している文献による研究の検索で，文書の参考文献の部分に指定された文献が含まれることが条件となる．

一方，文書の部分，その構成要素に対する検索意図は，文書を介するか否かで分類できる．文書を介さない検索意図は，あるキーワード，例えば，WordNet や相互情報量の定義や説明を知りたいというようなもので，その回答はどのような研究で使われているかに関係しない．これは質問応答技術が扱うような検索意図に近く，文書の構成要素毎にその特徴を表現し，適合するものを選択し，更に必要に応じてその一部を抽出して回答することが求められる．一方，文書を介するものは，前述のいずれかの方法で研究を指定し，それに関連する情報を求める．「〜研究における評価手法を知りたい」「〜研究においてよく参照される文献を知りたい」が例となる．この場合，それが文書に対して持つ役割に基づいて，構成要素が検索意図に適合するかを判断する必要がある．例えば，ある構成要素がその研究の評価手法についての部分であることが表現されていなければならない．

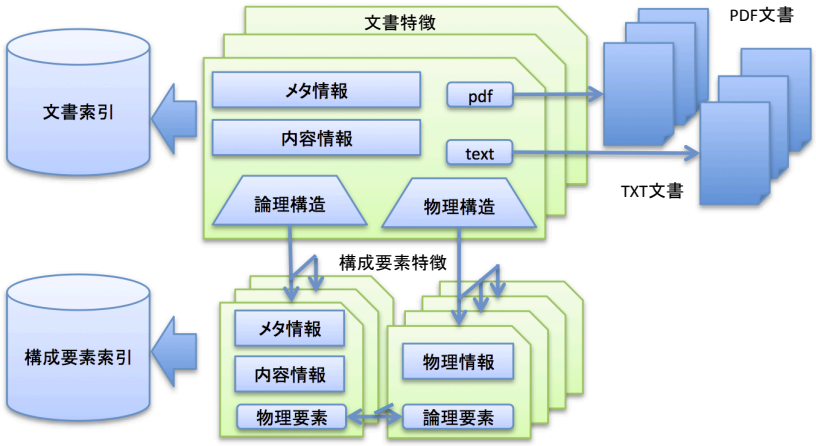

図 1: 文書の表現

このような様々な検索意図に対応するためには少なくともふたつのことが必要となる．ひとつは，表現された検索要求の背後にある検索意図の曖昧性の解消あるいは，その広がり（diversity）に配慮した検索方針で，例えば,「WordNet」という要求で表されている意図が,「WordNet についての研究」「WordNet を使った研究」「WordNet とは何か」等のいずれであるかを明らかにする必要がある．同様に「統計的機械翻訳の評価」は,「統計的機械翻訳の評価についての研究そのもの」や「統計的機械翻訳についての研究の評価」を求めている場合がある．

もうひとつは，そのような意図を満たすための文書表現と照合方式で，上で述べたように，文書の主題に関する表現だけでなく，メタ情報や，その構成要素に関する情報が必要となる．構成要素に関する情報としては，その主題に関する表現に加えて，文書における役割が明らかにされている必要がある．この役割情報は構成要素のメタ情報であり，それによって，文書を選択する条件に関連する部分であるかや，文書中の求められている部分であるかが判断される．これらを適切に使い分けて検索意図との照合を行う必要がある．

このような照合とその後の閲覧を考えた場合に，蓄積すべき文書表現と関連情報を図 1 に示す．文書はそのレイアウトを維持した PDF 文書とそこに含まれるテキストを抽出した TXT 文書として記憶され，そこから取り出された様々な情報が文書特徴として記述される．その中にその論理構造と物理構造の記述がある．論理構造と物理構造は対応づけられ，論理構造のそれぞれの要素については，そこに含まれるテキストについての内容情報と文書中での役割を示すメタ情報が記述され，物理構造の要素にはレイアウトにおける位置情報等が記述される．次節で述べるが，物理構造の要素（基本要素と呼ぶ）は論理構造と n:1 の対応を持つ．これらの文書特徴，構成要素特徴から検索に用いられる索引情報が生成される．

4 論理構造の抽出

4.1 方針

前節で述べた文書表現を獲得するために，文書からその物理構造と論理構造を抽出する検討を進めている．文書として予稿集等の PDF 文書を想定する．PDF 文書は LaTeX や MSWord 等の文書作成組版システムによって直接作成されるディジタル文書と紙媒体の文書をスキャンして得られるスキャン文書に分類される．言語処理学会年次大会予稿集においては，2003 年まではスキャン文書，それ以降はディジタル文書となっている．

スキャン文書から検索可能なテキスト情報と物理構造および論理構造を抽出するためには，OCR ソフトウェアを用いる．一般に OCR 処理はレイアウト認識と文字認識からなる．レイアウト認識は文書の各ページを矩形領域に分割した後，それらをテキスト，表，図等に分類し，位置や大きさの情報を得る．その後，テキストと分類された矩形領域を単位として，そこに含まれる文字の文字認識が行われ，テキスト情報が抽出される．日本語文書の OCR ソフトウェアにおいては，e-typist[2] とその上位製品である Win Reader Pro[3] が，認識結果を xhtml 形式で出力する機能を持ち，そこでは認識された矩形領域が xhtml の span 要素と対応し，その属性として，矩形の位置や大きさが表現される．OCR ソフトウェアのレイアウト認識と文字認識は，ともに完璧ではない．レイアウト認識の問題は後述するが，文字認識においても，特にスキャンの質が低い文書では誤

[2] http://mediadrive.jp/products/et/
[3] http://mediadrive.jp/products/wrp/index.html

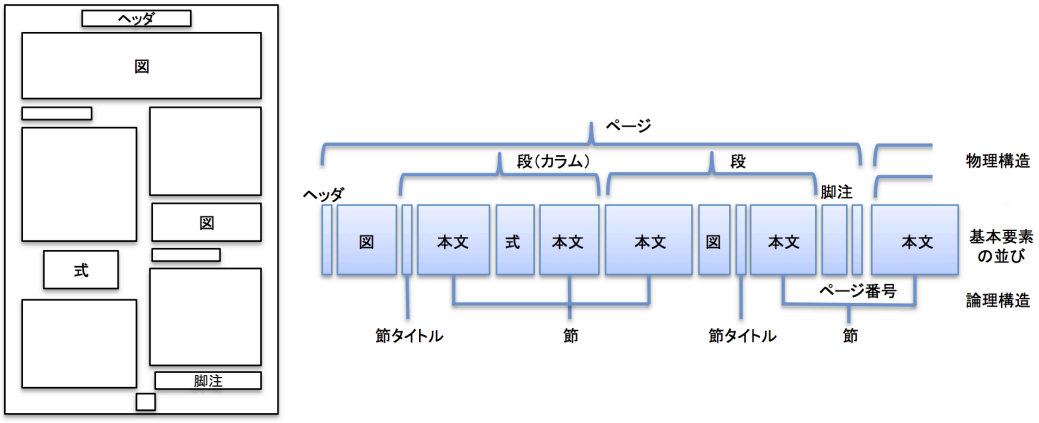

図 2: 論文のページレイアウトと物理構造と論理構造

りが多発するし，数式に使われるような記号としてのアルファベットは殆ど扱えない．このため，OCR 処理には人手介入が許されており，文字認識結果の後修正だけでなく，レイアウト認識を人手で修正した後に文字認識を行うことも可能となっている．

ディジタル文書は，その内部にテキスト情報を持っており，pdftotext[4]などのソフトウェアでこれを抽出することができる．この場合，抽出結果に OCR ソフトウェアの文字認識で生じるような誤りはない（ただし，[7]）．一方で，ほぼ行単位で抽出される文字列の順序は必ずしも文書作成者が意図したあるいは一般的な読者が読み進む順序とは一致しない．また，文字の位置についての情報は得ることができるが，OCR ソフトウェアのレイアウト認識で得られるような人間の直観にあった矩形領域への分割は取得できない．ディジタル文書を html 等に変換するものも配置されるのは行であり，OCR ソフトウェアのレイアウト認識における矩形のような概念は存在しない[5]．

OCR ソフトウェアのレイアウト認識は空白部分の存在（スペーシング）等の情報を用いて矩形領域を認識する．それらは文書の論理構造や意味内容を意識していない．一方，前節で述べた目的のためには，物理構造は論理構造と一定の関係をもつ必要がある．具体的には，論理構造の単位となるものが，紙面の物理的な制約の下で必要に応じて分割され，配置された構造を物理構造と考える．物理的な制約とは，多段組みにおける段の境界，ページの境界，図の挿入，脚注の挿入，ヘッダやフッタの存在などである．例えば，図 2 において，図の左に概念的に示すような論文の 1 ページについて，矩形で囲った部分それぞれを物理構造の基本要素と考える．これらの要素は 2 次元的に配置されているが，２段組の原稿であることを考慮すると，簡単な規則によって図の右に示す 1 次元の並びとすることができる．物理構造を考えた場合，並べられた基本要素が，段やページ等を構成していくし，論理構造を考えた場合は，節やそのタイトル等の物理要素が得られる．物理構造においては常に連続した要素がより大きな構造をなしていくが，論理構造は必ずしもそうではなく，図や脚注を間に挟んで一つの要素を構成する場合がある．物理構造と論理構造の関係をこのように位置づけると，物理構造と論理構造は共通の基本要素をもち，論理要素はひとつ以上の基本要素の並びから構成される．そして基本要素は，複数の論理要素を自分の中に含まないことがその条件となる．

OCR ソフトウェアのレイアウト認識の役割をこのような基本要素を矩形領域として抽出することと捉えた場合，その出力は様々な「誤り」を含む．それらは以下のように分類することができる．

1. 複数の論理構造の要素を含んだ矩形領域が抽出される．例えば，節のタイトルと節の本体，本文と脚注，図や表とそのタイトル，がひとつの矩形領域を構成する．

2. その一部にテキストを含むような図や表を多数の小さなテキスト矩形領域の集まりと認識する．

3. 多段組の文書を前提とすると不必要であるような過分割を行う．箇条書きやタイトルにおいて，中黒等の記号や番号等と本体部分との間隔が広かったり，文章中の句読点の配置等により，矩形の境界と誤認識されるような空白が生じることが原因である．

1. については，スキャンの品質が低く，段組みの間隔が狭い文書などに対しては２段組みの左右の段をひとつの矩形と認識するなど致命的な誤りを犯す場合もあ

[4] http://poppler.freedesktop.org
[5] 著者の調査不足であれば，ぜひご教示いただきたい．

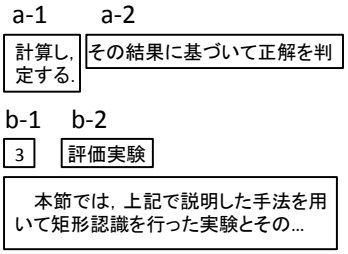

図 3: レイアウト認識の誤り例

る．3. は，図 3 に示すような場合で，a-1 と a-2，b-1 と b-2 は，それぞれひとつの要素とされるのが望ましい．

このような背景に基づき，図 4 に示すような手順で論理構造の抽出を行う．入力は，OCR ソフトウェアの処理結果とする．ディジタル文書の場合は，その文字認識結果にテキスト抽出の結果を重ねあわせて文字認識誤りの訂正を行うことを考えている．

基本要素抽出 OCR ソフトウェアのレイアウト認識の誤り訂正（上述した 3 種類の誤りの訂正）を行い，基本要素を抽出・作成する．

論理種別注釈 得られた基本要素に論理構造の観点からの種別を注釈づける．

論理構造構築 論理種別を注釈づけられた基本要素の並びから論理構造を得る．

4.2 コーパス

これらの処理の仕様検討と評価を目的に，小規模なコーパスを作成した．2003, 2006, 2009, 2013 年からほぼ同数をプログラム構成に基づく種別のバランスのみ考慮して無作為抽出した言語処理学会年次大会予稿 100 件を対象とし，まず，それら文書の e-typist のレイアウト認識の結果を人手により基本要素として適切なものを矩形領域とするように修正した．修正は，前述の「誤り」に対応して以下の 3 つの方針に基づく．

1. 改行で区切られた本文中の式や素性構造表現等については，本文と異なる領域とする，節のタイトルは本文から分離するなど，原則として分割の方向で，基本要素として適切な矩形領域へと修正する．適切な基本要素ということで，これらの矩形には論理種別（後述するように表 1 の `type` 属性の値として示される）のいずれかを付与することができる．

2. 図や表を，図に分類されるひとつの矩形領域とする．それぞれのタイトルは異なる領域とする．

表 1: 論理種別の注釈

属性	値	説明
type	header	ヘッダ
	page	ページ番号
	footer	ページ番号以外のフッタ
	title	論文タイトル
	auth	著者情報（所属等も含む）
	abst	梗概
	stitle	セクション（節）タイトル
	sstitle	サブセクションタイトル
	ssstitle	サブサブセクションタイトル
	body	本文
	list	箇条書き（全体）
	listitem	箇条書き項目
	footnote	脚注
	equ	数式
	fig	図
	tab	表
	figcap	図タイトル
	tabcap	表タイトル
	note	図表註釈
	ack	謝辞（全体）
	acktitle	謝辞タイトル
	ackbody	謝辞本文
	reftitle	参考文献タイトル
	refbody	参考文献本体（全体）
	refitem	参考文献項目
par	whole	全体（デフォルト値）
	first	先頭部分
	mid	中間部分
	last	末尾部分

3. 多段組を前提とした不必要な分割については，可能であれば統合を行う[6]．

その後，矩形領域（=基本要素）に表 1 に示す論理構造に関連するふたつの属性の注釈付を行った．第一の属性 `type` は論理構造における要素の種類（論理要種別）を示すものである．第二の属性 `par` は論理構造の観点ではひとつの要素となるべきものが，物理的制約で分割されているか，分割されている場合は，そのどの部分であるかを示している．

表 1 に示されているように，論理要素の種別においては，箇条書き部分を本文から区別する等，その後の利用で必要と思われるものに対してやや細かい区分がなされている．また，箇条書きや参考文献等において，その項目（listitem, refitem）と全体（list, refbody）の 2 種類の種別を設定している．粒度を揃えるということでは，両方を基本要素とすることは問題であるが，これは自動で行われるレイアウト認識の結果の修正を最小限とするための配慮である．つまり，箇条書きや参考文献の部分をレイアウト認識すると，文書のスペーシングにより，全体がひとつの矩形領域とされる場合

[6] 利用している e-typist では，テキストに分類される領域について，自動認識結果を更に分割することは自由に可能であるが，統合については実行できない場合があり，完璧な修正となっていない場合がある．

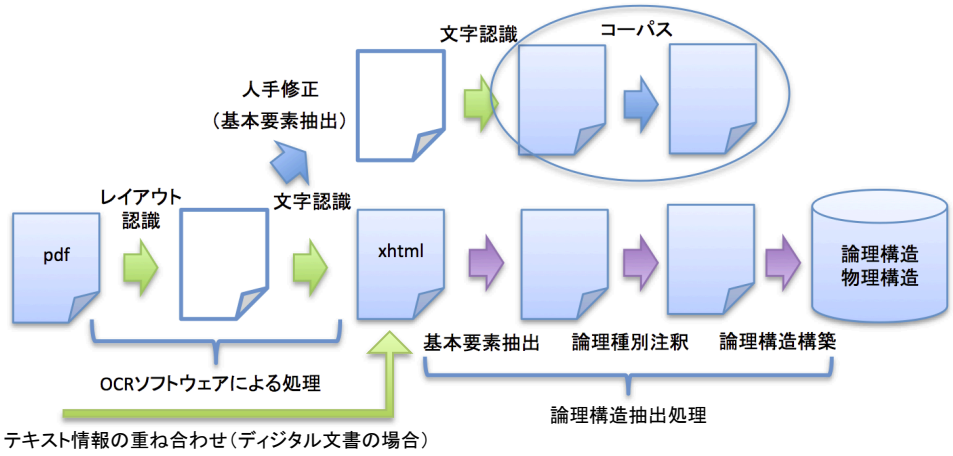

図 4: 論理構造抽出の枠組み

と，項目ごとに矩形領域とされる場合とがある．このいずれの場合も人手修正を行わず，異なる注釈を行うことで対応している．ただし，箇条書き部分が前後の本文と同じ領域とされてしまったり，一部の複数の項目だけがひとつの領域と認識された場合は，領域を分割することで修正を行っている（方針 1.）．

前述の論理構造抽出処理において，基本要素抽出は，レイアウト認識結果修正を模擬することに，論理種別注釈はその後の注釈の模擬に相当する．論理構造構築は，もしそこまでの処理が完璧であれば，単純なパージングであるが，そうでない場合は，処理誤りに起因するノイズへの対応や，場合によっては前段の処理へのフィードバックが必要になる．

4.3 実装

現在，基本要素抽出と論理種別注釈について実装を進めている．

基本要素抽出では，前述の 3 種類の誤りに対し，アルゴリズム的に修正を行っている．1. については，矩形の位置，先頭の文字種（先頭文字が空白であることによる字下げの認識を含む），行末における句点の存在，「謝辞」等のキーワードとの一致，等を用いて分割すべき境界の判定を行う．2. については，矩形の位置や大きさ，フォントの大きさ，矩形領域内の空白の割合等を用いて，テキスト領域ではない矩形を削除する．3. についても，同じ文書の別の部分の認識結果から推定される段組みのパラメータを前提として，不自然な横幅を持つ矩形が判定でき，その周囲にある矩形との位置関係から，統合すべきものが判断できることが多いので，それを用いて統合を行う．

テキストと分類された領域について，その効果を測ると，自動レイアウト認識の結果と人手修正後のコーパスとでは，文書毎のマクロ平均で，精度（修正が必要ない矩形数/自動認識結果での矩形数）が 0.58，再現率（修正されていない矩形数/人手修正後の矩形数）が 0.63 であるのに比較して，自動レイアウト認識結果に基本要素抽出を施したものは，人手修正後のコーパスに対して，精度（両者に共通する矩形数/基本要素抽出後の矩形数）は 0,79，再現率（両者に共通する矩形数/人手修正後の矩形数）は 0.75 と向上する．クローズドテストであり，2013 年のものを主に参照して開発したため，それらについては精度 0.89，再現率 0.90 と高い性能が得られる．一方で，2003 年のスキャン文書については，段組みを誤認識する等，致命的な誤りを含むものも多く，よい結果が得られていない．また図表や式については，複数のテキスト領域と誤って認識されたものから，そこに図表等が存在したことが復元される必要があるが，この処理は現時点では行っていない．

論理種別注釈は，コーパスを用いた機械学習を行い，CRF による系列ラベリングを行っている[7]．矩形領域の位置，先頭の文字種別等とバイグラムの情報を素性としている．10 分割交差検定で．表 2 に示す混同行列が得られている．ここでは，その後の応用を前提とした分類とし，list と listitem，stitle と sstitle 等はまとめている．また，コーパス中の論文には梗概（abst）を含むものが極めて少なかったため表に含めていない．全体の正解率は 87%である．

5 関連研究

PDF 文書からテキストを抽出し，検索を行う試みは幾つか行われている．阿辺川らは，抽出されたテキストと PDF 文書を用いて，参考文献へのリンクやキー

[7]CRF の実装は CRF++（http://taku910.github.io/crfpp/）を用いた．

表 2: 論理種別推定

正解＼推定	ack	at	au	bdy	equ	fig	fc	ft	fn	hd	lst	nt	pg	rb	rt	st	tab	tc	tt
ack	7			3							2								
acktitle (at)		1		1							1				9				
auth (au)			206	1												3			
body (bdy)	1			1860			7		7		157			8		5	1	9	
equ					159	27											10		
fig			1		24	217											36		
figcap (fc)			1	7			240				5							17	1
footer (ft)								107											
footnote (fn)				10					94		13					4			
header (hd)										76									
list (lst)		2		183		2	11		13		676			5	19	32		21	
note (nt)				1			3		1		6	4	1				1	3	
page (pg)							1						301						
refbody (rb)				7					1		19			96	1	1			
reftitle (rt)		1		2							11			1	85	3			
stitle (st)				2		1	1		3		11					1117			
tab					16	33											229		
tabcap (tc)				11			27				9							233	
title (tt)																			96

ワードへの脚注を備えた閲覧システムを実現している [1]．ACL Anthology[8]を対象に，統語解析可能なテキストを得るために，ディジタル文書，スキャン文書の解析が試みられている [3, 13, 14]．得られたテキストを統語意味解析し，意味に基づく検索を実現することがその目的である．増田らは，テキストマニングの対象として，OCR 読み取りを用いたテキストを利用している [10]．数式等を含めたより高精度な復元処理が磯崎によって検討されている [7]．

文書の構造認識については，Klink らや Luong らの研究がある [8, 9]．ここでも CRF を用いて，文書の構成要素からなる論理構造を明らかにしているが，検討されているのは論理種別注釈に相当する部分で，レイアウト認識の誤りに対する処理は含まれていない．文書の構造を利用するという点では前述の阿辺川のシステムに加えて，難波らが引用情報を解析して，その役割を利用した構造化を行っている [11]．

6 おわりに

文書構造に基づく対話的情報アクセスの枠組みを提案し，そのための文書表現を構築するために必要になる文書の論理構造解析について現状を報告した．提案した枠組みはまだ構想段階に留まっており，今後，以下の検討が必要と考えている．

研究論文等に対する検索意図の収集と分析 3 節で考察した検索意図の分類について，現実の検索意図を収集する等を通じて，詳細化を行い，それらの検索意図に応えるための照合方式を検討する．現在想定している文書表現がそのような照合方式に充分であるかを確認する．

閲覧等，インタラクションの枠組み設計 2 節の枠組みにおいて，まだ十分に検討されていない対話的な情報アクセスについて，文書とその部分の行き来や論理構造と物理構造の行き来等，これまでにはない焦点の移動について検討し，基本的な操作を明らかにする．

論理構造の抽出の精度向上と実現 4 節で提案している方式について引き続き検討を進め，どの程度の精度が得られるかの見通しを得る．それを受けて，文書表現の作成にどの程度の人手介入を必要とするか等を考慮に入れて，システム全体の設計を進める．また，現在では異なる方針で実装している基本要素抽出と論理種別注釈について枠組みの融合が可能かを検討する．

いずれも小さくはない課題であるが，順次検討を進めていきたい．

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

参考文献

[1] 阿辺川武, 相澤彰子: 脚注表示機能を備えた論文閲覧システム Sidenoter, 『言語処理学会第 20 回年次大会予稿集』, pp. 796–799 (2014).

[2] Bates, M.J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface, *Online Review*, Vol. 13, No. 5, pp. 407–424 (1989).

[3] Berg, Ø., Oepen, S., Read, J.: Towards High-Quality Text Stream Extraction from PDF. Technical Background to the ACL 2012 Contributed Task, *Proc. of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 98–103 (2012).

[4] Callan, J.P.: Passage-Level Evidence in Document Retrieval, *SIGIR '94*, pp. 302–310 (1994).

[5] Ellis, D.: A Behavioral Approach to Information Retrieval System Design, *Journal of Documentation*, Vol. 45 No. 3, pp. 171–212 (1989).

[6] Hearst, M.A., Plaunt, C.: Subtopic Structuring for Full-Length Document Access, *SIGIR '93*, pp. 59–68 (1993).

[7] 磯崎秀樹: PDF 中の TEX 記号の復元と ACL Anthology への適用, 『言語処理学会第 19 回年次大会予稿集』, pp. 956–959 (2013).

[8] Klink, S., Dengel, A., Kieninger, T.: Document Structure Analysis Based on Layout and Textual Features, *Proc. of International Workshop on Document Analysis Systems, DAS2000*, pp. 99–111 (2000).

[9] Luong, M., Nguyen, T., Kan, M.: Logical Structure Recovery in Scholarly Articles with Rich Document Features, *International Journal of Digital Library Systems*, Vol. 1, No. 4, pp. 1–23 (2010).

[10] 増田勝也, 丹治信, 植松すみれ, 美馬秀樹: 研究動向分析のための論文のデジタルテキスト化とマイニングシステム, 『言語処理学会第 20 回年次大会予稿集』, pp. 792–795 (2014).

[11] 難波英嗣, 神門典子, 奥村学: 論文間の参照情報を考慮した関連論文の組織化, 『情報処理学会論文誌』, Vol. 42, No. 11, pp. 2640–2649 (2001).

[12] Salton, G., Allan, J., Buckley, C.: Approaches to Passage Retrieval in Full Text Information Systems, *SIGIR '93*, pp. 49–58 (1993).

[13] Schäfer, U., Read, J., Oepen, J.: Towards an ACL Anthology Corpus with Logical Document Structure. An Overview of the ACL 2012 Contributed Task, *Proc. of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 88–97 (2012).

[14] Schäfer, U., Weitz, B.: Combining OCR Outputs for Logical Document Structure Markup. Technical Background to the ACL 2012 Contributed Task, *Proc. of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pp. 104–109 (2012).

[15] Tellex, S., Katz, B., Lin, J., Fernandes, A., Marton, G.: Quantitative Evaluation of Passage Retrieval Algorithms for Question Answering, *SIGIR '03*, pp. 41–47 (2003).

[16] 上田修一, 倉田敬子: 『図書館情報学』, 勁草書房, pp. 217–218 (2013).