NAIST-IS-DT9761019

博士論文

ディジタル図書館のための文書主題の抽出とその検索への応用

堀井 千夏

2000年3月24日

奈良先端科学技術大学院大学
情報科学研究科 情報処理学専攻

本論文は奈良先端科学技術大学院大学情報科学研究科に
博士 (工学) 授与の要件として提出した博士論文である.

堀井 千夏

審査委員： 千原 國宏 教授
植村 俊亮 教授
横矢 直和 教授
今井 正和 助教授
砂原 秀樹 助教授

ディジタル図書館のための文書主題の抽出とその検索への応用*

堀井 千夏

内容梗概

本研究は，単語の概念を求めることでその意味のルーツ (意味素性) を探索する新しい文書主題の抽出法を提案し，大量に氾濫する電子化情報を有効に活用することを目的とする．

近年，ネットワーク上で提供される情報源は急速に多様化，大規模化し，利用者が必要な情報だけを取出すことは非常に困難になっている．利用者が自分にとって有効な情報を探し出し，選別できなければその情報は存在しないに等しい．蓄積された情報を有効に活用するには，まず，その情報が表す内容を正確に抽出しておく必要がある．本手法では，単語情報に加えて概念情報を用いることで深層的な観点から文書の主題を抽出し，文書内容の表現についての充実を図った．文書の主題は，単語と概念の出現頻度分布から文書の強調度と表現度を求め，両者に基づいた特徴ベクトルを用いて定量的に表現した．提案手法の有効性を検証するために，技術論文の主題を概念で獲得する評価実験を行った．その結果，従来の文字列に着目した抽出法では単語の意味まで絞り込むことは困難であったが，単語の概念情報から筆者が論文中で用いた意味を推測することが出来た．また，アンケート質問 “本手法で抽出した主題が妥当であるかどうか” に対する調査の結果，76%の被験者が “excellent”，“good” と評価しており，本手法で抽出した主題が被験者の想定する主題とほぼ一致していることがわかった．

さらに，２１世紀型図書館として注目されている「ディジタル図書館」の文献検索を取り上げ，文書の主題や利用者の検索意図を推測した新しい検索を実現し

*奈良先端科学技術大学院大学 情報科学研究科 情報処理学専攻 博士論文，NAIST-IS-DT9761019, 2000年3月24日.

た．文献の電子化が活発に進められているディジタル図書館では，膨大な文献情報と利用者との仲介役として検索意図に適した文書を提供する文献検索の実現が大きな課題とされている．本論文では，提案する内容抽出法の応用として，文書の主題や利用者の検索意図を推測する意味検索を提案し，本学のディジタル図書館で保管している文書データに適用した試作システムについて述べた．

キーワード

文書主題の抽出，概念情報，ディジタル図書館，意味検索

Extraction of Subject from Documents for Digital Libraries and its Application for Information Ritrieval*

Chinatsu Horii

Abstract

This paper describes a new method to extract subjects from documents with searching for the roots of a word in the meaning of conceptual information. The purpose of the method is to make good use of the scattering huge vault of document on the network.

Recently, as the volume and the variety of information resources on the network is increasing rapidly, it becomes very difficult to obtain desired information for the user's needs. It is necessary and indispensable to know what the documents mention in order to use of information effectively. In this approach, an emphasis factor and an expression factor are defined as a characteristic of the document. Both of them are estimated from the frequency of the appearance of the words and their concepts. The subject is determined by means of the characteristic vector based on these factors. As an evaluation of this approach, I performed the several experiments to extract the subject of technical papers with concepts. The results show that the proposed method has capability to express the meanings of author's intention. To evaluate effectiveness of this method, I investigate the questionnaire survey, whether the subject of these papers with concepts are represented or not. 76% of answers admit that the proposed method extracts the subject of technical papers with concepts.

*Doctor's Thesis, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DT9761019, March 24, 2000.

In addition, I discuss how to process the documents for information retrieval in digital library, which is a new library in the 21st century. This paper presents a new retrieval method which is one of the examples applied the proposed algorithm of extraction of subject. This retrieval method realize the semantic retrieval estimating the subject of document and retrieval intention. I present a prototype system implemented into digital document data stored in NAIST Digital Library.

Keywords:

Extraction of Subject,Conceptual Information,Digital Library, Semantic Retrieval

目 次

図目次

図目次

表目次

第1章

序論

本章では，本研究の目的を明らかにし，提案する手法の概要と論文の構成について述べる．

1.1 大量文書データからの情報収集

近年における蓄積メディアの大容量化により，雑誌や書籍などの文書情報が大量に電子化されている．また，インターネットに代表されるコンピュータネットワークの普及やマルチメディア技術の著しい発展に伴い，文書だけでなく映像や音声等といった多種多様な情報へのアクセスが可能になってきた．電子化による情報の提供は，紙による印刷物に比べて軽量かつ品質に劣化がないだけでなく，時間や場所に拘束されることなく情報を即座に入手することができる．この利便性から利用者は急増し，その分野も情報関連から様々な分野へと広がっている．情報の電子化による利点は多いが，その一方で，様々な場所で蓄積された電子化情報は複雑で大規模なものとなり，その情報量は急増している．すでにネットワーク上は情報過多状態であり，利用者は蓄積情報の中から自分に必要な情報だけを取り出すことが困難になっている．情報の探索には WWW 上で提供されている様々な大規模情報検索システムが利用されているが，その検索結果として収集された情報量は非常に多く，その中から必要な情報だけを分別することはもはや人間の処理能力では不可能である．特に，情報の記録には記述形式の柔軟性や蓄積データ量の観点から文書形式が最も利用され，ネットワーク上で電子化されている情報の大部分は文書形式である．そのため，利用者は情報の内容を読んで確認する必要があり，大きな負担となる．これはネットワーク上だけではなく，個人が保管している情報に対しても同様なことが言える．長期間に渡って情報を蓄積

していれば，当然，個人で保管する情報量も増えてしまう．収集した情報の適切な整理・分類には大変な労力と時間が必要であり，もはや個人レベルでさえ情報が氾濫し，利用者は迷子になっている始末である．

このような大量文書情報の取り扱いに関する問題を解消するために，情報の収集（情報検索や情報フィルタリング）や情報の統合（情報分類や組織化）に関する様々な研究が行われている [1][2][3][4][5]．これらの研究では文書データの内容をシステムが適切に解析することが共通した課題であり，文書内容の分析とその抽出法が重要な鍵となっている．本論文では，大量の文書データを適切に処理するために必要な文書主題の抽出法を提案する．

1.2 本研究の目的と提案手法

1.2.1 研究目的

本研究の目的は，単語の概念情報から単語がもつ意味のルーツ（意味素性）を探索する新しい文書主題の抽出法を提案し，大量に氾濫する電子化情報を有効に活用することである．さらに，21 世紀型図書館として話題になっているディジタル図書館の文献検索に提案手法を適用し，文書主題や利用者の検索意図を推測する新しい検索を実現する．

1.2.2 提案する文書主題の抽出法

本研究では，大量な文書データから利用者に適切な情報を提供するために必要な文書主題の抽出法を提案する．

文書の主題を抽出するには，多義語や同義語といった単語の意味についての曖昧性を解消し，単語が文書中でどのような意味で使用されているのかを求める必要がある．例えば多義語である“スプリング”という単語は，1）ばね　2）春の二つの意味が存在し，文書中で使用されている意味を解析しなければならない．また，同義語である“コンピュータ”は，1） 電子計算機 2）パーソナル・コンピュータ 3）PC などで言い換えることができ，これらを同一の単語として処理する必要がある．本手法では，このような単語がもつ意味の曖昧性を解消するた

めに，単語の概念情報を利用した手法を提案する．概念情報の取得にはEDR電子化辞書[6]を用い，文書の出現単語がもつ意味を概念で表現する．単語が表現する具体的な内容が“単語の意味”であるのに対し，“単語の概念”とは，単語の同義・類義語が共通して表す内容を指す．例えば，“馬車”や“バス”，“自動車”の単語はそれぞれ複数の意味をもつが（バスは「乗り合い自動車」,「風呂」,「コントラバス」の意味をもつ），その中で共通して表現される内容は“乗り物”である．単語の概念は，さらに概念で表現することが出来る．例えば，“乗り物”の概念は“ものを運搬する器具”となる．本手法では，単語がもつ全ての意味について概念を再帰的に求め，概念空間を構築する．この概念空間を上位層へたどることで文書の出現単語がもつ意味のルーツを探索し，概念を最小単位とした文書主題の解析を行う．具体的には，概念の出現頻度分布から主題が表す内容の“表現度”として単語の概念に関する重みを求め，文書の内容として重要な概念だけに絞り込む．さらに，主題の内容をどれぐらい強調しているかを示すために，主題の“主張度”として文字列の出現頻度分布から単語の出現に関する重みを求め，表現度と主張度の両者を用いた特徴ベクトルから文書の内容を定量的に評価することで文書の主題を抽出する．

提案する文書主題の抽出法は，様々な情報活用に適用することができる．本論文では，その一つとして，ディジタル図書館の情報検索を取り上げ，概念を用いた検索手法として“意味検索”について述べる．意味検索は，従来の全文検索手法のような単語体系に基づいた検索手法ではなく，文書の主題や利用者の検索意図を推測して検索に用いる．そのため，単語のうらに隠された意味や検索意図に対して不足している情報を推測することが可能になり，図書館司書の協力なしに利用者の要求に適切な検索結果を提供することが期待される．

なお，文書内容の抽出法については第3章で，検索への応用については第4章で詳しく述べる．

1.3 本論文の構成

本論文は以下の5章からなる．

第2章では情報活用の現場としてディジタル図書館を取り上げ，その現状と問題点を明らかにする．ディジタル図書館における問題のうち，その要となる文献検索に焦点をあて，文書の主題抽出の必要性と関連研究について述べる．

第3章では提案する文書主題の抽出法について述べる．本手法では，概念情報を用いることで深層的な観点から文書の主題抽出を行ない，文書内容の表現についての充実を図る．文書の主題は，単語と概念の出現頻度分布から文書の主張度と表現度を求め，両者に基づいた特徴ベクトルを用いて定量的に表現する．提案手法の有効性を検証するために，技術論文の主題を概念で獲得する実験を行い，実験結果についての評価と考察を行う．

第4章では提案手法の応用として，意味検索と文書構造に基づいた内容抽出の2例について述べる．意味検索では，検索質問を概念で表現する検索式部，科学技術論文の主題を選定する索引語作成部，および検索式と索引語を照合する照合部から成る検索システムの試作について述べる．また，文書構造に基づいた内容抽出では，文書を六つの構造に分割し，各構造部分の特徴抽出について述べる．

最後に，第5章では概念に基づいた文書主題の抽出法に関して総括する．

第2章
文書の主題抽出における背景とその関連研究

本章では，新しい情報活用の場として注目されているディジタル図書館を取り上げ，その特徴と問題点について明らかにする．さらに，ディジタル図書館における最も重要な問題点の一つである文献検索に焦点をあて，文書主題の抽出の必要性と関連研究について述べる．

2.1 ディジタル図書館の現状と問題

近年におけるネットワークやマルチメディア技術の著しい発達に伴い，情報提供の窓口となる「ディジタル図書館」の機能向上と実用性に関する研究が盛んに行われている．本学（奈良先端科学技術大学院大学）においても，平成8年4月よりディジタル図書館の運営が開始され，データベースの構築や検索技術の開発を中心とした研究が進められている [7]．以下では，本研究の背景となるディジタル図書館の現状について述べ，現在抱えている問題点について検討する．

2.1.1 ディジタル図書館の目的と新しい機能

1990年代に入り，米国は国家プロジェクトとして情報ハイウェイ構想を提案した．この構想は2000年までに米国全土における教育機関や医療機関をインターネットで接続するというNII（National Information Infrastructure：国家情報基盤）計画と，これを世界的規模に発展させようとするGII(Global Information Infrastructure：世界情報基盤) 計画の二つからなり，「全ての人が適正な料金で，

場所・距離等に関係なく最高の授業が聞け，必要な時に必要な場所で順番を待つことなく医療サービスを受けることが出来る社会が到来する」と説明されている[8]．この情報ハイウェイ構想は，高速・広域ネットワーク上における多種多様な情報を世界中でやり取りすることが目的である．ネットワークを介して，いつでも，どこからでも，あらゆる情報を利用できることは，まさに，その場に居ながらにして利用できる大規模な図書館の実現であり，この構想を有意義に利用できる分野の一つとしてディジタル図書館が上げられている．

従来型の図書館における主な業務内容は以下の6項目であるが[9]，ディジタル図書館では，この従来業務には無い新しいサービスの提供や運営形態の改革を目指している．

- 図書・資料の貸し出し／返却
- 検索・分類のための目録作成
- 書籍の管理・保管
- 利用者の検索要求への対応
- サービス拡大への対応（CD-ROM, ビデオ，マイクロフィルム等）
- 図書・資料の複製等による劣化防止

世界中の大学や公立図書館，あるいは個人が各々のディジタル図書館を運営しており，そこで所蔵するコンテンツも図書や雑誌だけでなく写真，絵画，地図など多岐に渡っているため，ディジタル図書館の活動は一様ではないが，これらは共通して「利用者の利便性」と「管理・蓄積の効率化」に基づいて多様化している．本学のような大学附属図書館について述べれば，新しいサービスとその内容は図2.1に挙げるような6項目が挙げられ，図書・資料等の電子化による情報の共有化・非劣化や窓口業務の自動化など，従来型の図書館の業務体系を大きく改善している[10]．近年では，CD-ROMやネットワークを媒体とした電子ジャーナルの出版や販売が盛んに行われ，電子化作業をすることなく保管することが可能であり，また，ネットワーク上で公開されているOPAC(Online Public Access Catalog)の電子化目録を利用することもできる．この他にもディジタル図書館の運営や機能の充実に向けて世界中で活発な研究開発が行われている．

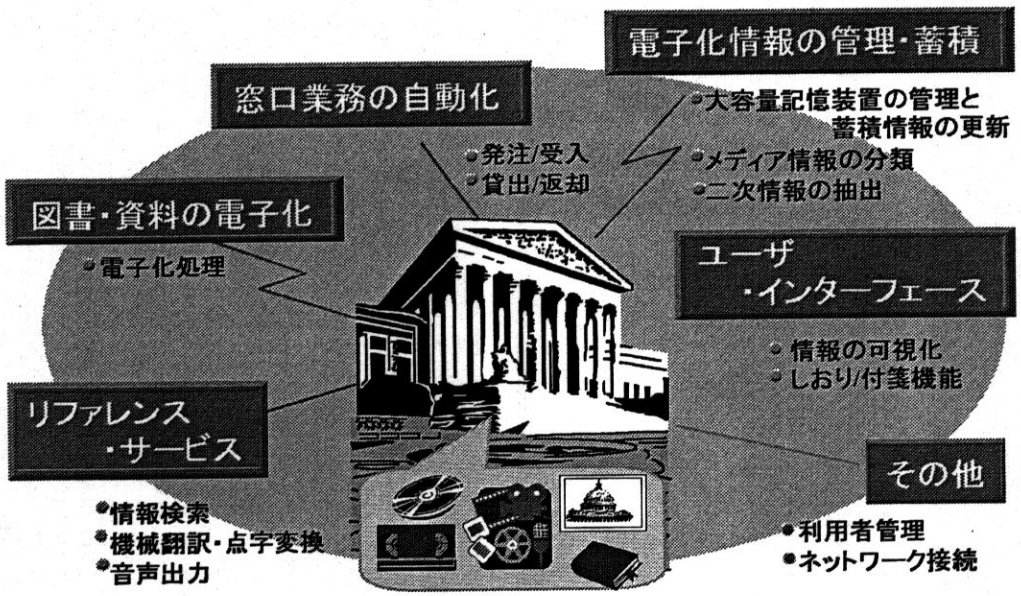


図 2.1 ディジタル図書館の新しいサービスとその内容

2.1.2 ディジタル図書館プロジェクト

ディジタル図書館の実現に向けて多くのプロジェクトや大学図書館が活動を進めている．その中から情報収集やデータベースの活用に特徴のある国内外の主要な活動をいくつか取り上げ，その特徴について以下に紹介する．

国内のディジタル図書館

青空文庫（図 2.2）

ジャーナリストの富田倫生氏らによって，芥川龍之介や夏目漱石などの著作権が消滅した（著者の死後 50 年経過）文学作品や権利継承者の了解を得た図書データを中心にデータベースを構築している．これらの文書データは，作家名や作品名で検索することができ，自由にダウンロードして読むことができる．また，数多くの作家が残した日記や創作ノート，下書き，草稿，未完成作品など全集以外では読むことができない作品についても収集しており，出版者の観点からデータベースを構築している [10][11].

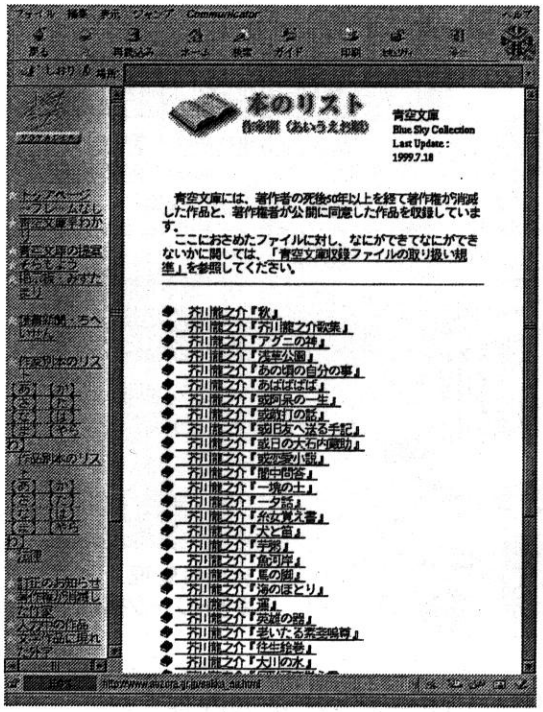


図 2.2 国内のディジタル図書館 (1)：青空文庫

図 2.3 国内のディジタル図書館 (2)：ディジタルミュージアム

国会図書館

国立国会図書館は,「パイロット電子図書館プロジェクト」,「児童書の電子図書館」,「アジア文献情報データベース提供システム」,「G 8 電子図書館プロジェクト」,「「電子化英文政府刊行物公開実験」,「国会会議録全文データベース」の六つのプロジェクトを行っており，高速・広帯域ネットワークを用いて，遠隔地から都道府県立および政令指定都市図書館の総合目録のデータベースにアクセスすることができる．また，国会図書館が所蔵する明治刊行図書や第二次対戦前後の刊行図書，国会審議用調査資料など 1000 万ページにものぼる貴重書についての文書データを閲覧することが可能になる予定である．現在，その実現に向けて実用性や問題点について検証実験を行っている [12].

次世代電子図書館システム研究開発事業

国策プロジェクトとして通商産業省が主導している次世代電子図書館システム研究開発事業は，NEC，沖電気，東芝など民間企業9社との共同で，次世代の情報流通基盤の研究開発とその具体的なアプリケーションの研究開発を行っている．その特徴としては，利用者のフィルタリング条件に従って不要な情報を取り除く情報フィルタリング機能や映像情報をMPEG7で保管した映像情報検索，メモ・しおり機能などの電子読書支援などがあげられる．また，利用者の資格チェックやクレジットカード番号の暗号化など，知的財産権やプライバシーの保護などの観点からも検討を進めている [13][14].

奈良先端科学技術大学院大学

本学，奈良先端科学技術大学院大学は，三つの研究科（情報科学，バイオサイエンス，物質創成）から成り，附属図書館では各分野の図書や学術雑誌，学内論文を中心に保管している．記憶システムには，アクセス速度の速い磁気ディスクと低コストの磁気テープを併用したマイグレーションシステムを用いており，あらかじめ利用者がキーワードを登録しておけば，附属図書館の新着情報を電子メールで報せてくれる「アラート機能」などがある [15][16].

東京大学 ディジタルミュージアム（図2.3）

従来，書籍は図書館，美術品は美術館，自然・文化的な資料は博物館が所蔵するなどその区別は比較的はっきりしていたが，ディジタル図書館で動画や写真，絵画などが所蔵できるようになり，美術館，博物館の役割をディジタル図書館が担うことも可能になってきた．美術館や博物館はただ所蔵品を展示するのではなく，独自の解説や閲覧サービスを提供する必要がある．東京大学の「ディジタルミュージアム」は，ディジタル技術を活用した新しい博物館を目指している．所蔵品の3次元データ，CADデータ，表面のテクスチャー・データ等といった外観再現データに加えて，X線CT，釉薬の化学分析に関するデータベースの実現を試みている．現在，電子化された貴重な古文書や生体標本など600万点を保存し，その紹介論文や解説を博物館の観点で掲載している [17].

その他

九州大学付属図書館では,「OPAC 横断検索サーバ」として他大学のディジタル図書館40ヶ所に対して目録検索を提供している [18]. また，学術情報センターでは，電子図書館の試行実験を経て 1997 年 4 月から事業としての運営を開始しており，検索ソフトウェアの無料配布，利用者を限定した課金制度，利用者から依頼された資料をコピー・郵送するドキュメントデリバリー (Document Delivery) などが試行されている [19][20].

海外のディジタル図書館

プロジェクト・グーテンベルク（図 2.4）

プロジェクト・グーテンベルクは，1971 年，米国イリノイ大学 マテリアルリサーチ・ラボのマイケル・ハート教授によって始められた.「不思議の国のアリス」,「イソップ物語」などの童話や「聖書」，シェークスピアの作品の他に，シソーラスや百科事典など 2000 タイトルの古今書物が電子化され，ネットワーク上や CD-ROM で提供されている [10][21].

カーネギーメロン大学（図 2.5）

NSF（国立科学財団），DARPA（国防総省），NASA（宇宙開発局）から研究助成を受けているプロジェクトの一つであり [22]，オンラインビデオライブラリを実現している．音声認識，画像理解，自然言語処理技術を用いたビデオコンテンツの分類や索引づけの他に，ビデオストーリの分割，要約，重要な画像フレームの抽出を行っている．[23].

カリフォルニア大学（バークレイ校）（図 2.6）

カリフォルニア州に関する大量な環境データベースを構築している．23,664 点の植物や動物，人，景色の画像を検索することができ，カリフォルニアで生息する植物について 300,000 ページの解説文書が電子化されている．この他にカリフォルニアの上空から撮影した航空写真を保管し，画像中の色やテクスチャ，シンメトリなどの画像特質を用いてオブジェクトを抽出・認識し，これを画像検索に用いている．画像処理の解説論文についても提供している [24].

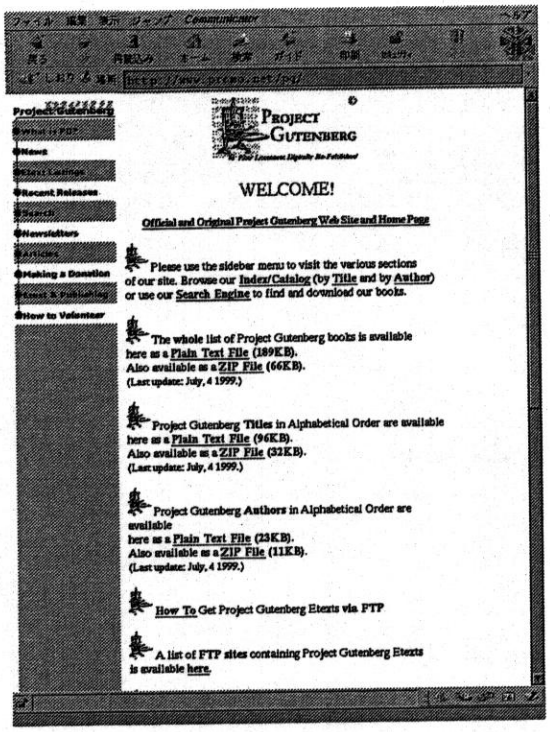


図 2.4 海外のディジタル図書館 (1)：プロジェクト・グーテンベルク

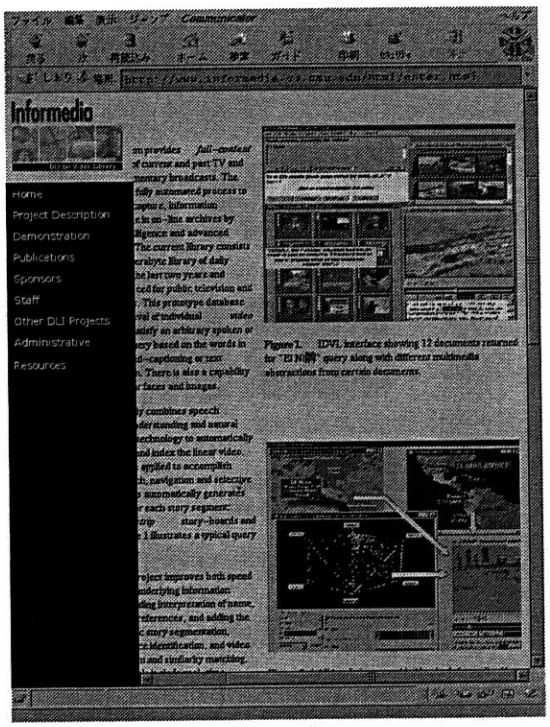


図 2.5 海外のディジタル図書館 (2)：カーネギーメロン大学

カリフォルニア大学（サンタバーバラ校）

カリフォルニア大学サンダーバーバラ校を中心とする Alexandria Digital Library は地図と画像を中心とした地理情報データベース Alexadria を開発し，地名索引や緯度，経度による検索システムや位置指定を地図上で選択する検索支援システムを提供している．この他に，地理情報に関するシソーラスの構築や地図データのためのメタデータについての研究が盛んである [25].

ミシガン大学

ミシガン大学では，情報に対して概要（話題，形式，利用者等）を付加し，インターフェース・エージェント，メディエータ，コレクション・エージェント三つで構成される知的エージェントシステムを用いて，利用者に応じた

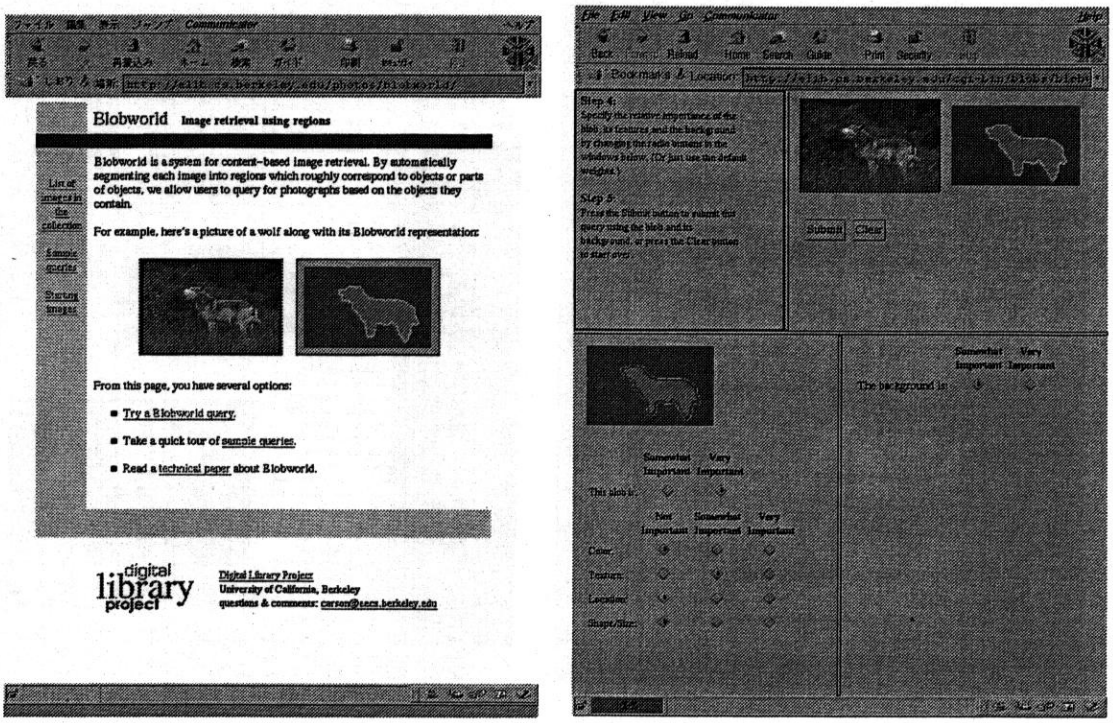


図 2.6 海外のディジタル図書館 (3)：カリフォルニア大学 (バークレイ校)

検索を試みている．電子化されたドキュメントには，研究教育を目的とした地球・宇宙科学分野，人文科学分野などが幅広く蓄積・提供している．また，Elsevier 社による 1100 冊の雑誌文献 (TULIP) のための索引・書誌情報による検索システムも提供している [26].

The WWW Virtual Library

バーチャル・ライブラリは，World Wide Web の発案者ティム・バーナーズ・リーが「最古のカタログ」として始めた．現在はジェラード・マニングによって管理され，ボランティアの協力によって支えられている．情報収集されたカタログには，コンピュータや教育，ビジネス．法律，社会など非常に多くのカテゴリーが用意されており，これらのカテゴリーは，各分野の専門家や研究者によって分類され，その内容も学術色の濃いものである [10][27].

アメリカン・メモリー

世界最大の地図コレクションを所蔵する米国議会図書館のプロジェクトである．2000年までに500万枚の地図を電子化し，全米の主要な図書館と連携して巨大なディジタル図書館を目指している．すでに，10,000枚の地図データを収集している．この他に，流行歌の楽譜や南北戦争の写真，歴史的建造物など人々の暮らしをうかがうことのできる画像が多く保管されている．[10][28].

その他

カトリックの総本山バチカンでは，カトリック関連書など20,000点のコレクションを電子化し[29]，カナダ[30]，イギリス[31]，フランス[32]，中国[33]など世界中の国立図書館がディジタル図書館のプロジェクトを立ち上げ，国有コレクション等の電子化情報を公開している．

2.1.3 ディジタル図書館が抱える問題

世界中で，ディジタル図書館の実現に向けて大規模なプロジェクトが進められ，その規模に関わらず，運営基盤を標準化する活発な動きもある．しかし，ディジタル図書館は情報活用の場として新しい枠組みであり，また，様々な分野の最新技術を土台とした大規模な応用システムであるため，未だ解決されない問題が多岐に渡って残されている．この問題は大きく分て「計算機・通信技術」,「情報の蓄積・管理」,「情報検索」,「ユーザインターファエース」,「知的財産権」の五つに分類することができ，以下にその問題の内容について詳しく述べる[10] [34] [35].

計算機・通信技術

ディジタル図書館では，書籍のページ画像だけでなくビデオの動画といった容量の大きなマルチメディアデータを大量に蓄積するため，そのデータを高速に通信・処理できることが必要不可欠である．多くのディジタル図書館では，最新の高速・広域ネットワーク技術やマルチメディア通信技術が駆使されているが，保管されるデータは増加する一方であり，通信・処理技術に満足することはできない．また，最新技術は試験的な部分が多く，そのプロトコルデザインやデータ圧縮伸張技術などは標準化されていない．そのため何

をディジタル図書館に適用するべきかについて試行錯誤的な要素を拭えない．

情報の蓄積・管理

一次情報の電子化

現在のディジタル図書館は，従来から所蔵している図書・資料と，新たに電子化した一次情報とが混在している．最新刊の図書については，電子形態で出版されたものを利用すれば良いが，既存のものについては，電子化作業により作成しなければならない．ディジタル図書館は，図書や資料の他に，ビデオや音声などのマルチメディアデータも多く所蔵するため，多様な形式に対応した電子化作業が必要である．また，その作業内容や行程の標準化についての課題も多く，効率のよい電子化技術の開発が望まれている．

大規模分散データベースの構築

従来型の図書館では，同一の図書・資料を各図書館で個別に所蔵していなければならない．冊子を共有することはできないため，利用者は所蔵する図書館まで足をはこぶか，図書館に代行して借りてもらうしかない．一方，ディジタル図書館では，冊子体ではなく電子化された一次情報を扱うため，他のディジタル図書館と相互に参照することが期待される．しかし，ディジタル図書館が保管する一次情報はネットワーク上で広域に分散し，そのデータは従来型のように重複しているのが現状である．そのため，エージェントやインテリジェント・ゲートキーパーなどを用いて，世界中に分散した大量の情報の中から必要な情報を見つけ出し，大規模分散データベースを構築する必要がある．

情報の蓄積と管理

多くの利用者が図書館に対して求める情報は，図書や資料の名称ではなく，その内容についてである．従来型の図書館では取り扱う情報の最小単位が冊子体やメディアであり，その内容に関する情報は提供されていない．一方，ディジタル図書館では，一次情報が電子化されているので，文書理解や自然言語処理，画像理解，音声認識技術等を用いてその内容を解析し，利用者が

必要とする内容に再構築できる可能性があり，内容に関する情報提供の実現が期待される．しかし，従来とは異なる情報の蓄積・管理方法が必要となるため，情報認識・理解技術の他にも幾つかの課題が残されている．特に，図書館は一次情報と共に「一次情報の情報」として二次情報を保管するが，この二次情報の構造についても従来の書誌情報や目次といった形態ではなく，その内容に関する目録をつくらなければならない．また，こうした一次情報や二次情報の詳細化によりディジタル図書館のデータ構造はより複雑になり，それを記憶する装置も一層の大容量化を計らなければならない．現時点では効率のよい記憶方法に頼らなければならないが，近年にみられるような記憶媒体の技術進歩によれば，近い将来，大容量記憶装置に関する問題は解決できると考えられる．

情報検索

一次情報の電子化が進むにつれて，利用者が膨大な情報の中から必要な情報を取得することが困難になっている．従来型の電子図書館では，図書館司書との対話から利用者は情報を取得していた．しかし，ディジタル図書館では，検索システムによる情報に頼るしかない．ディジタル図書館が検索の際に参照する情報は，メディアの種類，書誌名，目録であり，一次情報の内容について触れていないのが現状である．利用者は必ずしも検索内容に詳しいとは限らない．新しい分野に直面して図書館を利用する場合も多いはずである．このような利用者は適切な検索質問を生成できないため，情報を探すことは不可能である．ディジタル図書館では，司書に代って内容に関する情報を提示できる検索システムが必要であり，また，利用者が検索質問を推測することができるような思考支援や，膨大な検索結果の中から利用者が目的とする情報を絞り込めることが求められる．

ユーザインターフェース

ディジタル図書館では，図書館司書による検索支援がないため，リファレンス・サービスには，あらゆる利用者を想定した支援機能が必要である．具体的には，計算機に不慣れな子供や老人のための技術支援や，目・耳・手の不

自由な人のための作業支援（点字変換，音声入出力機能），外国語または日本語が理解できない人のための多言語文書閲覧支援などが挙げられる．また，情報の可視化による利用者の理解支援やアニメーションなどによる馴染みやすさ，人間工学的な観点からのインターフェースなど，多くの工夫が必要である．

知的財産権

知的財産権はディジタル図書館を運営する上で避けることのできない問題である．ネットワーク上で情報を読むということはそれをコピーすることとほぼ同義だからである．多くのディジタル図書館のプロジェクトは，出版社に使用許可を得て公開している．そのため，従来の図書・資料の貸し出しは無料であったが，ディジタル図書館では情報課金システムを導入し，使用料を徴収せざるを得ない．しかし，情報の適切な価格設定や個人情報の保護に関する問題が多く残されている．

ディジタル図書館が抱える問題点は，上記にあげたように技術問題や運営問題，社会的問題など多岐に渡っているが，本論文では技術的な問題に着目する．ディジタル図書館における技術的な問題は，以下の四つにまとめることができる．

- 高速通信技術の実現
- 効率のよい情報収集
- 情報検索の充実
- ユーザインターフェースの開発

本研究では，ディジタル図書館を実現していく上で，最優先に解決しなければならない問題の一つとして「情報検索」を取り上げる．情報検索技術は，電子化情報と利用者とをつなぐ唯一の仲介役であり，また，その技術内容は情報の構造や管理形態，分類等に対して相互に関与している．そのため，情報検索の充実を図ることは，ディジタル図書館の機能を大きく前進させることであり，同時に，情報を処理する上で難しいとされている多くの問題を解決する糸口となる．次節で，文献検索技術の動向と問題について詳しく述べる．

2.2 ディジタル図書館で用いられる文献検索

2.2.1 キーワード検索と全文検索

ディジタル図書館は保管する情報量やメディアの多様化に対し，高度かつ高速な情報検索システムの実現を目指している．具体的には，以下のような技術開発が求められている．

文献検索技術 利用者の検索要求に対して妥当な文献を提示する技術である．蓄積ファイル構造，検索モデル，検索質問の支援，条件操作などが含まれる．大量な検索結果の絞り込み法やシソーラス辞書による関連語句への展開などの知的概念検索 [36] が期待されている．

マルチメディア検索技術 静止画像や動画像，音楽などのマルチメディア情報を検索する技術である．利用者の視覚や聴覚によって判断される内容を検索質問で表現する方法や，情報の特徴抽出に関する技術が求められる．

情報フィルタリング技術 検索結果の中から利用者が作成したプロファイルに従って，必要な情報だけをあらかじめ取り出して提示する技術である．フィルタリング結果を利用者が再検討し，その結果を新しいプロファイルとして更新していく．一時的に必要な情報ではなく，日頃から必要とする情報を得るのが目的である．

マルチプロトコル横断検索技術 マルチプロトコル横断検索とは，http や telnet といった異なるプロトコルを利用し，ネットワーク上で公開されているディジタル図書館へ検索サーバが利用者に代行して検索するシステムである．利用者はディジタル図書館の場所を意識すること無く検索結果を得ることができる．広域に分散した情報の統合化に関する技術が求められる．

エージェント応用技術 情報フィルタリングやマルチプロトコル横断検索の実現には，エージジェントと呼ばれる分散・協調作業を処理毎に行う技術が考えられている．ディジタル図書館においては，エージェント毎に検索処理を行い，エージェント間で問い合わせをすることで情報の統合を図る．エージェントに円滑な作業をさせるために，シソーラスを共通知識として与え，

自立したエージェント群（マルチエージェント）の実現を目指している．近年では，シソーラスの構築など知識獲得に関する研究が盛んに行われている[37][5][38].

高速並列検索技術　メディアの種類，情報の種類別のデータベースに対して，並列に検索処理を実行することで，大容量情報を高速に検索，閲覧できる技術である．

情報検索技術において，ディジタル図書館で保管する形態が図書や資料などの文書形式であることや，情報フィルタリングや横断検索などの基盤技術であること，利用頻度の高い技術であることから，特に，文献検索技術に関する発展が望まれている．

現在，ディジタル図書館で用いられている文献検索手法は，以下の二つに分類することができる．

- キーワード検索
- 全文検索（full-text search）

文献の主題を表す単語／句は，従来型では図書館司書が付与し，ディジタル図書館では検索システムが付与する．この語句を索引語（Index Term）といい，索引語の一覧を記録したファイルを索引ファイル（Index File）と呼ぶ．キーワード検索は，この索引語を“題名”や“概要”など文献の「一部の情報」を対象にして求めるのに対し，全文検索は，文献の「全ての文字列情報」を対象にして索引語を求めるといった違いをもつ．キーワード検索は時間やコストの省力，作業の簡略化などの利点から従来の図書館では広く用いられていたが，索引語が文献の一部にしか起因しないため，情報の欠損を原因とする検索もれが多く生じてしまう．一方，全文検索は，全文データの蓄積やその作業にかかるコスト，計算機能力の不足から避けられていたが，近年における比較的安価な大容量記憶媒体と高性能な計算機の普及や電子化作業の改善に伴い，キーワード検索に代って精度のよい検索手法として急速に拡大している．現在のディジタル図書館では全文検索が主に用いられており，文献全体を対象に質の良い索引語を抽出するための研究が広く行われている[39][40][41][42]．以下本論文では，全文検索を対象とする文献検索について述べる．

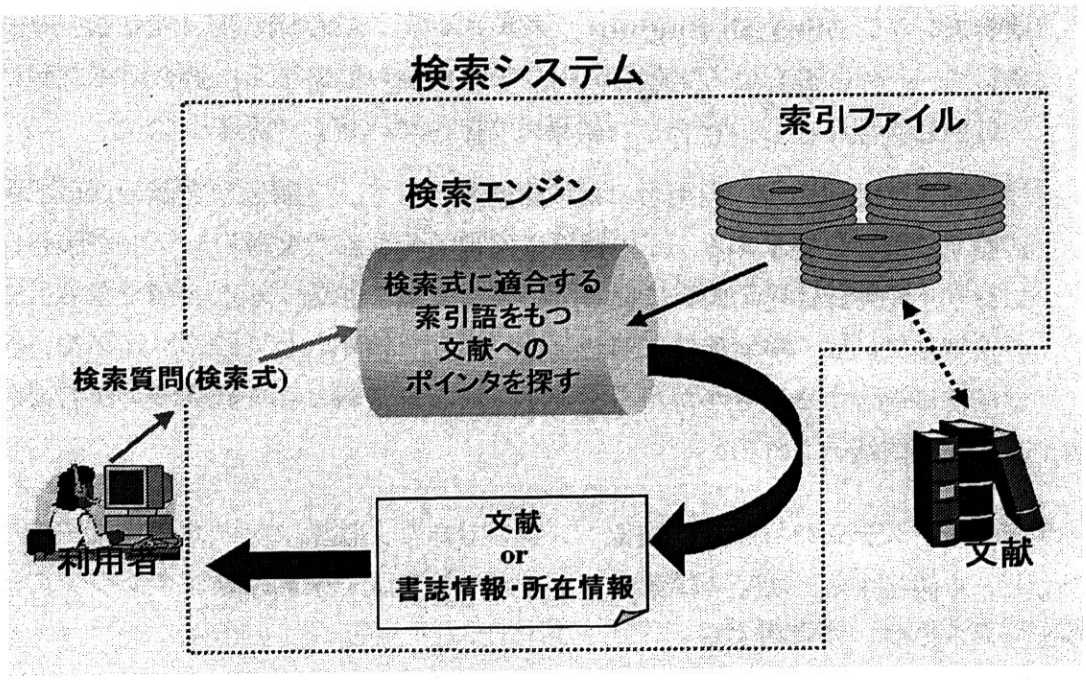


図 2.7 情報検索システムの構造

2.2.2 文献検索システムの構造

文献検索の一般的なシステム構造は図 2.7のように表され，各部の役割を以下に示す [43][44].

利用者（User） 利用者は自分が欲しい文献を特定する記述として検索質問（Request）を検索システムに与える．このときの検索質問は，1 個以上の語や句（Term），あるいは自然言語文で記述された文である．検索質問は，検索エンジンが問い合わせを行うための形式として検索式（Query）に変換される．システムが予め検索質問を定式化した場合，検索質問はそのまま検索式として用いられ，検索質問と検索式は同一となる．

索引ファイル（Index File） 検索対象となる各文献は，文書の特徴を的確に表す複数の語や句を索引語として付与する．この作業を索引付け (Indexing) と言い，索引語を効率良く検索に用いるために予め索引ファイルとして保管しておく．

検索エンジン（Search Engine） 検索エンジンは検索質問（検索式）を解釈して，それに適合する文献を索引ファイルから検索する．適合する文献は一般的に複数存在し，適合した結果は文献へのポインタ集合となる．

検索結果（Result） 利用者には検索結果として，文献名や文献の所蔵場所に関する情報を提示する．この情報には電子化された文献にリンクが張られており，利用者は提示内容に従って必要な文献を閲覧することができる．この検索結果には，検索条件に適合しない余分な情報として検索ノイズや，検索条件に適合するにもかかわらず提示されない情報として検索モレが含まれることが問題とされる．

文献検索システムの心臓部は検索エンジンであり，検索ファイルや検索式の設計に大きく関与する．次節では索引ファイルや検索式を含めた検索エンジンの理論とモデルについて説明する．

2.2.3 従来手法の検索理論とモデル

従来の文献検索の手法には，全文の文字列に対して検索式とパターンマッチングを高速に行う手法と，全文から必要な単語や語句を抽出し，これに対して検索式との照合を行う手法とに分類することができる．前者には，「文字成分表」や「N-gram」，「パトリシア木」などの高速文字列検索方式があり，後者には，「拡張ブーリアン」や「ベクトル空間モデル」，「確率モデル」，「LSI（Latent Semantic Indexing）」などがある．文献検索において質の良い検索結果を提供するためには，検索エンジンが検索式と索引語の適合度を算出し，結果に優先順位を設ける必要がある．優先度を求めるためには，前者のような文字列と検索式とのパターンマッチングではなく，後者の全文から必要な単語や語句を抽出しておく手法を用いなければならない．本論文の第 4 章で提案する意味検索手法は優先度を求める点で，この後者に該当する．以下，従来法の理論とモデルについて説明する．[43][45][46][47].

	あ	い	…	山	…
文献 1	1	1	…	0	…
文献 2	0	1	…	1	…
…					
文献 n	1	0	…	0	…

表 2.1 基本文字成分表

文字列	学位	位論	論文	文を	を書	書く	くの	のは	は大	大変
出現位置	001	002	003	004	005	006	007	008	009	010

表 2.2 N-gram 方式による文字列の切り出しとその出現位置

高速文字列検索

全文検索における文献の索引語は，出現するすべての文字列の組み合わせが対象となるため，索引語の数は非常に大きくなる．そこで，検索質問（検索式）に対する高速な索引語の問い合わせを行うために，効率のよい索引語の抽出とそれを保存する索引ファイルの構造に関する方法が用いられている．以下に主な方法について述べる [44][45].

文字成分表方式

文献中に現れるすべての文字をビット情報で表し，これを文字成分表として検索に用いる．例えば，文献中に存在する「山」という文字は，表 2.1に示すような文字成分表の「山」の欄に出現すれば「1」が立てられ，存在しなければ「0」のままとなる．この表のような 1 文字に関する成分表では，検索文字の出現順序関係の情報をもたないため，かなりの検索ノイズが生じる．そこで，2 文字ごとに成分表をつくる連接文字成分表や 1 文字飛ばしで 2 文字ずつの成分表を作るスキップ文字成分表などの工夫がなされている．しかし，出現頻度に関する情報をもたないため，候補となる文献の数が多くなると高速性が失われてしまう．大量文献を検索対象とした場合，絞り込みによる候補文献の削減方法が必要である．

索引番号	半無限部分文字列	ビット列
001	いろはにほへと	000101
002	ろはにほへと	001000
003	はにほへと	011001
004	にほへと	100010
005	ほへと	110010
…	…	…

索引番号	半無限部分文字
001	いろはにほへと
007	と
004	にほへと
003	はにほへと
006	へと
…	…

表 2.3 半無限部分文字列のリスト

N-gram 方式 N-gram 方式とは，文献中のすべての文字に対してある固定の長さ N の連続する文字列 (N-gram) を索引語として登録し，これを検索に用いる手法である．文頭から機械的に 1 文字ずつずらして長さ N の文字列を順に切り出す．切り出したすべての文字列を索引語として登録し，文献番号と文字列の出現位置を登録する．ただし，固定長 N は言語や文字の種類によって適切な値を選ぶものとする．例えば，N=2 の場合，ある文献に含まれる「学位論文を書くのは大変」という文では，表 2.2のように分割され，これらを索引見出して，文献番号と出現位置に関する情報を登録する．検索質問に対しても長さ N の連鎖する文字列に分割し，出現情報が連続する組合わせを検索結果とする．N-gram 方式は高速に索引登録，検索ができるが，文を形態素解析することなくすべての文字列に対して処理するため，容量が大きくなることが欠点である．

パトリシア木方式 パトリシア木方式とは，文中のすべての文字から始まる文字列（半無限部分文字列）を索引語として登録する．例えば，「いろはにほへと」という文は，表 2.3のように分割される．この半無限部分文字列を ASCII 文字コード順に並べ替えて登録する．具体的な処理としては，文字列をビット列として扱い，「0」か「1」で分岐する木構造の二分探索で文字列検索を行う．この方式は検索質問と完全一致した文字列を高速に検索することができるため，大規模な文献検索に向いている．しかし，日本語で書かれた

文献の場合，空白文字などの区切り文字が少ないため，索引語の数が非常に多くなり，それに応じて生成時間も長くなる．また，この応用として，文字列のビット列ではなく，文中の何文字目かを表す数字を索引語として用いたSuffix array方式がある[48].

パターン認識方式 パターン認識方式とは，文献中に含まれる文字列をバイトコード列として扱い，このパターンについて比較する方式である．この手法には日本語処理のための単語辞書，文法辞書等を一切必要とせず，言語に依存しないテキストパターンにより検索を行う．しかし，バイトコードの類似パターンには，多くのノイズを含むため，文献情報との照合が必要であり，文献数が多くなるに従い検索に時間がかかる．

拡張ブーリアン

検索質問をあらかじめ定式化した検索式には，ブーリアン方式が古くから用いられている．この方式は，論理式で利用者の要求を表現する．具体的な検索条件には，二つ以上の条件を同時に満たすＡＮＤ演算（論理積）や二つ以上の条件のうち少なくとも一つを満たすＯＲ演算（論理和），条件を満たすものを除くＮＯＴ演算（論理差）などを用いて表す．基本的には，利用者が入力した検索質問の文字列に完全に一致する文献を該当文献とするが，トランケーション（前方一致検索，後方一致検索，中間一致検索）法として，検索質問の一部分に不特定の文字が存在する場合でも一括して検索できる方法もある[43][45][46][47].

ブーリアン方式による検索式の表現法は，論理式を用いることで利用者がもつ表現の幅を広げる．しかし，検索質問に完全一致する検索結果のみが提示されるため，単語の選び方や検索結果の絞り込み技術が問われる[49]．そこで，ブーリアン検索の枠組を拡張し，適合度合によって検索結果に順位を付け，上位から順に利用者に提示する拡張ブーリアン検索が用いられるようになっている．以下にその拡張方法「Mixed Min and Max」と「P-normモデル」について述べる．

Mixed Min and Max 索引語 A, B と文献 D との関連度を表す重み d_A, d_B を求め，and と or の論理式を用いた検索質問の優先度を以下のように定義する．

$$q_{AandB} = min(d_A, d_B) \tag{2.1}$$

$$q_{AorB} = max(d_A, d_B) \tag{2.2}$$

ただし，q_{AandB} は検索質問の条件式 $AandB$ についての関連度とする．検索条件のうち，and は関連性の低い方を，or は関連性の高い方を選ぶという評価法である．

P-norm モデル 文献 D に対する索引語 A_i の重みを $d_{Ai}(\leq 1)$，検索質問に対する索引語 A_i の重みを q_i とする．このとき，or と and の質問に対する類似度 sim(質問, D) を以下のように定義する．

$$sim((A1orA2or..orAn), D) = \left(\frac{(d_{A1}q_1)^P + ... + (d_{An}q_n)^P}{q_1^P + ... + q_n^P}\right)^{\frac{1}{P}} \tag{2.3}$$

$$sim((A1andA2and..andAn), D) = 1 - \left(\frac{((1-d_{A1})q_1)^P + ... + ((1-d_{An})q_n)^P}{q_1^P + ... + q_n^P}\right)^{\frac{1}{P}} \tag{2.4}$$

ただし，P は 1 より大きいとするが，類似度計算に時間がかかるという問題がある．類似度 sim の値が大きい順に適合結果として提供する．

ベクトル空間モデル

単語をベクトル空間上の 1 点として扱った検索モデルにおいては，文献はいくつかの単語の集合によって特徴付けられると考えられる．同様に検索質問もいくつかの単語からなるので，ベクトル空間内の点として扱うことができる．各単語に対するベクトルの大きさを求めるために，文献 D_j における単語 t_i の重み w_j^i は以下のように定義される．

$$w_j^i = tf_j^i \cdot log\frac{N}{df_j} \tag{2.5}$$

ただし，tf_j^i を文献 D_j における単語 t_i の出現頻度，df_i を単語 t_i が出現する文献数とする．N は文献集合に含まれる文献の数であり，df_j を文献集合毎に正規化してい

る．単語t_iの重みw_j^iを用いることで，文献D_jのベクトルは$\vec{D}_j = (w_j^1, w_j^2, ...w_j^m)$と表される．ただし，$m$はベクトル空間の次元数とする．今，検索質問$Q$のベクトルを$\vec{Q} = (q_1, q_2, ..., q_m)$とすると，文献$D_j$と検索質問$Q$の類似度$sim(D_j, Q)$は以下の式より求まる．

$$sim(D_j, Q) = \frac{\vec{D}_j \cdot \vec{Q}}{|\vec{D}_j||\vec{Q}|} \tag{2.6}$$

$$= \frac{\Sigma_{i=1}^m (w_j^i \cdot q_i)}{(\Sigma_{i=1}^m (w_j^i)^2 \cdot \Sigma_{i=1}^m (q_i)^2)^{1/2}} \tag{2.7}$$

この式より，検索質問Qがベクトル空間モデルによる検索エンジンに与えられると，$sim(D_j, Q)$の大きい順に並んだに文献名のリストが得られる[50].

出現頻度分布による確率モデル

文献における単語の出現頻度分布をポアソン分布で近似するモデルがある．文献Dを短い間隔でn個に分割し，そこにある単語tが出現する確率をpとする．文献Dにおける出現回数の期待値λの単語がx回出現する確率は次式のポアソン分布$p(x;\lambda)$で近似できる．

$$p(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!} \tag{2.8}$$

$$\tag{2.9}$$

この分布によれば，文献集合において，単語tの出現する文献数の期待値dfは，$N(1-p(0;\lambda_t))$から予測することができる．idfとポアソン分布から予測される出現頻度の差をRIDFとして次式で定義される．

$$RIDF = idf - log_2(1 - p(0;\lambda_t)) \tag{2.10}$$

ただし，$idf = log_2\frac{N}{df}$，$\lambda_t = \frac{cf_t}{N}$である．$1-p(0;\lambda_t)$は単語$t$が少なくとも1回は現れる文献のポアソン分布における確率である．ポアソン分布は文献の意味内容に直接係わらない単語の出現頻度分布をよく近似するため，idfから差し引くことで，意味内容を表す単語の特徴を近似する．

LSI（Latent Semantic Indexing）

LSI とは，ベクトル空間の行列に特異値分解（Singular Value Decomposition）を適用し，ベクトル空間を正規直交空間として表すことで，多次元ベクトルの適当な重み付けから新たな変数を求め，この空間の統計量の次元を下げる方法である．この変数は，元の多次元ベクトル空間を集約し，少ない変数で元の空間を近似的に表すことができる．近似によるベクトルの損失については，最小二乗誤差用いて最適化する必要がある．文献間で共起する単語を機械的に同じ次元とすることができるため，近年，有効な検索手法として用いられているが，システムの性能は （元の単語 × 文献 のデータ量）に依存するため，SVD の計算量が大変重くなるという問題がある [51][52].

2.2.4 最近の動向とその問題点

従来の文献検索は，ベクトル空間モデルや確率モデル等の統計モデルに基づいた研究開発が主流である．統計モデルは文献を単なる文字列の羅列として扱い，この文字列の出現頻度情報から関連度を求める．文献中の単語が表す意味については求めないため，検索意図を解釈することはなく，また，検索意図に従った検索も難しい．そこで，近年においては，検索意図に着目した検索手法として，以下に示すような知識処理や学習アルゴリズム，文書クラスタ法といった知的検索の研究が盛んに行われている [53].

知識処理 検索対象の分野に関する知識情報を利用する手法である．階層構造を成すメニューに従って，利用者が求める情報が何かを絞り込んでいく手法 [54] や，複数のエキスパートシステムを用いて利用者の要求をパラメータで表現する手法 [55] が提案されている．

学習アルゴリズム 利用者と検索システムのインタラクションにより，検索質問を推敲していく手法である．利用者が検索結果を確認して検索質問を再入力することで，求める情報へと絞り込んでいく適合性フィードバック法がある．この手法に遺伝的アルゴリズムを組み合わせて，検索質問を再編成する手法などが提案されている [56][57][58].

文書クラスタ法 文書データベースの構成内容をあらかじめ利用者に提示し，検索対象の全体像からどのような検索質問を選ぶべきかを支援する手法である．教師なし学習方式に自己組織化マップを組み合わせた方法などが提案されている [59][60].

知識処理は特定した分野に限定した手法である．分野の拡張は困難であり，大規模な知識情報の獲得が必要となる．学習アルゴリズムを用いた手法は，修正回数が多くなるに従って利用者に煩わしさを与えることになる．また，利用者の不適切な絞り込みにより最終的には結果が得られない可能性もある．文書クラスタ法では精度の良い文書分類法が必要となるが，現在は文献中の頻出単語で特徴抽出および分類評価する手法が用いられており，統計モデルと同様な問題が残されている [61][62].

上記の他に知的検索を用いた製品として，ジャストシステム社の ConceptBase[63] やコマツソフト社の VextSearch[64] のような文献中に含まれる単語の関連性を数値化し，この関連度に基づいてベクトル空間を構築して検索に用いる手法がある．しかし，n 次元のベクトル空間を構築することは非常に大きなデータベースを抱えることになり，また，その構築方法が文書中の使用単語に大きく依存することは文書の質が問われることになる．

従来の知的検索法に共通した問題点は，効率のよい知識情報の獲得やその利用法，質のよい検索質問・索引語の抽出法などが未だ確立されていないことにある．この問題を解決するために，本研究では，知識情報を意味的な絞り込み，文書情報からその主題を抽出する手法を提案する．この手法を用いて，検索質問と索引語を抽出し，検索意図や文書主題に着目した意味検索法を実現する．知識情報については，既存のシソーラス辞書を使用するものとする．大規模な知識獲得の手法はオントロジーなどを用いて確立されつつあり，本手法では知識体系の利用方法に焦点をあてているためである．以下，本手法で提案する文書データからの主題抽出に関連する研究について述べる．

2.3 文書からの主題抽出に関する手法

文書データからの主題抽出法は，あらかじめ決められた抽出テンプレート（抜き出す項目）に対する情報を文書から抽出する手法と，抽出テンプレートを使わずに主題を抽出する手法とに分類することができる．前者は，文書理解を行わず，必要な情報だけを効率的に取出すことに重点を置いており，電子ニュースのダイジェスト作成 [65] や IICA[5]，METIS[66] などの手法が提案されている．これらの手法は情報の構造が決まっている場合において，効率よく，正確にその内容を抽出することができるが，必要とする情報の分野毎に抽出テンプレートを作成する必要がある．そこで，後者のような文書の分野に関係なく主題を抽出する手法として文書の重要個所を抽出する手法が提案されている．本手法は，ディジタル図書館で保管されている様々な分野の文書を対象としているため，後者の抽出テンプレートを作成しない主題抽出法を提案している．以下，両者の手法について述べる．

2.3.1 抽出テンプレートに基づいた情報抽出

文書に対して，あらかじめ抽出テンプレートを作成する手法は，特定の情報を簡単に調べたい場合に非常に有効である．抽出する内容が決まっており，抽出テンプレートを事前に作成することが出来れば，広範囲で高速に必要な情報を得ることができる．その用途は，電子ニュースや報告書，新製品の情報，医療カルテと様々である．以下に抽出テンプレートを用いた具体的な手法について述べる [4][67].

電子ニュースダイジェスト自動生成 電子ニュースの会議情報について抽出テンプレートを作成し，ダイジェストを自動生成する．この手法は，電子ニュースのうち，会議告知記事 fj.meetings からから「タイトル」,「開催期日」,「開催場所」等を抽出するシステムであり，その抽出方法は，文書のスタイル（センタリング，境界線，箇条書きなど）と言語パターン（「ご案内」「開催」「x 年 y 月 z 日」などの文字列パターン）およびそれらの順序情報を利用している [65].

IICA IICA は WWW からオントロジーと呼ばれる基本語彙の体系を用いて情報抽出を行うシステムである．オントロジー上の各概念に対して，属性情報（たとえば,「温泉」についての属性情報は “温泉名”, “効能”, “泉質” など）を定義し，各属性情報に特有の言語表現パターンから抽出ルール（たとえば, “@痛” や “@症” の表現パターンが存在すれば，その内容は「傷病」を指す）を決定し，このパターンに従って文書内容を抽出する手法である [5].

METIS METIS は HTML 化した金属材料論文の要約とサーベイを自動的に行うシステムである．金属材料系の論文を対象とし,「試験材料」,「実験方法」,「実験結果」などの抽出テンプレートから内容抽出を行い，ヒューリスティックな自然言語処理で得られる知識片 KP(Knowledge Piece) を用いて技術情報空間を構築している [66].

これらの手法は，基本的に抽出テンプレートと文書データとのパターンマッチング技術により情報を抽出しているが，共通した問題として，テンプレートを情報抽出の課題ごとに作成しなければいけないことが挙げられる．つまり，電子ニュースのテンプレートは電子ニュースに対してしか活用することが出来ず，他の文書に対して内容抽出を行うときには新たに作成しなければならない．そこで抽出テンプレートを使用しない様々な手法が提案されており，次節でこれらの手法を紹介する．

2.3.2 文書における重要個所の抽出法

重要個所の抽出法は，文書の自動要約の必要性からも様々な研究がなされている．これらの研究は，基本的に文書の個所（単語，文，段落）に対して重要度を与え，その順序から重要個所を抽出する．以下に重要個所を抽出するために必要な重要度を求める手法について示し，その問題について述べる．[68].

文書中の単語の出現頻度に基づく手法 文書中における頻出単語は文書の主題を示すという仮定が情報処理の分野でよく用いられる．出現頻度の高い名詞をキーワードと考える tf 法や，tf 法に加えて文書数も考慮する tf*idf 法などを用いた単語の重み付け技法が代表的である [69]．文書中の出現頻度に基いて単語に重要度を与えるという考え方を利用し，単語の重要度を基に文に

重要度を付与するという重要文抽出手法も提案されている [70][71]. 単語の重要度から文の重要度を計算する手法は文中に出現する単語の重要度の総和を文の重要度としている.

文書中での位置情報に基づく手法 文書は, その構造に規則性があると仮定することができる. たとえば, 学術論文は, 序論, 本論, 結論のような構造を持ち, 新聞は, 見出し, 小見出し, 本文の構造をもつことが多い. この文書構造における位置から重要度を求め, 重要な箇所を決定する手法が提案されている. 文書全体のまとめは書き出しや結び近くにあり, 重要な文は段落の先頭, 最後, 節の見出しの直後にあるとする手法 [72] や, 新聞記事の要約は本文の始めにあるとする lead 手法 [73] などが提案されている. この他に, 本文以外の情報 (文書の章や節のタイトルや, 新聞の見出し等) から重要個所を抽出する手法も提案され, 最近では, 見出しに含まれる名詞を多く含む文が重要だとする手法がある [74][75].

手がかり表現に基づく手法 文書中には, 重要箇所を探す手がかりとなる表現がいくつか存在する. たとえば, 学術論文では, 'this method', や 'our approach' などの表現が論文の主題を表す個所に出現し, 'for example' などの例示を示す接続語で始まる文は重要度が低いと言える. このような手がかり表現を利用して, 文書中の重要文を抽出する手法が提案されている. [72].

文書中の単語／文の類似性に基づく手法 文書に出現する単語の重みのベクトルを用いて, 文書間の類似度をベクトル間の内積等で計算する手法や, 文書中の文を単位として, 同様な類似度から文間のつながりの度合を計算し, 重要と考えられる文を抽出する手法などが提案されている [76]. これらの手法は, 共通した単語が文書に出現する度合に基づいて文間の結合度を求め, 重要個所を抽出する. 最近では, 段落をノードとして類似度の高い段落同士をリンクで結んだマップを生成し, その中から重要段落を抽出する手法 [77] や 2 文間に出現する共通な単語に対して計算した文間関連度の平均と, あるカバレジに基づいて文の重要度を計算する手法などがある [74].

文書における重要個所の抽出は，文書の主題を把握する上で大きな役割を担う．しかし，上記に挙げた手法を用いれば文書の内容を全て理解できるわけではない．重要個所の抽出における問題には，文書中の抽出箇所を単に集めてもその内容につながりがないことが挙げられる．この問題は，抽出した個所を文字列の羅列として処理しているため，著者が抱いた意味を充分に表現できないために生じる．特に，本研究の目的は文書の要約ではなく，様々な情報活用の場で汎用性のある主題の抽出を実現することであるため，抽出結果における不明瞭な内容を利用者によって類推してもらうわけにはいかない．この問題点を解消するために本手法は，単語の出現頻度だけでなく単語の概念情報についての出現頻度を用い，文字列ではなく概念を最小単位として文書の主題を解析する．本手法は出現頻度を用いるため上記で紹介した“出現頻度に基づく手法”と類似しているが，出現頻度を求める対象が単語ではなく，単語の概念情報を用いている点が異なり，意味を考慮した処理である点が新しい．具体的には，概念の出現頻度分布から主題が表す内容の“表現度”として単語の概念に関する重みを求め，この表現度に基づいて文書の内容として重要な概念だけに絞り込む．また，主題の内容をどれぐらい強調している文書であるかを表すために，主題の“主張度”として文字列の出現頻度分布から単語の出現に関する重みを求め，両者を用いた特徴ベクトルから文書の内容を定量的に評価して文書の主題を抽出する．

さらに，本論文では，提案手法の適用として先述した意味検索の他に，文書構造に基づいた内容の抽出について述べる．これは，上記した“位置情報に基づく手法”と類似しているが，概念情報を用いることで構造部分が表現する内容の意味を提示する点に新規性がある．

第3章
文書主題の抽出法

本章では，提案する文書主題の抽出法について述べ，その評価実験と考察について報告する．

3.1 文書特徴量の抽出

文書の主題を単語体系に基づいた特徴抽出から求めたのでは内容的特質を充分に表現することは困難である．本手法では，単語情報に加えて概念情報を用いることで深層的な観点から特徴抽出を行ない，内容表現の充実を図る．

文書の特徴は，以下に示す2種類の出現頻度分布に基づいた特徴ベクトルを用いて定量的に表現する．

- 出現単語の文字列に関する出現頻度分布
- 出現単語の概念に関する出現頻度分布

文字列の出現頻度分布からは文書がもつ主題の“主張度”として単語の出現に関する重みを求め，概念の出現頻度分布からは主題が表す内容の“表現度”として単語の概念に関する重みを求める．

技術論文などの比較的狭い分野を対象とする場合，単語の出現頻度を文書集合から相対的に評価して抽出語を決めたのでは，その分野で多く見られる重要な単語を取り除いてしまう．本手法では，論文毎に表現度を求めて主題を明らかにし，その主題の主張度から主題を述べる強さを求める．この両者の重みに基づいて文書の主題を表す特徴ベクトルを決定する．技術論文の主題を獲得する処理の流れを図3.1に示す．

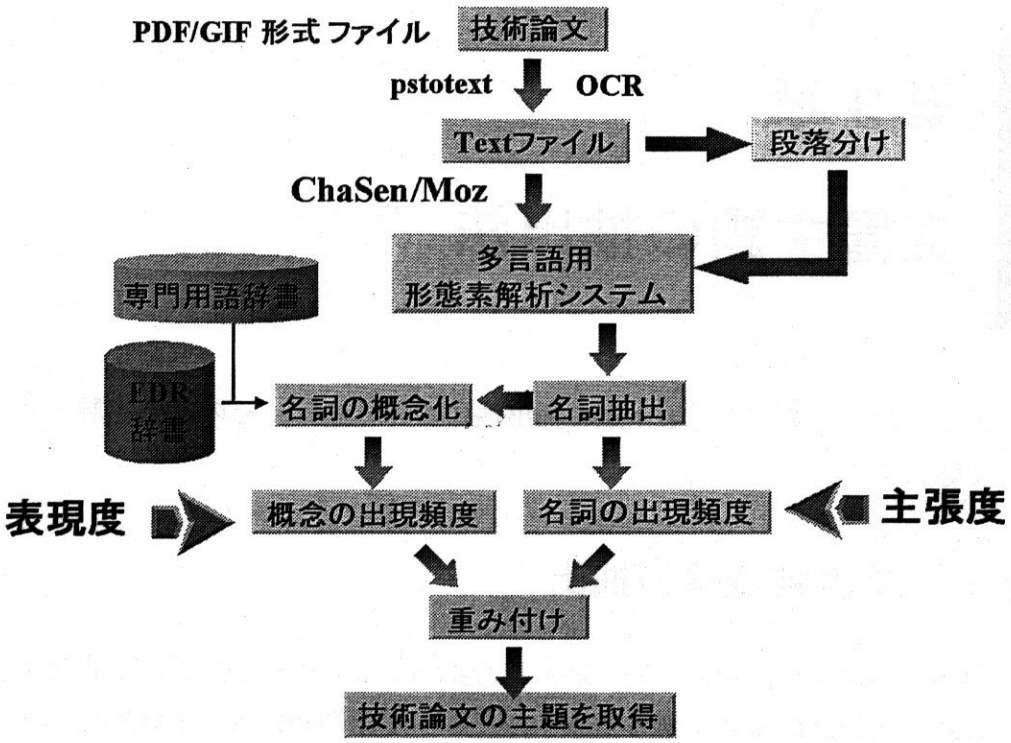


図 3.1 技術論文の主題抽出処理

3.1.1 単語の文字列に関する出現頻度分布

抽出した文書の主題を情報検索に用いる場合，その主題は利用者が要求する内容に沿っているだけではなく，その内容を述べる強さを求める必要がある．本手法では，文書毎に主題を述べる強さとして主張度を求め，文書間でその強さを比較する．

今，文書 $d_j\{j = 1, 2, 3, ...n\}$ から重複語を取り除いた単語 $w_i\{i = 1, 2, 3, ..., n\}$ を入力とし，対象とする全ての文書 d_j の集合を次式のように文書データベース DB と表すものとする．

$$DB = \{d_1, d_2, d_3, ..., d_j, ...\} \tag{3.1}$$

このとき，文書 d_j 内にある単語 w_i に関する主張度 $I_w(w_i, d_j)$ を次式のように定

義する．

$$I_w(w_i, d_j) = \frac{f_w(w_i, d_j)}{max_{f_w(w_l, d_j)}} \tag{3.2}$$

ただし，$f_w(w_i, d_j)$ は 文書 d_j 内にある単語 w_i の出現頻度であり，$max_{f_w(w_l, d_j)}$ は文書 d_j 内で最も多く出現した単語 w_l の出現頻度とする．

3.1.2 単語の概念に関する出現頻度分布

文書は出現単語の集合であり，単語の集合が表す意味によって文書内容が表現される．そのため，文書の特徴を抽出するには文書の出現単語がどのような意味で使用されているのかを求める必要がある．本手法では，出現単語の概念情報を用いて意味のルーツ（意味素性）を明らかにし，深層的な観点から単語が表す内容を求める．

以下ではまず，主題の特徴を表す概念情報を獲得するために必要な概念空間の構築法について述べる．

EDR 電子化辞書を用いた概念空間の構築

本研究では，単語の概念情報を用いることで，文書の出現単語がもつ深層的な意味を求める．単語を概念で表現するにはシソーラスを用いることが一般的である．近年では国語辞書からシソーラスを自動構築する手法 [37] やオントロジーによる手法 [5] などが用いられ，比較的精度のよい情報の構造化が試みられている．本手法では，単語を概念で表現することに加えて単語から生成される概念空間を構築し，この空間を上位層へたどることで単語がもつ意味のルーツを探索する．単語から概念空間を構築するには概念の上位-下位関係について詳細に記載しているシソーラスが必要であり，本研究ではこの条件を満たす既存のシソーラスとして EDR 電子化辞書を採用する．EDR 電子化辞書は 11 のサブ辞書から成るが，ここでは表 3.1に示すような記述形式による単語辞書および概念体系辞書，概念見出し辞書の 3 種類の辞書を用いる．各辞書中における概念の記述は 16 進数の概念識別子で表現されているため，概念間の比較が容易であり，言語の種別に関係なく検索できるといった特徴をもつ．

具体的には，概念体系辞書を用いて単語の上位概念を再帰的に検索することで，

表 3.1 EDR 電子化辞書の記述形式

日本語単語辞書

[形式] レコード番号 \t 単語見出し \t 不変化部-連接属性対 \t かな表記 \t 発音 \t 品詞 \t 構文木 \t 活用情報 \t 表層格情報 \t 相情報 \t 機能語情報 \t 概念識別子 \t 英語概念見出し \t 日本語概念見出し \t 英語概念説明 \t 日本語概念説明 \t 用法 \t 頻度 \t 管理情報 \n “”

[実例] JWD0372940 \t IR[アイアール] \t IR(JLN1,JRN1) \t アイアール \t アイア’ール \t JN1 \t “” \t “” \t “” \t “” \t “” \t **3bcd88** \t “information retrieval” \t 情報検索 [ジョウホウケンサク] \t “the act of retrieving necessary information (from files)” \t 必要な情報を検索すること \t AB \t 0/0 \t DATE=“93/3/10”

概念体系辞書

[形式] レコード番号 \t 上位概念識別子 \t 下位概念識別子 \t 管理情報 \n

[実例] CPC0380737 \t **3c7ddf** \t **3bcd88** \t DATE=“95/6/7”

概念見出し辞書

[形式] レコード番号 \t 概念識別子 \t 英語概念見出し \t 日本語概念見出し \t 英語概念説明 \t 日本語概念説明 \t 管理情報 \n

[実例] CPH0402128 \t **3c7ddf** \t **output** \t 出力する [シュツリョク・スル] \t **“an act of taking out data from a computer”** \t コンピューターからデータをとりだすこと \t DATE=“95/6/6”

注）\t：タブコード，\n：改行コード

階層構造をもった概念の集合を概念空間として得る．図 3.2は，概念空間の構造を示しており，五つの基本概念と一つの専門用語を次元とした空間で表現される．概念体系辞書では,「概念」を最上位概念とし，その下位概念を基本概念（「人間または人間と似た振る舞いをする主体」,「ものごと」,「事象」,「位置」,「時」の 5 つの概念）として，概念の上位-下位関係を階層構造として体系化している．図 3.3に基本概念の下位概念を記す．各単語はこの階層構造の節として表現され，各節の上位概念の集合は一意に決まる．また，概念には多重継承が存在し，一つの概念が二つ以上の上位概念をもつ場合があるため，正確には木構造をとらないが，

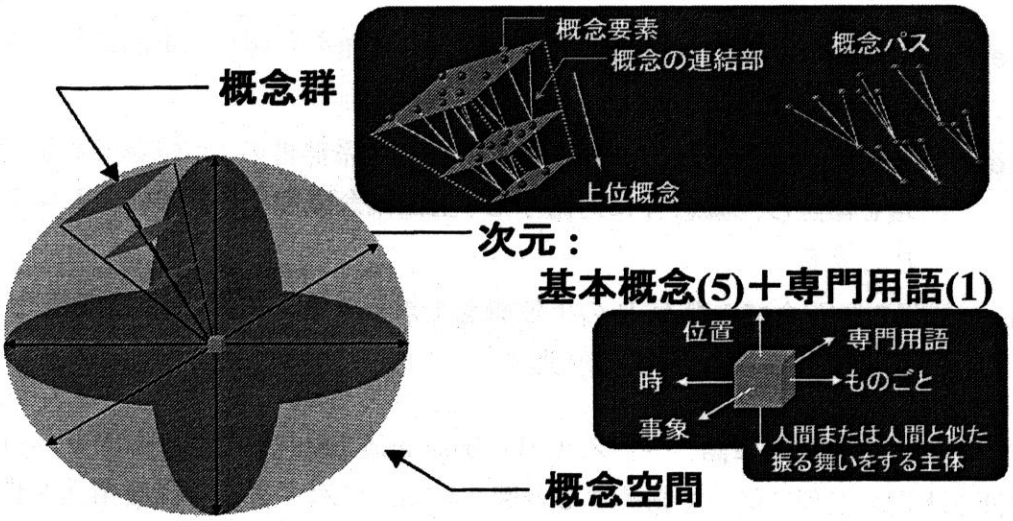


図 3.2 概念空間の構造

[基本語概念体系の上位2レベル目までの項目一覧]

1 人間または人間と似た振る舞いをする主体 3aa911
- 1-1 人間 30f6b0
- 1-2 動物 30f6bf
- 1-3 自立活動体 3aa912
- 1-4 神，仏，霊，天使など，抽象的な有意志体 4444b6

2 ものごと 3d017c
- 2-1 もの 444d8
- 2-2 事柄 444ab5
- 2-3 識別名 444daa
- 2-4 客観的な対象 0e7faa

3 事象 30f7e4
- 3-1 現象 30f7e5
- 3-2 行為 30f83e
- 3-3 移動 30f801
- 3-4 変化 3f9856
- 3-5 状態 3aa963

4 位置 30f751
- 4-1 場所 3aa938
- 4-2 複数のものの関係によって決まる位置 30f753
- 4-3 領域 30f767
- 4-4 部分 3f9651
- 4-5 方向 3f9658
- 4-6 抽象的位置 444a9d

5 時 30f776
- 5-1 時間点 3f9882
- 5-2 周期的に訪れる時間 444dd2
- 5-3 過去から現在，未来までの時間の流れの中で捉えた時間 444dd3
- 5-4 時間 30f77b
- 5-5 単位で長さを示した時間 444dd4
- 5-6 周期 4449e2
- 5-7 経過・歴史 30f7d6

図 3.3 上位レベル 2 までの基本概念

便宜上，木構造の場合の用語を用いるものとする．

以下に単語から生成される概念空間の構築手順を示す．

step 1 日本語単語辞書の“単語見出し”の項で単語を参照し，単語に対応する“概念識別子”として概念Aを求める．

step 2 概念Aの上位概念を求めるために概念体系辞書の“下位概念識別子”の項を参照し，概念Aに対応する“上位概念識別子”を求め，これを概念Bとする．

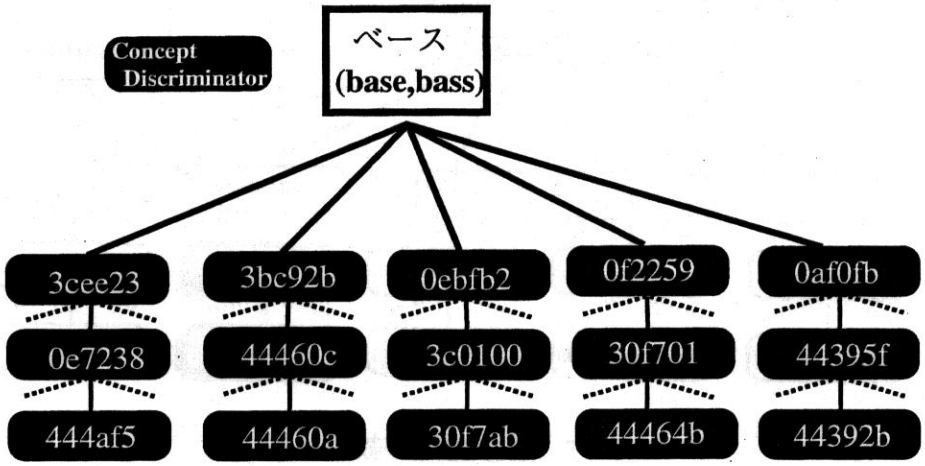
step 3 得られた全ての概念Bの上位概念を求めるために，概念Bを概念Aと置き換えて**step 2**を繰り返す．

例えば，日本語の単語“ベース（ the base and bass ）”から生成される概念空間を上記の処理に従って構築すると図3.4（a）に示すような階層構造が得られる．ただし，理解を容易にするために概念空間の主要な枝だけを示している．また，図3.4（a）の概念識別子は概念見出し辞書から概念の見出し語とその説明を検索し，概念識別子が示す概念の内容を知ることができる．図3.4（b）は図3.4（a）の各概念識別子を英語概念見出しとその説明で表したものである．

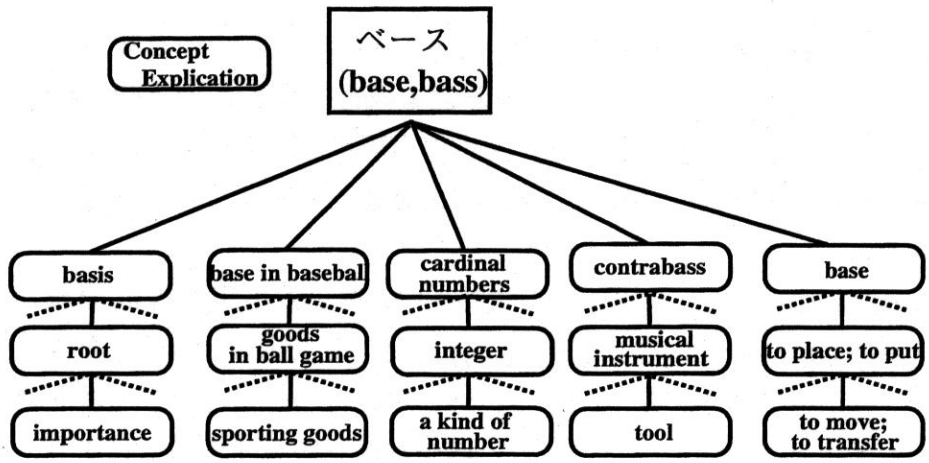
本研究では，概念空間の各節となる概念を根とする階層構造を部分概念空間とおく．今，任意の単語 w を根として上位概念の検索を k 回繰り返して得られた概念のうち t 番目の概念を $c(w,k,t)$ とすると，単語 w から生成される概念空間における部分概念空間の集合 $A(w)$ は，概念 $c(w,k,t)$ から生成された部分概念空間 P_{wkt} を用いて次式のように表される．

$$A(w) = \{P_{w01},\ P_{w11},\ P_{w12},\ \ldots P_{wkt},\ \ldots\} \tag{3}$$

ただし，$k=0$ の部分概念空間 P_{w01} は，単語を根とする概念空間であり，この空間を部分概念空間の集合 $A(w)$ に含めるものとする．



(a)



(b)

図 3.4 単語 “ベース” から生成される概念空間（(a) 概念識別子による表示 (b) 英語概念説明による表示）

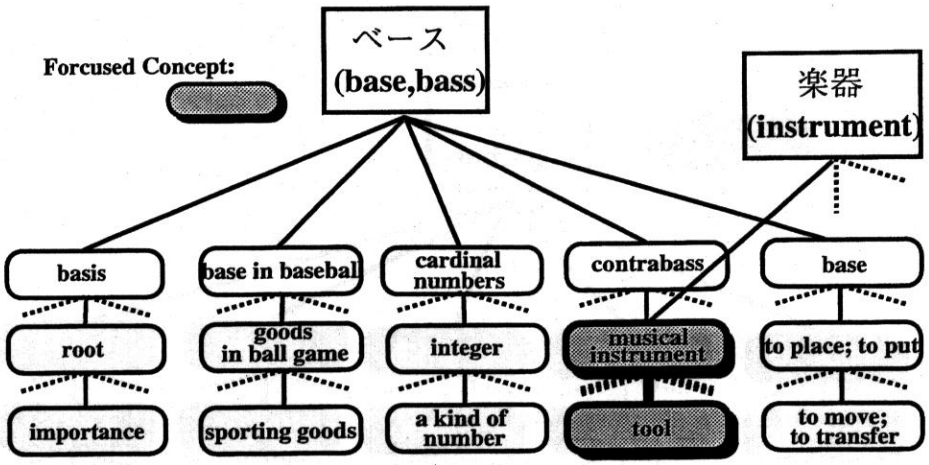


図 3.5 概念空間の絞り込み

3.1.3 概念空間の照合によって絞り込まれた部分概念空間とその出現頻度

1 語から生成された概念空間には，文書の主題とは異なる意味も含むといった問題がある．上述の “ベース” という単語も概念空間のどの意味で使用されているのか断定できない．本手法では，文書の出現単語から生成された概念空間を照合することで，文書内容に沿って概念空間を絞り込み，その空間についての出現頻度を求める．

概念空間の照合とは，文書から重複語を取り除いた出現単語 w_i を根とした概念空間における部分概念空間の集合 $A(w_i)$ に対して共通要素 $P_{w_i kt}$ を求めることである．文書中の単語は各々がいくつかの意味を持つが，その多くは共通して文書の主題を表現する意味を含んでいる．概念空間を照合することにより，単語が共通して表現する意味に概念空間を絞り込み，これを部分概念空間として獲得する．

例えば，文書中に “ベース” という単語の他に “楽器” という単語が出現したとする．図 3.4 (b) の概念空間に “楽器 (instrument) ” から生成される概念空間を照合すると，図 3.5に示すような同じ意味から成る部分概念空間が得られる．この結果から “ベース” は “道具 (tool) としての楽器 (musical instrument)” を意味していると推測することができる．

また，概念空間の照合から得られた部分概念空間を要素としてもつ集合 $A(w_i)$ の数をその部分概念空間の出現頻度とする．出現頻度の高い部分概念空間が，文書の主題を強力に表現していることになる．以下に，概念空間の照合により絞り込まれた部分概念空間とその出現頻度を獲得するまでの手順をまとめる．

step 1 文書から重複語を取り除いた単語 w_i について，概念空間を構築する．

step 2 概念空間について部分概念空間の集合 $A(w_i)$ を求める．

step 3 各部分概念空間の集合 $A(w_i)$ において，共通要素となる部分概念空間 P_{w_ikt} を抽出する．この P_{w_ikt} が照合により絞り込まれた部分概念空間となる．

step 4 P_{w_ikt} を要素とする部分概念空間の集合 $A(w_i)$ の数を出現頻度として求める．ただし，出現頻度が 1 回限りの部分概念空間は明らかにノイズであるため，あらかじめ取り除いておくものとする．

3.1.4 部分概念空間における連結数と概念パス

本手法は単語の上位概念を求め，その経路（パス）を探索することで，意味のルーツを探る．しかし，その内容が全く具体性に欠けてしまっては文書の主旨を表現することはできない．そのため，概念空間を絞り込んだ結果，得られた部分概念空間は，シソーラスの階層構造において下位層部に存在する必要がある．つまり，単語が具体的な事象を示し，かつ，部分概念空間の根のレベルが低いことが必要である．ここで，部分概念空間の根のレベルとは，概念空間内で単語の値を “0” としたときのレベル数のことである．部分概念空間のレベルは上位概念を検索する際に求めることが可能であるが，大規模なシソーラスの階層構造から，単語の具体性を明らかにすることは非常に困難である．そのため，両者を満たす指針が必要となる．

本手法では，部分概念空間の内容が抽象的な表現に偏らないために，部分概念空間の根から外部節（external node：子をもたない節）までの相対レベルとして連結数を設定し，根から外部節までの経路を概念パスとして定義する．

今，図3.6において，任意の単語 w を根として上位概念の検索を k 回繰り返して得られた概念のうち t 番目の概念を $c(w,k,t)$ とする．この $c(w,k,t)$ を根とする部

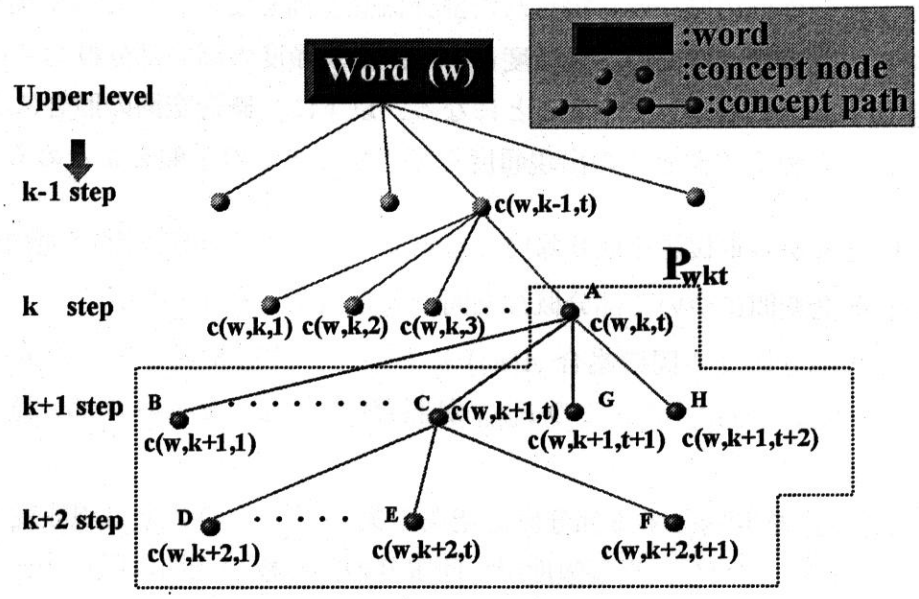


図 3.6 部分概念空間と概念パス

分概念空間 P_{wkt}（点線で囲んだ部分）に対する概念パスは，概念 $c(w,k,t)$ を先頭として，各外部節までの上位概念をレベル順に並べたリストであり，外部節の数だけ存在する．$c(w,k,t)$ と g 番目の外部節までの概念パスを $h(c(w,k,t),g)\{g = 1, 2, 3, ...\}$ とおくと，概念 $c(w,k,t)$ を根とする部分概念空間 P_{wkt} は，次式のように概念パス $h(c(w,k,t),g)$ の集合として表現される．

$$\begin{aligned} P_{wkt} \;=\; & \{h(c(w,k,t),1),\ h(c(w,k,t),2), \\ & ...,\ h(c(w,k,t),g),\ ...\} \end{aligned} \tag{3.3}$$

連結数の多い概念パスは，そのリストに具体的な内容を表す概念を要素として含んでおり，その先頭リストは概念空間においてレベルが低いことを表す．一方，連結数の少ない概念パスは，先頭リストのレベルが高いか，または単語がシソーラスの階層構造の中で上位層に位置し，その内容が抽象的であることを示している．本手法では部分概念空間を概念パスで表し，連結数が少ない概念パスについてはその内容が抽象的な表現に偏っているため，本手法での有効性は低いとする．

シソーラスがもつ階層の高さは，そのシソーラスの規模によって異なる．大規

模なシソーラスにおいて，階層構造の上位層では抽象的な内容の微妙なニュアンスについて記述してあり，本手法での有効性は低く，データ量を増やす原因にもなる．そのため，シソーラス全域にわたって概念空間を構築する必要はなく，シソーラスの規模や特性等を考慮して，適宜，階層の高さに上限を決定する必要がある．このとき，上限が少なすぎると連結数に違いがなくなるため具体性を連結数から推測することが難しくなるが，適度な上限はデータ量を減らし，抽象的な内容をあらかじめ省いておくことができる．本手法で使用したEDR電子化辞書では経験則から概念空間の階層の高さを“レベル5”とした．そのため，外部節とは概念空間においてレベル5までの末端の節を示すものとする．

3.1.5 文書内容の表現度

概念空間の照合によって求まる部分概念空間の出現頻度を用いて，その空間が表す文書内容の表現度を算出する．今，文書 d_j の重複語を取り除いた単語 w_i から上位概念の検索を再帰的に k 回繰り返して得られる概念集合の t 番目の概念 $c(w_i,k,t)$ から構築された部分概念空間 P_{w_ikt} に関する表現度 $E_c(P_{w_ikt},d_j)$ を次式のように定義する．

$$E_c(P_{w_ikt},d_j)=\frac{f_c(P_{w_ikt},d_j)}{max_{f_c(P_{w_ims},d_j)}} \tag{3.4}$$

ただし，$f_c(P_{w_ikt},d_j)$ は，文書 d_j の重複語を取り除いた単語 w_i から構築された部分概念空間 P_{w_ikt} の出現頻度であり，$max_{f_c(P_{w_ims},d_j)}$ は，文書 d_j 内で最も多く出現した部分概念空間 P_{w_ims} の出現頻度とする．このとき，上位概念集合は一意に決定されるため，部分概念空間についての表現度はそのまま部分概念空間がもつ概念パスの表現度となる．

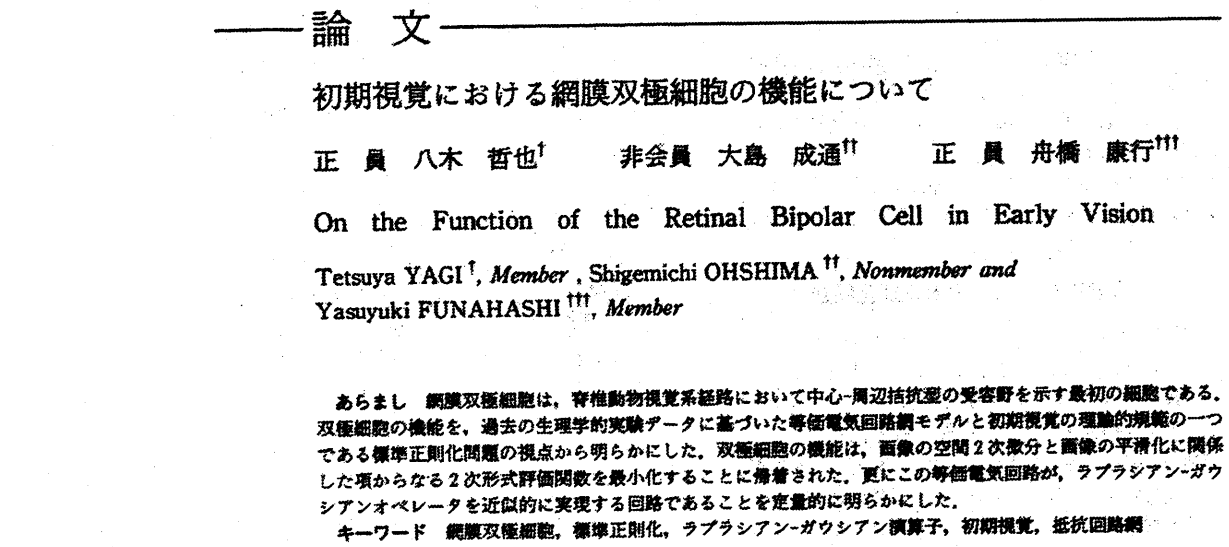
3.1.6 特徴量

情報検索などの文書処理に適用する場合，文書の主題について内容を抽出するだけでなく，主張の強さについても考慮する必要がある．本手法では，単語の文字列に関する出現頻度から主題の主張度を求め，単語の概念に関する出現頻度から主題の表現度を求める．この両者を用いて文書の主題を表す特徴ベクトルの特徴量を決定する．

特徴量は概念パスを用いて表現する．表現度の高い概念パスが文書の主題を表し，その概念パスが主張度の高い出現単語から求められていれば，文書はその主題を強調していることになる．表現度および主張度を用いて文書の特徴量を決定するための指針を求める．さらに，抽象的な内容を表す概念パスを文書の特徴としないために，この指針の結果に対して連結数をキーとして降順にソートする．今，概念パスの連結数が r である概念パス $h_p(c(w_i,k,t),g,r)$ の表現度を $E_c(h_p(c(w_i,k,t),g,r),d_j)$，概念パス $h_p(c(w_i,k,t),g,r)$ を概念空間に含む文書 d_j の出現単語 w_q の主張度を $I_w(w_q,d_j)$ とすると，文書の特徴量を決定するための指針 $G(h_p(c(w_i,k,t),g,r),d_j)$ は，次式のように定義される．

$$G(h_p(c(w_i,k,t),g,r),d_j) = \sum_q I_w(w_q,d_j) + E_c(h_p(c(w_i,k,t),g,r),d_j) \tag{3.5}$$

連結数 r および $G(h_p(c(w_i,k,t),g,r),d_j)$ の値が大きい概念パス $h_p(c(w_i,k,t),g,r)$ が文書の特徴を強力に表現していることになる．この特徴量を索引語として情報検索システムに適用し，概念での検索を実現した検索システムのモデルについては 4 章で述べる．



論　文

初期視覚における網膜双極細胞の機能について

正員　八木　哲也[†]　　非会員　大島　成通[††]　　正員　舟橋　康行[†††]

On the Function of the Retinal Bipolar Cell in Early Vision

Tetsuya YAGI[†], *Member*, Shigemichi OHSHIMA[††], *Nonmember and*
Yasuyuki FUNAHASHI[†††], *Member*

あらまし　網膜双極細胞は，脊椎動物視覚系経路において中心-周辺拮抗型の受容野を示す最初の細胞である．双極細胞の機能を，過去の生理学的実験データに基づいた等価電気回路網モデルと初期視覚の理論的規範の一つである標準正則化問題の視点から明らかにした．双極細胞の機能は，画像の空間２次微分と画像の平滑化に関係した項からなる２次形式評価関数を最小化することに帰着された．更にこの等価電気回路が，ラプラシアン-ガウシアンオペレータを近似的に実現する回路であることを定量的に明らかにした．

キーワード　網膜双極細胞，標準正則化，ラプラシアン-ガウシアン演算子，初期視覚，抵抗回路網

図 3.7 主題抽出実験に用いた技術論文“初期視覚における網膜双曲細胞”

3.2 評価実験

3.2.1 技術論文の主題を概念を用いて抽出する実験

本手法においては文書の主題を概念で表現することが焦点となる．そこで，文書をディジタル図書館で保管している技術論文とし，このテキストデータから論文の主題を概念パスの集合として獲得する実験を行なった．上記の実験において設定した条件は次の通りである．

- テキストデータには，本学付属図書館が保管している技術論文の OCR 結果を使用する．
- 概念空間における階層の高さをレベル５とする．
- 概念空間の絞り込みに閾値は設定せず，一度しか出てこない概念パスのみを切り捨てる．

表 3.2に実験結果を示す．これは，電子情報通信学会論文誌に掲載された図 3.7 の主題を概念で抽出した結果である．この表は，概念パスの特徴量を決定するた

めの指針 $G(h_p(c(w_i,k,t),g,r),d_j)$（図 3.2では G と略す.）を値の大きな順番に並べ，G の値が同じものについては連結数の多い順に並べた結果から上位の概念パスを抜粋したものである．ここでは便宜上，3連結までの概念パスを紹介する．表において，三つずつの区切りはそれぞれ概念パスを示しており，下段へいくに従って上位レベルの概念を表す．また，各概念ごとに EDR 電子化辞書での概念識別子とその見出し語を日本語と英語で示す．

表3.2の結果より，$G(h_p(c(w_i,k,t),g,r),d_j)$ の値が最も大きい“網膜”は，概念パスの意味から生物学的な名称ではなく，視覚器官として使用されていることがわかる．また，“双曲細胞”においても概念パスが情報分野を示していることから，著者が生物学用語を情報分野の視点で使用していることがわかる．この考察は，通信や回路といった他の概念パスが情報分野を指していることからも確認され，この論文の主題は情報分野における視覚および通信，回路といった概念で表現されることがわかる．このように，従来の文字列検索では単語の意味まで絞り込むことは困難であったが，概念パスから単語がもつ意味のルーツを探索することで，文書の著者が用いる単語の意味を推測することが可能となる．

3.2.2 精度評価とアンケート調査

本手法の有効性を検証するために，電子情報通信学会論文誌 VOL.J78-DII NO.7 に掲載されている 15 本の論文を対象に上記の実験を行い，その結果をもとに精度を求めた．また，さらに利用者の視点から本手法を評価するためにアンケート調査を行った．精度については，実験結果とあらかじめ人手で主題として抽出した単語の結果との整合性から評価した．

精度評価とアンケート調査

精度の評価には，再現率（R：Recall）と適合率（P：Precision）の指標を用いた．各指標を次のように定義する．ここで，本手法が主題として抽出した概念を「実験結果」とし，人手によって抽出された単語を「正しい結果」とする．また，単語と概念との整合性をとるために，概念には見出し語を参照した．

表 3.2 概念パスによる技術論文“初期視覚における網膜双曲細胞”の主題表現

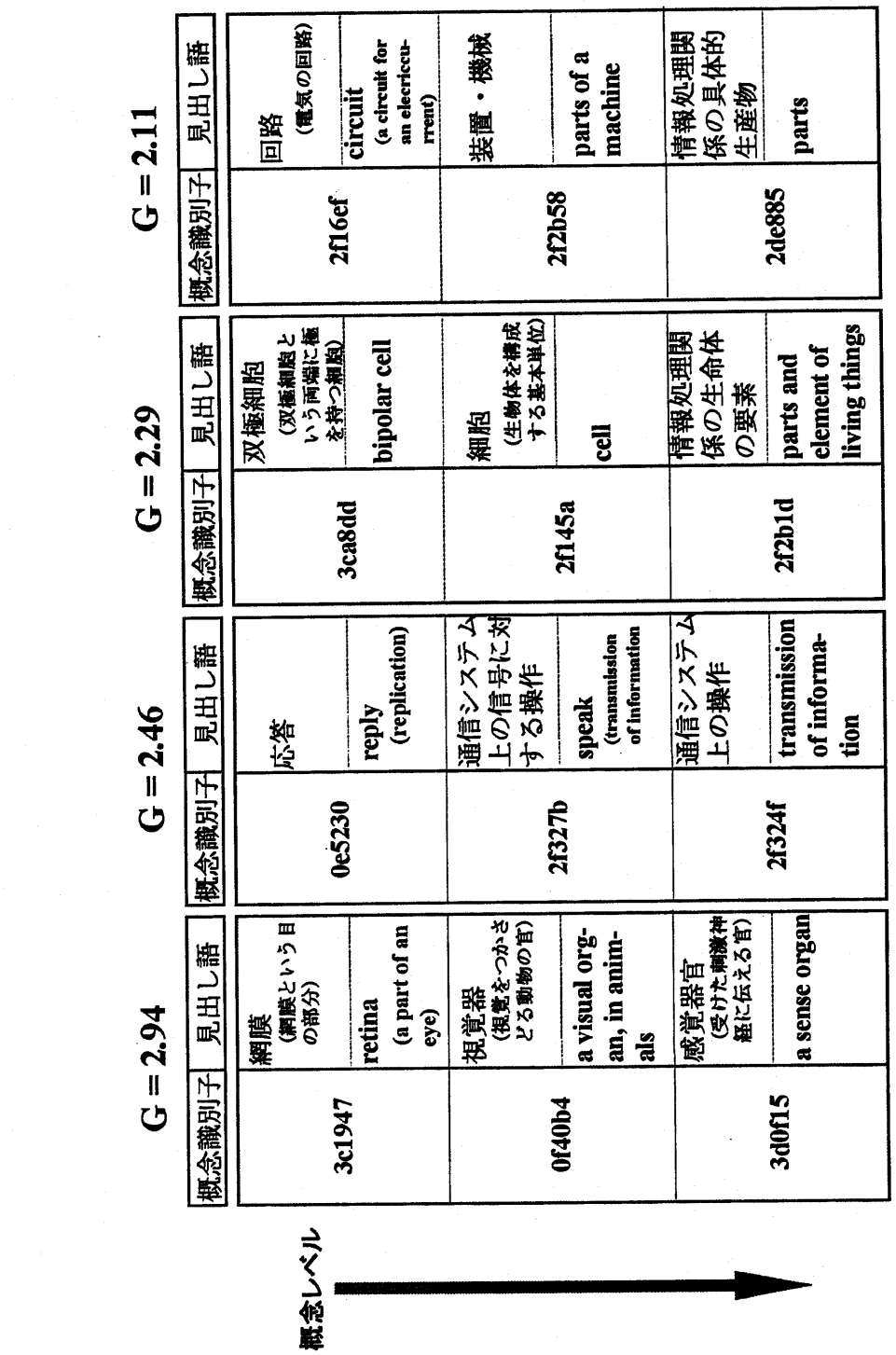


表 3.3 論文誌に対する精度評価

	Recall	Precision
IEICE VOL.78-DII NO.7 (15 papers)	3728/5136 =72.6 %	3728/7456 =50.0%

$$R = \frac{\text{実験結果に含まれる正しい結果の数}}{\text{正しい結果の数}} \tag{3.6}$$

$$P = \frac{\text{実験結果に含まれる正しい結果の数}}{\text{実験結果の数}} \tag{3.7}$$

この指標を用いて，本手法の精度を論文誌について評価した結果を表3.3に示す．ただし，本手法は文書の特徴を概念を抽出するのが目的であるため，人手による抽出では文書の出現単語だけを選び出すのではなく，主題の内容が該当する分野についても考慮して抽出語を選び出した．具体的には，以下の三つを参考にして抽出語を選びだした．

- EDR 電子化辞書
- シソーラス辞書検索　[http://search.kcs.ne.jp/the/]
- 専門書籍のコンテンツやインデックス

アンケート調査

利用者の視点から実験結果を評価するために，以下の専門分野で研究する本学の学生5人を対象にアンケート調査を行なった．

- 画像処理と認識（1人：博士後期課程，3人：博士前期課程）
- 自然言語処理（1人：博士前期課程）

アンケートは，まず，上述の論文誌に掲載されている15本の論文を読んでもらい，“本手法で抽出した論文の主題が妥当であるかどうか”という問いに対して“excellent”，“good”，“fair”，“poor”の4段階で評価してらうものである．アンケートの結果を表3.8に示す．その結果，76%の被験者が“excellent”または，

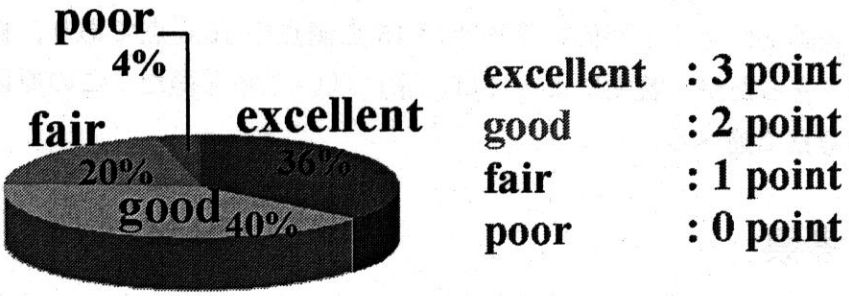


図 3.8 アンケート質問 “本手法で抽出した論文の主題が妥当であるかどうか” に対する結果

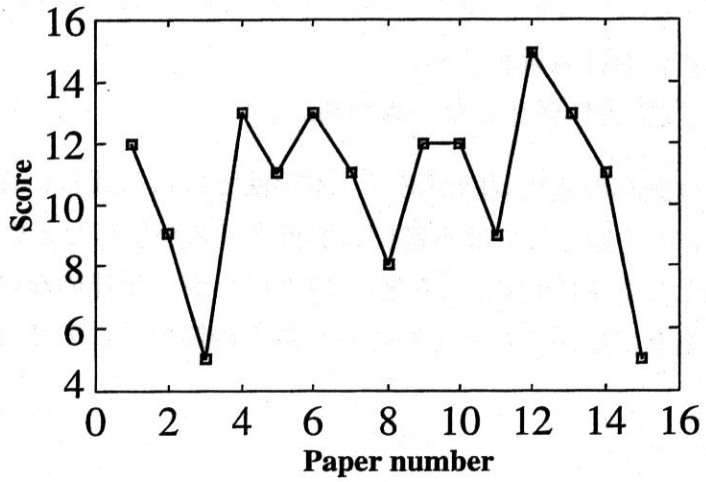


図 3.9 アンケートによる各論文の評価結果

“good” と評価しており，概念で表現された論文の主題が被験者の想定する主題とほぼ一致していることがわかる．

実験結果の状況をさらに細かく見るために，“excellent” を 3 点，“good” を 2 点，“fair” を 1 点，“poor” を 0 点とし，各論文の評価を点数化したグラフを図 3.9 に示す．各論文に対する評価の平均点は 15 点満点中 10.6 点であり，総合的には高いものであったが，論文によっては非常に低い評価を得た．この原因については次節の考察で述べる．

3.3 考察

評価結果において，再現率およびアンケートの結果は総合的に満足のいくものであったが，適合率はあまり良くなかった．これは，各抽出語に重みとして G の値を設けているが，この重みに対して閾値を設定していないためだと考えられる．適合率を上げるには，論文毎に最適な閾値を設定する必要があるが，不適切な閾値はかえって精度を落とすため使用しなかった．

また，アンケートの結果において，論文の主題によって非常に低い評価がなされたものがあった．この評価の原因には，次の 2 点が考えられる．

- 形態素解析における解析ノイズ
- EDR 電子化辞書に存在しない専門用語

形態素解析には既存の連接規則がある．本手法ではこの規則に従った解析結果を用いたため文書の主題としては不適切な解析ノイズが生じたと思われる．また，既存の知識情報として EDR 電子化辞書を利用したが，記述のない単語の概念に対する処理が妥当ではなかったと思われる．以下の節で，この 2 点についての考察を述べる．

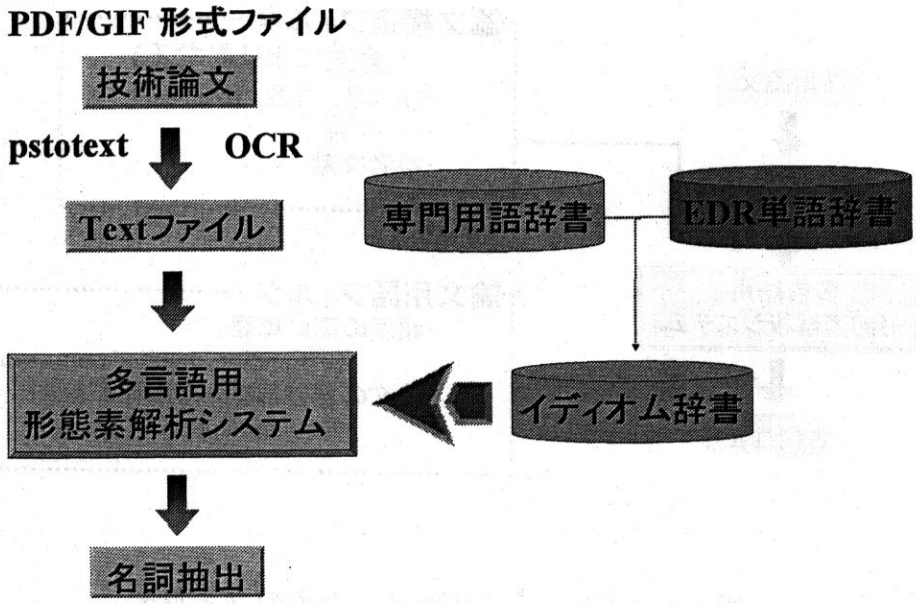


図 3.10 名詞抽出処理における辞書の参照

3.3.1 解析ノイズ除去フィルタ

本手法は概念情報を文書中の単語に対して求めるため，文書の単語が不適切に解析されると正しい概念情報を得ることができない．この不適切な形態素解析は，以下の二つに対する品詞分解が原因だと考えられる．

- 複合名詞や略式名
- 文書内容に関係のない文字列

本手法では形態素解析に既存の連接規則を用いたため，複合名詞に関係なく品詞分解してしまう．例えば,「コンピュータビジョン」は“コンピュータ”と“ビジョン”の二つの普通名詞として解釈されてしまう．このような複合名詞への対処として，文中で連続して出現する普通名詞の並びを一つの複合名詞として解釈するようにした．また，固有名詞，地名，人名，未定義語などの単語に対しても同様に連続する場合は複合名詞として取り扱った．しかし，連続して抽出される単語の並びを全て一つの複合名詞として扱っては抽出結果が冗長な名詞の羅列

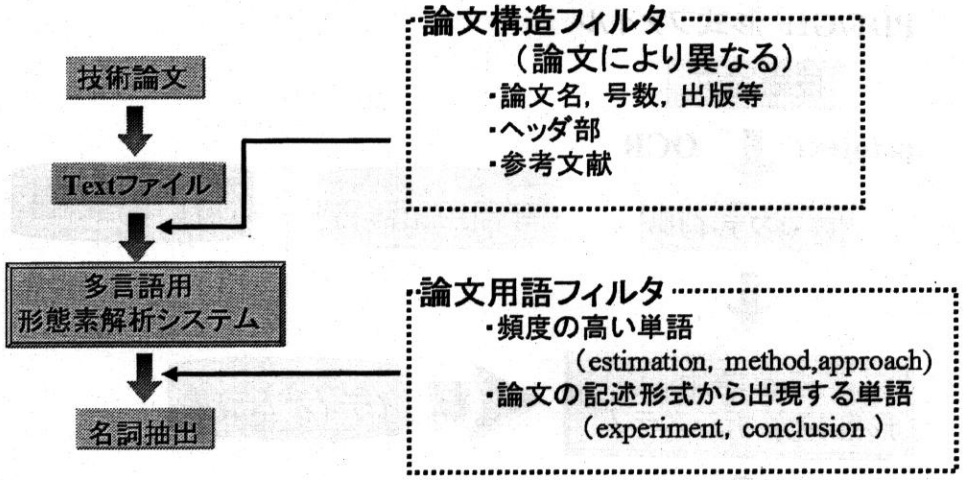


図 3.11 不要文字を削除するためのフィルタ

となり，その内容は不明瞭になりかねない．例えば，“ computer vision field ”や“ image processing applications ”のような連続する名詞から成る複合名詞において，「field」や「applications」は文意的に不要である．“ computer vision field ”は“ computer vision ”という複合名詞と“ field ”という普通名詞として解釈されなければならない．“ image processing applications ”についても同様である．そこで，既存の単語辞書と専門用語辞書から複合名詞を抽出し，これをイディオム辞書として作成した．図 3.10に示すように，品詞分解の際にこの辞書を参照することで冗長な複合名詞を取り除いた名詞の抽出を行うことができる．また，略式名などの複数の表現をもつ複合名詞は，このイディオム辞書に記載してある名称に予め統一しておく必要がある．例えば，「3 次元」は“ 3D ”や“ 3 dimension ”など複数の表現をもつが，イディオム辞書にあるのが“ 3D ”であるとすると，文書の中にある“ 3 次元 ”は“ 3 ”と“ 次元 ”に品詞分解されてしまう．このような表記の揺れを解消するために，品詞分解の前に略式名のある単語を修正しておくフィルタを設置した．

文書に記載している全ての文字列に対して形態素解析を行った結果には，その文書の主題としては不適切な単語が多く含まれている．その要因として以下の 2

つが考えられる．

・本文以外の構造部に記載された単語

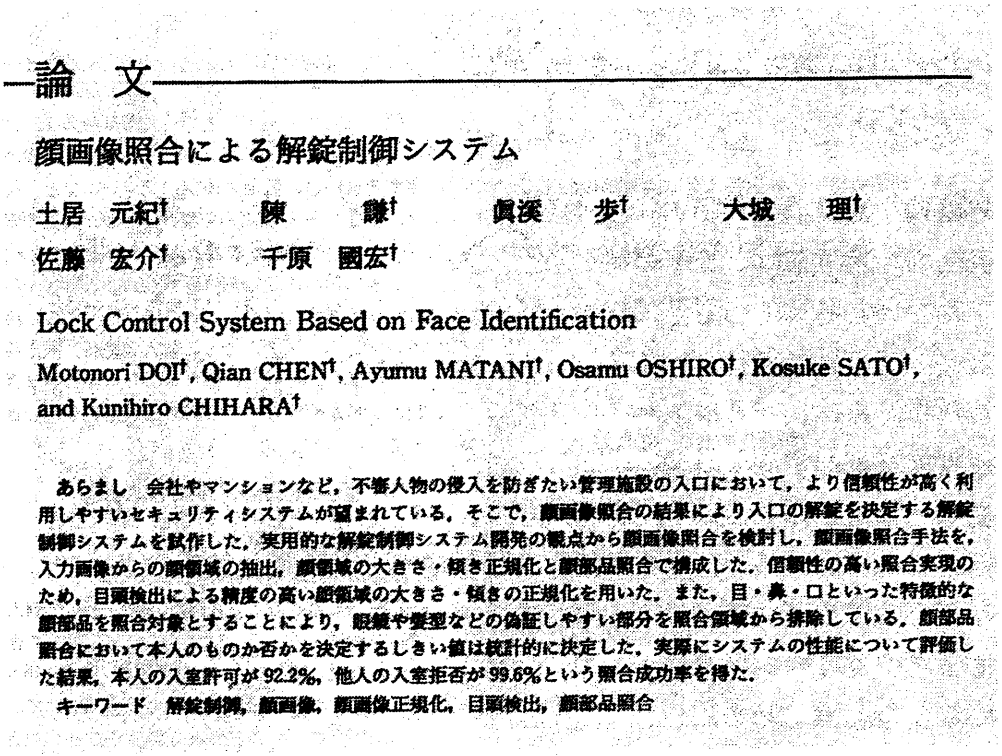
・記述形式として用いられる単語

本文以外の構造部に記載された単語としては，ヘッダ部分に書かれている号数や出版社名，参考文献などが挙げられる．これらの補足情報は本文と同等に扱うべきではない．号数や出版社名は不要単語であり，参考文献には著者名や書籍名など本文とは直接関係しない単語が多く含まれている．また，記述形式として用いられる単語には，技術論文を例にすると“estimation”や“method”などの技術的な内容を説明するために使用される単語や，“experiment”や“discussion”など文書の構造上用いられる単語が挙げられる．これらの単語は文書内容の説明に使用される単語であり，それ自体に主題としての意味をもたない不要単語である．このような不要単語は予め取り除いておく必要がある．対象とする文書のジャンルによって取り除く内容は大きく変わるが，本手法はディジタル図書館で保管されている技術論文を対象に，図3.11に示すような「論文構造フィルタ」と「論文用語フィルタ」の二つを用いた．効率的に処理するために，論文フィルタでは形態素解析の前に号数や出版社名などの不要な単語を取り除き，論文用語フィルタでは形態素解析の後で取り除いた．

上記にあげるようなイディオム辞書とフィルタの使用によって形態素解析におけるノイズを取り除くことができると考える．

3.3.2 概念で表現できない用語に対する処理

EDR電子化辞書は大規模な知識情報からなるシソーラスであるが，その内容は全てを網羅しているわけではない．本手法は知識体系をこの辞書を用いて構築しているため，掲載されていない用語に対する概念が妥当ではなかったと考えられる．例えば，電子通信情報学会に掲載された技術論文“顔画像照合による解錠制御システム（図3.12）”を例にとってみると，主題の抽出処理を行った結果は表3.4のようになる．この表は，主題として抽出した概念を文書の主題を表す順番に示したものであり，太字はEDR電子化辞書に掲載されていなかった単語である．ただし，掲載されていない単語についての重みGは主張度のみから求め



論　文

顔画像照合による解錠制御システム

土居　元紀† 陳　謙† 眞溪　歩† 大城　理†
佐藤　宏介† 千原　國宏†

Lock Control System Based on Face Identification

Motonori DOI†, Qian CHEN†, Ayumu MATANI†, Osamu OSHIRO†, Kosuke SATO†, and Kunihiro CHIHARA†

あらまし　会社やマンションなど，不審人物の侵入を防ぎたい管理施設の入口において，より信頼性が高く利用しやすいセキュリティシステムが望まれている．そこで，顔画像照合の結果により入口の解錠を決定する解錠制御システムを試作した．実用的な解錠制御システム開発の観点から顔画像照合を検討し，顔画像照合手法を，入力画像からの顔領域の抽出，顔領域の大きさ・傾き正規化と顔部品照合で構成した．信頼性の高い照合実現のため，目頭検出による精度の高い顔領域の大きさ・傾きの正規化を用いた．また，目・鼻・口といった特徴的な顔部品を照合対象とすることにより，眼鏡や髪型などの偽証しやすい部分を照合領域から排除している．顔部品照合において本人のものか否かを決定するしきい値は統計的に決定した．実際にシステムの性能について評価した結果，本人の入室許可が92.2%，他人の入室拒否が99.6%という照合成功率を得た．

キーワード　解錠制御，顔画像，顔画像正規化，目頭検出，顔部品照合

図 3.12 主題抽出に用いた技術論文“顔画像照合による解錠制御システム”

た．このような辞書から概念で表現できない文書の頻出単語については，文書の主題に関連深い用語として蓄積することが必要である．図3.13は，概念で表現することができない単語に対する処理の流れを示す．技術論文を題材とした場合，辞書に記載されていない頻出単語は専門用語として解釈することができ，これを蓄積することで著者によって新しく定義された専門用語への対処も可能となる．また，専門用語に対する概念空間の構築には，電子化された専門書籍のコンテンツや構成等から関連分野の用語と関連性を利用することで解決されると考える．

表 3.4 論文主題の抽出結果における概念で表現できない用語

特徴量の重み G	特徴量
1.90	画, 画像, 画像 映像
1.50	こと, 分布, 方 事象の状態
1.30	こと, もの 概念 , 画像, 図 具体物の形
1.00	顔, 鼻 動物の部分, いろいろな具体物の属性 照合
0.90	しきい値, しきい値決定
0.93	登録, 顔画像
0.80	フレーム, システム 構造
0.62	顔部品
0.56	正規化
0.43	抽出, 相関値, 口
0.37	顔領域, テンプレート, テンプレートマッチング
0.31	装備, 背景 静物 physical object 画, 背景 絵を主とする表現物, 処理
0.31	顔部品照合, 顔画像照合
0.25	類似, 利用, 排除, 入室許可, 照合領域, 照合評価, 照合成功, 処理時間, 解錠制御システム, 右目

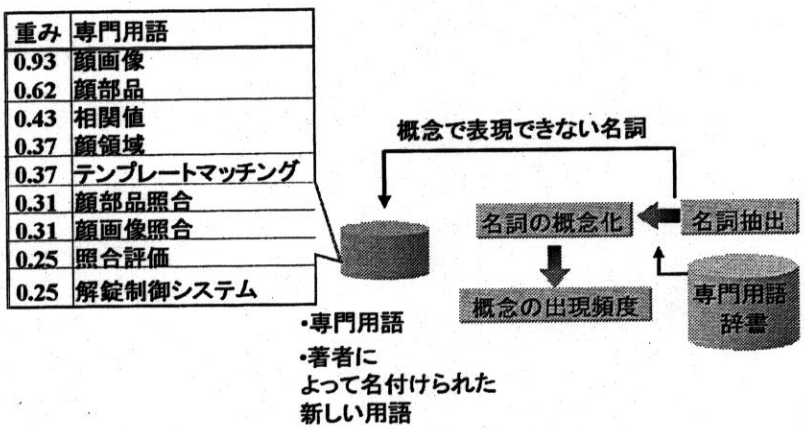


図 3.13 概念で表現できない専門用語に対する処理

第4章
主題抽出法の応用

本章では，提案した文書主題の抽出法の応用として意味検索と文書構造に基づいた内容抽出について述べる．

4.1 意味検索

提案した主題の抽出法は，様々な情報活用に適用することができる．本論文では，その適用の一つとしてディジタル図書館の情報検索を取り上げ，“意味検索”を実現する．意味検索は，従来の全文検索のような単語体系に基づいた検索手法ではなく，単語の深層的な意味として単語の概念を求め，文書の主題や利用者の検索意図を推測して検索に用いる．そのため，単語のうらに隠された意味や検索意図に対する不足情報を推測することが可能になり，図書館司書の協力なしに利用者の要求に適切な検索結果を提供することが期待される．

意味検索の処理の流れを図4.1に示す．この図にあるように，意味検索は，文書の主題を概念を用いて選定する索引作成部，検索質問を概念で表現する検索式部，および検索式と索引語を照合する照合部の3部から構成される．ただし，ここでの検索対象はディジタル図書館で保管されている技術論文とする．以下，それぞれの処理について詳しく述べる．

4.1.1 索引語作成部における技術論文の概念獲得

意味検索における索引語作成部では，まず，入力となる文書テキストに対して形態素解析を行い，その結果から名詞を抽出する．この名詞に対して本手法を適用し，概念で索引語を獲得する．以下，獲得処理について説明する．

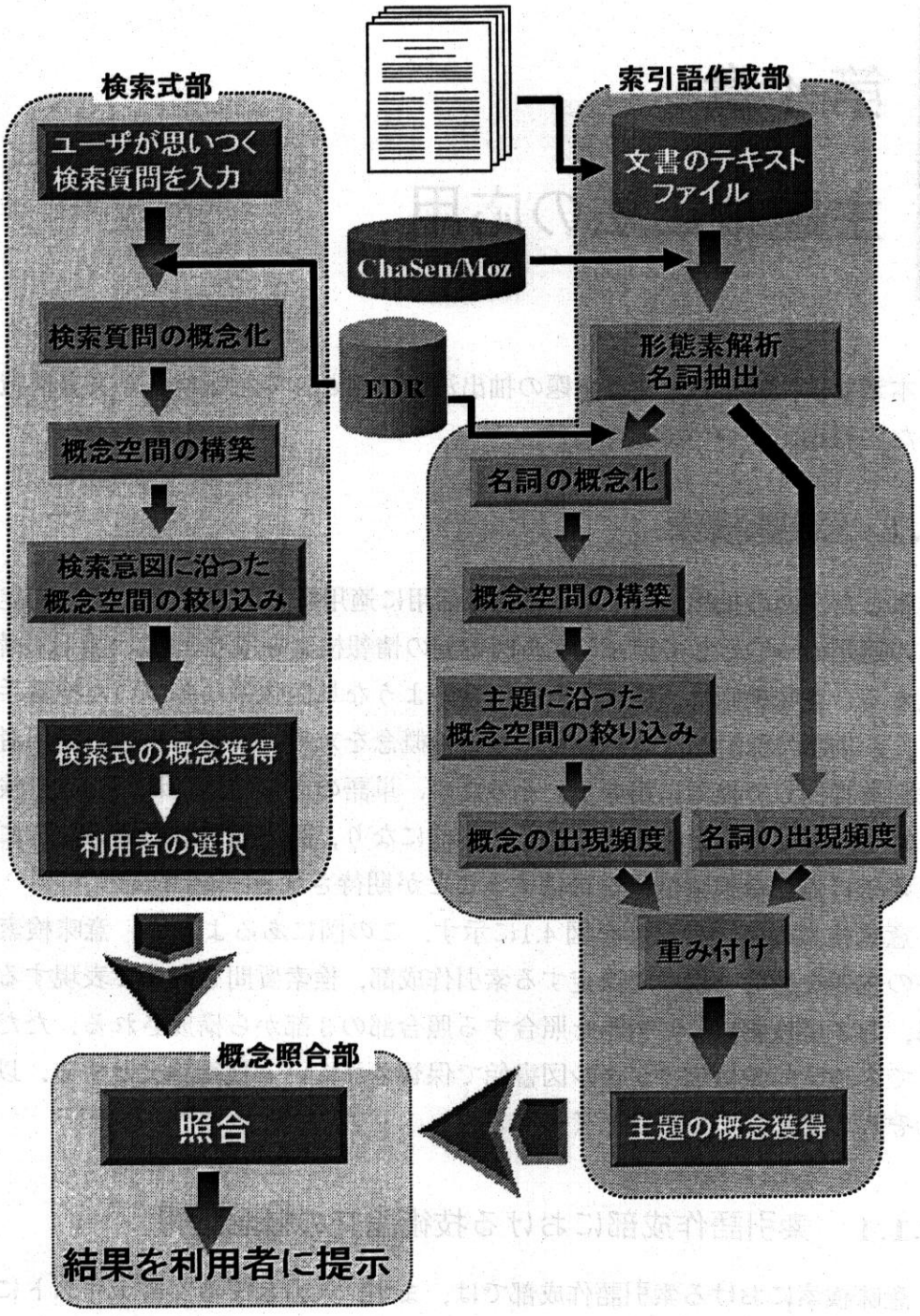


図 4.1 概念情報に基づいた意味検索モデルの概要

論文画像のテキスト処理

意味検索ではディジタル図書館への実装を前提としているため，本論文ではディジタル図書館が実際に保管している論文画像に対して文書処理を行う．

本学附属図書館が運営するディジタル図書館において，図書や資料の電子化作業は自動ページ送りが可能なスキャナによって半自動的に電子メディアとして保存されている．電子化された一次情報は，400DPI と 100DPI の 2 種類の解像度をもった画像として保管する．400DPI の画像は OCR ソフトウェアを用いてテキスト情報を抽出するのに用いられ，100DPI の画像は閲覧用として用いられる[7][16]．ディジタル図書館が稼働し始めた当初，文献は一般的に画像で提供されていたが，OCR 処理によるテキストには図 4.2のような多くの誤字脱字を含むため，近年においては，画像に代って PDF(Portable Document Format) 形式で文献を提供するディジタル図書館が増加している．PDF 形式は，テキストを文字コードで表現するため，オリジナルの文献に忠実なテキストを得ることが出来る．また，文献に含まれるあらゆる内容（文字、画像、レイアウト情報など）を一つのファイルに収めることや，パスワードの設定により利用者を限定することが可能である．本学のディジタル図書館においても，電子メディアの形式が画像から PDF へと移行しつつあるが，現段階では両方を取り扱っている．そのため，本研究でも画像と PDF の両形式を入力として扱うものとする．

形態素解析における名詞の抽出

図 4.2に示すようなテキストに対して形態素解析を行い，その結果から名詞を抽出する．日本語形態素解析システムには，本学の松本らによる『ChaSen』[78] を用い，多言語形態素解析システムとしては山下による『Moz』[79] を用いる．名詞抽出の処理について以下に示す．

(1) **前処理** 入力をプレーンのテキストとする．ChaSen/Moz は入力に対し一行ごとに形態素解析を行なうため，論文を一文ごとに区切り，改行コードを挿入する．

(2) **ChaSen/Moz による品詞分解** ChaSen/Moz を用いて形態素解析を行なう．図 4.2のテキストデータを前処理し，形態素解析を行なった結果の一部

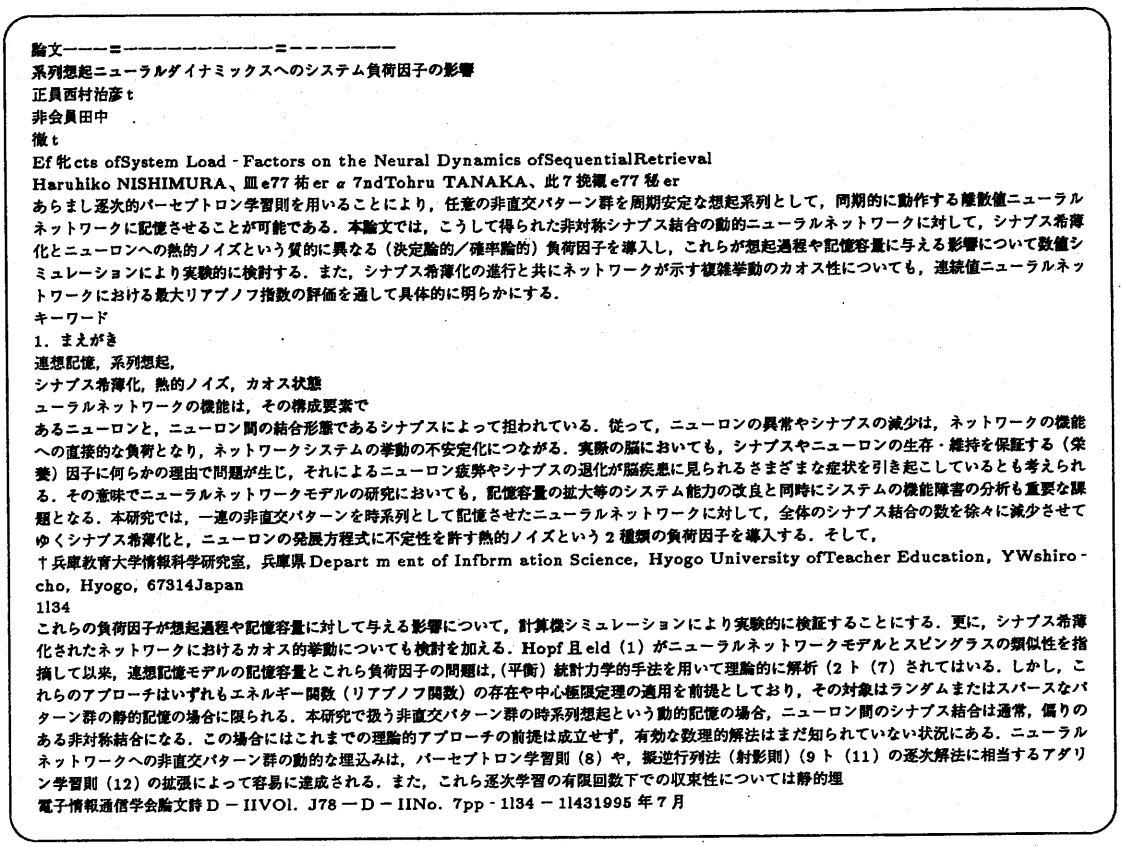

論文−−−=−−−−−−−−−−−−=−−−−−−−−
系列想起ニューラルダイナミックスへのシステム負荷因子の影響
正員西村治彦 t
非会員田中
徹 t
Ef 牝cts ofSystem Load - Factors on the Neural Dynamics ofSequentialRetrieval
Haruhiko NISHIMURA、皿e77 祐 er α 7ndTohru TANAKA、此 7 挽禰 e77 秘 er
あらまし逐次的パーセプトロン学習則を用いることにより，任意の非直交パターン群を周期安定な想起系列として，同期的に動作する離散値ニューラルネットワークに記憶させることが可能である．本論文では，こうして得られた非対称シナプス結合の動的ニューラルネットワークに対して，シナプス希薄化とニューロンへの熱的ノイズという質的に異なる（決定論的／確率論的）負荷因子を導入し，これらが想起過程や記憶容量に与える影響について数値シミュレーションにより実験的に検討する．また，シナプス希薄化の進行と共にネットワークが示す複雑挙動のカオス性についても，連続値ニューラルネットワークにおける最大リアプノフ指数の評価を通して具体的に明らかにする．
キーワード
1．まえがき
連想記憶，系列想起，
シナプス希薄化，熱的ノイズ，カオス状態
ューラルネットワークの機能は，その構成要素で
あるニューロンと，ニューロン間の結合形態であるシナプスによって担われている．従って，ニューロンの異常やシナプスの減少は，ネットワークの機能への直接的な負荷となり，ネットワークシステムの挙動の不安定化につながる．実際の脳においても，シナプスやニューロンの生存・維持を保証する（栄養）因子に何らかの理由で問題が生じ，それによるニューロン疲弊やシナプスの退化が脳疾患に見られるさまざまな症状を引き起こしているとも考えられる．その意味でニューラルネットワークモデルの研究においても，記憶容量の拡大等のシステム能力の改良と同時にシステムの機能障害の分析も重要な課題となる．本研究では，一連の非直交パターンを時系列として記憶させたニューラルネットワークに対して，全体のシナプス結合の数を徐々に減少させてゆくシナプス希薄化と，ニューロンの発展方程式に不定性を許す熱的ノイズという 2 種類の負荷因子を導入する．そして，
† 兵庫教育大学情報科学研究室，兵庫県 Depart m ent of Infbrm ation Science, Hyogo University ofTeacher Education, YWshiro - cho, Hyogo, 67314Japan
1134
これらの負荷因子が想起過程や記憶容量に対して与える影響について，計算機シミュレーションにより実験的に検証することにする．更に，シナプス希薄化されたネットワークにおけるカオス的挙動についても検討を加える．Hopf 且 eld（1）がニューラルネットワークモデルとスピングラスの類似性を指摘して以来，連想記憶モデルの記憶容量とこれら負荷因子の問題は，(平衡）統計力学的手法を用いて理論的に解析（2 ト（7）されてはいる．しかし，これらのアプローチはいずれもエネルギー関数（リアプノフ関数）の存在や中心極限定理の適用を前提としており，その対象はランダムまたはスパースなパターン群の静的記憶の場合に限られる．本研究で扱う非直交パターン群の時系列想起という動的記憶の場合，ニューロン間のシナプス結合は通常，偏りのある非対称結合になる．この場合にはこれまでの理論的アプローチの前提は成立せず，有効な数理的解法はまだ知られていない状況にある．ニューラルネットワークへの非直交パターン群の動的な埋込みは，パーセプトロン学習則（8）や，擬逆行列法（射影則）（9 ト（11）の逐次解法に相当するアダリン学習則（12）の拡張によって容易に達成される．また，これら逐次学習の有限回数下での収束性については静的埋
電子情報通信学会論文詩 D − IIVOl. J78 −D − IINo. 7pp - 1134 − 11431995 年 7 月

図 4.2 論文画像に対する OCR 処理結果の例

を図 4.3に示す．また，日本語の例文に対する ChaSen の形態素解析結果を図 4.4に，英語の例文に対する Moz の形態素解析結果を図 4.5に示す．

(3) **名詞類を抽出** ChaSen/Moz が名詞と判断するものを分類すると，普通名詞/NN，サ変名詞，固有名詞/NNP，地名/NNP，人名/NNP，数詞/CD，形式名詞/JJ，副詞的名詞/RB がある．また，辞書に載っていない単語は未定義語 (null) と判断する．本手法では，ChaSen/Moz が名詞と判断した単語の中から，普通名詞/NN，サ変名詞，固有名詞/NNP，地名/NNP，人名/NNP，形式名詞及び未定義語/null を単語として抽出する．ただし，形態素解析の誤りである未定義語を名詞に含まないために，「漢字以外の 1 文字」と「平仮名だけからなる単語」は予め取り除いておく．

あら [Y:あら POS:感動詞]
ま [Y:ま POS:名詞接頭辞]
し [Y:し POS:述語接続助詞---し]
逐次 [Y:ちくじ POS:副詞]
的 [Y:てき POS:形容詞性名詞接尾辞-ナ形容詞-X]
パーセプトロン [Y:ぱーせぷとろん POS:普通名詞]
学習 [Y:がくしゅう POS:サ変名詞]
則 [Y:そく POS:動詞-サ変動詞-語幹]
を [Y:を POS:格助詞---を]
用い [Y:もちい POS:動詞-母音動詞-X]
る [Y:る POS:動詞性接尾辞-母音動詞-X-る]
ことに [Y:ことに POS:副詞]
より [Y:より POS:格助詞---より]
, [Y:, POS:読点]
任意 [Y:にんい POS:普通名詞]
の [Y:の POS:助動詞-ナ形容詞-X-のだ]
非 [Y:ひ POS:ナ形容詞接頭辞]
直交 [Y:ちょっこう POS:サ変名詞]
パターン [Y:ぱたーん POS:普通名詞]
群 [Y:ぐん POS:普通名詞]
を [Y:を POS:格助詞---を]
周期 [Y:しゅうき POS:普通名詞]
安定 [Y:あんてい POS:サ変名詞]
な [Y:な POS:終助詞---な]
想起 [Y:そうき POS:サ変名詞]
系列 [Y:けいれつ POS:普通名詞]
と [Y:と POS:名詞接続助詞---と]
し [Y:し POS:述語接続助詞---し]
て [Y:て POS:引用助詞]
, [Y:, POS:読点]

図 4.3 図 4.2 のテキストに対する形態素解析結果

(例文 1) コンピュータグラフィックス(CG)で物体を表現するには，物体の形状と光学特性の情報が必要である.

形態素解析「ChaSen」の結果

```
コンピュータグラフィックス
[Y:こんぴゅーたぐらふぃっくす POS:普通名詞]
( (null)
CG (null)
) (null)
で [Y:で POS:判定詞-判定詞-ダ列タ系連用テ形]
物体 [Y:ぶったい POS:普通名詞]
を [Y:を POS:格助詞--を]
表現 [Y:ひょうげん POS:サ変名詞]
する [Y:する POS:動詞-サ変動詞-基本形-する]
に [Y:に POS:名詞接続助詞]
は [Y:は POS:副助詞--は]
, [Y:, POS:読点]
物体 [Y:ぶったい POS:普通名詞]
の [Y:の POS:格助詞]
形状 [Y:けいじょう POS:普通名詞]
と [Y:と POS:格助詞--と]
光 [Y:ひかる POS:固有名詞]
学 [Y:まなぶ POS:固有名詞]
特性 [Y:とくせい POS:普通名詞]
の [Y:の POS:格助詞]
情報 [Y:じょうほう POS:普通名詞]
が [Y:が POS:述語接続助詞--が]
必要 [Y:ひつよう POS:形容詞-ナノ形容詞-X]
で [Y:で POS:動詞-母音動詞-基本連用形]
ある [Y:ある POS:連体詞]
. [Y:てん POS:数詞]
(null)
EOS
```

図 4.4 ChaSen による形態素解析結果の例

索引語の抽出

検索時の適合性を向上させるには，検索質問の妥当性に加えて，文書の主題となる索引語を正確に抽出することが必要である．意味検索では，文書から重複語を取り除いた出現単語に対して概念空間を生成し，この空間から索引語を選定する．索引語の選定には 3 章で提案した文書主題の抽出法を用いる．

以下に索引語を概念で獲得する処理の流れを示す.

(1) 技術論文のテキストデータに対する形態素解析結果から名詞を抽出し，その文字列に関する出現頻度を求める.

(2) 名詞を概念識別子に変換し，概念空間を構築する.

(3) 名詞から生成された概念空間の照合を行い，絞り込んだ部分概念空間の概念パスとその出現頻度を求める.

(4) 名詞の文字列および概念に関する出現頻度分布から論文の特徴量を獲得する.

(5) 論文の主題を概念パスの集合として保管する.

(例文 2) Accurate estimation of the optical flow field of an image sequence is critically important to a number of computer vision and image processing applications.

形態素解析「Moz」の結果

```
Accurate (null)
estimation [NN Pr:1/179722]
of [IN Pr:30999/134926]
the [DT Pr:56300/111243]
optical [JJ Pr:18/82405]
flow [NN Pr:69/179722]
field [NN Pr:94/179722]
of [IN Pr:30999/134926]
an [DT Pr:4210/111243]
image [NN Pr:86/179722]
sequence [NN Pr:2/179722]
is [VBZ Pr:8789/27619]
critically [RB Pr:3/37798]
important [JJ Pr:233/82405]
to [TO Pr:29973/30195]
a [DT Pr:25820/111243]
number [NN Pr:428/179722]
of [IN Pr:30999/134926]
computer [NN Pr:409/179722]
vision [NN Pr:16/179722]
and [CC Pr:21623/32235]
image [NN Pr:86/179722]
processing [NN Pr:35/179722]
applications [NNS Pr:41/81122
                BF:application]
. [. Pr:52726/53362]
EOS
```

図 4.5 Moz による形態素解析の例

4.1.2 検索式部における検索式の概念獲得

検索時の適合性問題は，利用者が入力する検索質問の妥当性に大きく依存する．そのため，検索の精度を上げるには単語のうらに隠された意味や利用者の表現不足から生じた検索意図と検索質問との違いを推測する必要がある．本手法では，単語がもつ概念情報を用いて深層的な観点から検索式を導く．3 章で提案した主題の抽出法は文書を対象にした手法であるが，検索質問を文書中の単語に置き換えることで検索意図を探索する手法として用い，適切な単語を選べない利用者の検索意図を解析し，検索質問における意味の曖昧性を解消する．

以下に検索式を概念で獲得する処理の流れを示す．

(1) 利用者が入力した検索質問（複数可）を概念識別子に変換する．
(2) 各概念識別子について上位概念を再帰的に検索し，概念空間を構築する．
(3) 概念空間を絞りこみ，利用者の要求に添った概念パスを獲得する．
(4) 絞り込まれた概念パスとその内容を利用者に提示する．

上記の処理から得られた概念パスは検索式の候補である．利用者が入力した検索質問の数や質により，この候補の数は異なる．最終的には，この候補の中から利用者に検索意図に沿ったものかどうかを確認してもらい，その結果を検索式として用いる．利用者に確認することなくこの候補を全て検索式としてもよいが，検索意図の真偽は利用者のもつ個人的な概念に依存するため，検索結果に多くのノイズを含む原因となる．そのため，本研究では検索式の選択については支援的な立場をとった．

4.1.3 概念照合部

概念照合部では，上記の概念パスで表現された検索式と概念パスの集合として保管されている技術論文とを照合し，この結果を WWW ブラウザを用いて利用者に提示する．照合する概念は 16 進数で表された概念識別子で表されるため，文字列に比べてその照合は非常に容易である．また，概念識別子は日本語と英語とで共通し，異なる言語間における検索が可能である．

4.1.4 実装モデル

意味検索を試験的に実装したものを図 4.6～図 4.9に示す．図 4.6に検索条件の入力画面を示す．ここで利用者は，検索したい単語を検索質問とて入力する．更に，検索精度として本手法で用いる概念空間の階層の高さを選択する．ここでは “circuit” という検索質問と “5-RANK” の階層レベル数を検索条件として入力した検索例について紹介する．この検索条件に基づいて検索式を求めた結果の一覧を図 4.7に示す．“circuit” から得られる各概念については見出し語を表示し，利用者はこの見出し語を参照して自分が想定した検索意図を表す項目を選択する．各概念パスの先頭にはチェック欄があり，利用者はこの欄にチェックを入れる．図 4.8は “circuit” という単語を “電気回路” という意味で選択した場合である．利用者が入力した検索質問の概念と予め論文の主題として保管してある概念とを照合し，その結果を検索意図に適合する論文として図 4.9に示すように利用者へ提示する．

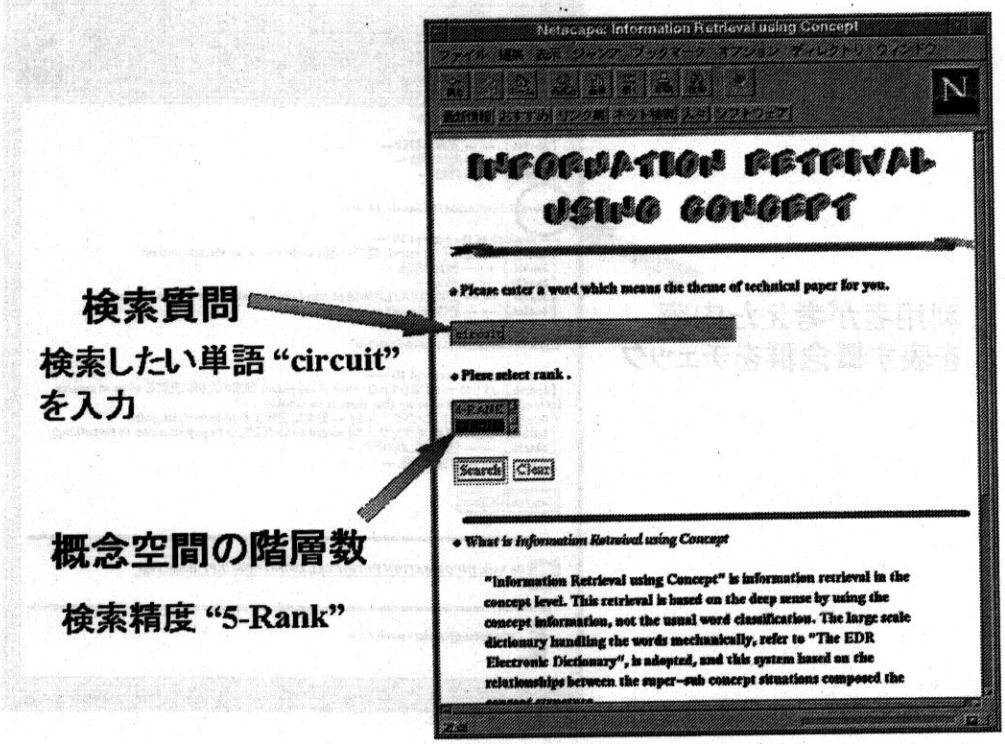


図 4.6 検索画面：検索質問と概念空間の階層数を入力

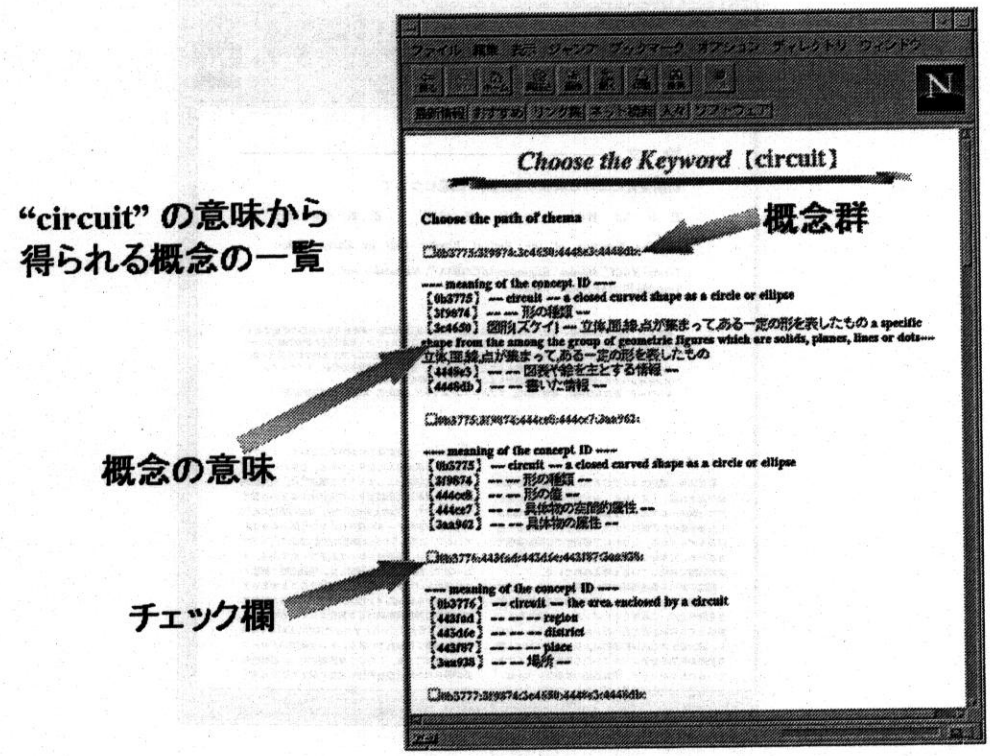


図 4.7 検索質問から得られた概念パスの一覧 (1)

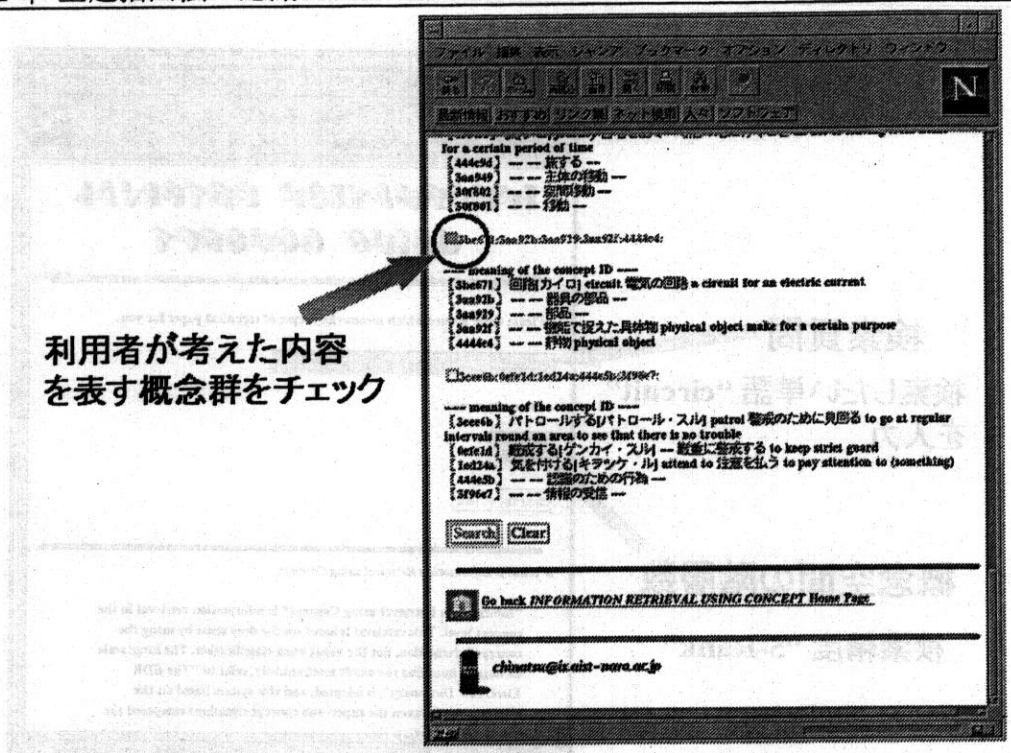


図 4.8 検索質問から得られた概念パスの一覧 (2)

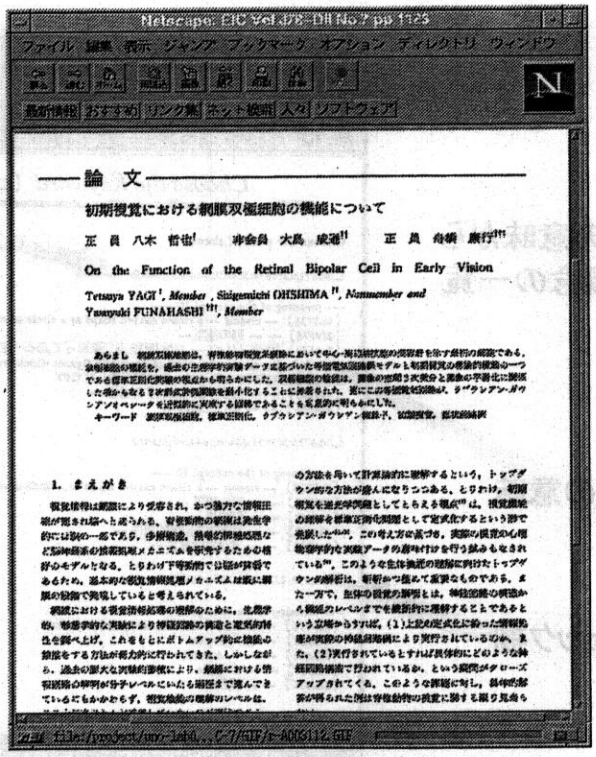


図 4.9 検索結果である論文の閲覧画面

3 次元画像処理を用いた骨梁構造定量的評価法の開発*

大松浩一郎

内容梗概

寝たきり老人の原因疾病として脳卒中についで第2位を占める骨折は骨粗鬆症を背景に発生すると考えられ，老人性骨疾患の診断・治療に必要な新技術の開発が待たれている．現在，この骨粗鬆症は骨密度を基準に診断されている．しかし同程度の骨密度でも骨折する場合としない場合があり，近年では骨梁構造をはじめとする骨質評価も重要とされている．本研究では緻密な骨構造が高精細に観測可能なμX線CT画像複数枚を用い，骨梁構造の新たな3次元的定量評価指標を提案する．また別に骨密度の計測を行ない，提案指標と骨密度指標とを骨強度に関して比較検討した．

まず，健常および骨粗鬆症のラット腰椎横断面画像から皮質骨領域と骨梁領域

図 4.10 比較実験の対象論文 “3 次元画像処理を用いた骨梁構造定量的評価法の開発”

4.1.5 意味検索と全文検索の比較実験

本研究では，現在のディジタル図書館が用いる全文検索のような文字列検索では不可能であった文書主題の検索を意味検索として実現した．文字列検索では，検索質問と文書中の文字列とのパターンマッチングから検索結果を求めているため，検索質問は必ず欲しい文書中の単語でなければその文書を入手することはできない．しかし，利用者は必ずしも検索したい分野の専門家ではなく，また，文書中で著者が使用する単語が適切であるという保証もないため，検索システムに対して適切な検索質問を入力することは非常に困難だといえる．これに対して，文書主題に対する検索手法は，検索質問の表現に多少の揺らぎがあってもその内容が検索対象の主題を指していれば検索もれが生じることはなく，利用者は文書中の単語が思い浮かばなくても欲しい文書を検索結果として得ることができる．

検索手法＼検索質問	X線	レントゲン
全文検索	○	×
提案手法 (意味検索)	○	○

(a)

検索手法＼検索質問	骨梁	骨組織	骨組み
全文検索	○	○	×
提案手法 (意味検索)	○	○	○

(b)

図 4.11 意味検索と全文検索の比較結果

本論文で実現した意味検索が全文検索に対して有効であることを検証するために，比較実験を行った．実験内容は，図4.10に示す修士論文「3次元画像処理を用いた骨梁構造定量的評価法の開発」を入手するために想定したいくつかの検索質問で意味検索と全文検索を行い，この論文を入手することができるか否かを比較するものである．ただし，この実験は検索結果の適合度を問わないため全文検索には高速文字列検索のパトリシア方式を用いる．実験結果を図4.11に示す．図4.11(a)は同義語である“X線”と“レントゲン”の検索質問で全文検索と意味検索を行った結果である．その結果，意味検索では“X線”と“レントゲン”の両方の検索質問で論文を入手できたが，全文検索では“レントゲン”で入手することはできなかった．概して技術論文では“レントゲン”という単語は用いられず，“X線”が使用される．図4.10の論文においても一貫して“X線”が用いられている．これらの単語は同じ内容を指しており，どちらの単語でも同じ論文が検索

されなければならない．この点で全文検索より意味検索の方が有効であるといえる．また，図4.12(b)では，“骨梁”という医学用語が思い付かないために連想した“骨組織”と“骨組み”の検索質問で全文検索と意味検索を行った結果である．やはり，全文検索では文書中に存在しない単語“骨組み”で論文を検索することができなかったが，意味検索では他の単語と同様に論文を入手することができた．ただし，意味検索で使用するEDR電子化辞書には“骨梁”という専門用語が掲載されておらず，“骨”と“梁（はり）”とに分けて概念化している．この場合，検索質問についても“骨梁”を分割して検索に用いるため不適合になることはない．比較実験からわかるように，文書主題に基づいた意味検索は全文検索における文字列検索の問題点である表記の揺れを補った検索手法であるといえる．

意味検索の入力は，全文検索と同様に文書中の単語であるが，検索に用いる蓄積データの形式は両者で大きく異なっている．全文検索の蓄積データがテキスト文書の文字列集合であるのに対し，意味検索で蓄積されるデータは文書の主題を表す概念情報の集合となる．意味検索における蓄積データの内容とその照合法を説明するために，同義語である“X線”と“レントゲン”を例として取り上げ，各単語の概念空間を図4.12と図4.13を示す．図4.12は単語“X線”から構築された概念空間を表している．下段に行くに従って各ノードは上段のノードの上位概念を表している．“X線”の上位概念は“電波”であり，その上位概念は“電気”と“波の形をとって伝わる現象”である．概念情報は上位にいくに従って体系的な内容を表す．意味検索では，文書中の単語“X線”を図4.12に示す概念空間の情報として蓄積している．これに対して，検索質問となる“レントゲン”は，図4.13に示すような枝分れをもつ概念空間で表される．その内容は図4.13に示すように，“X線”と“レントゲン”，“レントゲン写真”の3本の柱で構成されている．利用者はこの中から必要な概念情報のグループを選択し，選択された概念情報と蓄積されている主題の概念情報とを照合することで検索結果を求める．上記の実験では，利用者は“X線”の意味で“レントゲン”を用いており，予め文書の主題として蓄積されている“X線”の概念と適合するため，検索結果として論文を入手することができた．

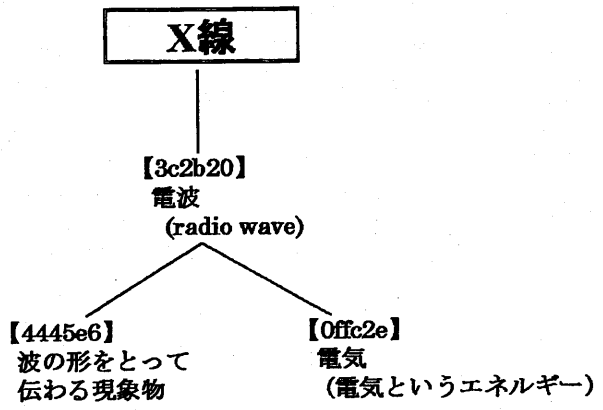


図 4.12 単語“X線”から構築される概念空間

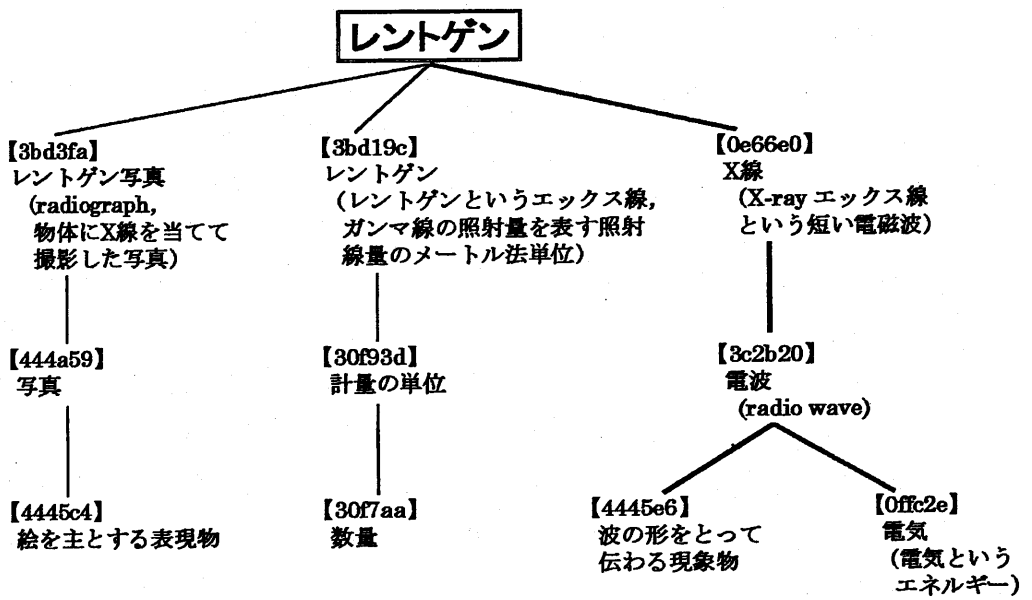


図 4.13 単語“レントゲン”から構築される概念空間

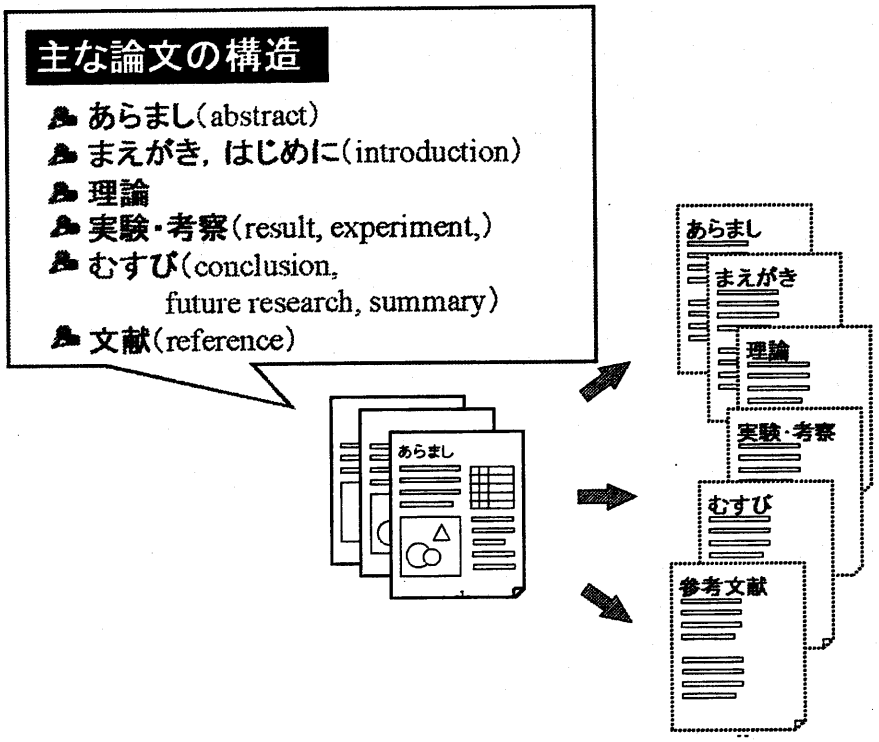


図 4.14 技術論文の文書構造

4.2 文書構造に基づいた内容抽出

ディジタル図書館などの情報検索では，文書を最小単位とした検索結果を利用者に提供しているが，文書はいくつかの事柄について論じた部分の集合であるため，文書を解体して利用者が必要とする単位で検索結果を提供することが求められている [9]．本論文では，構造形式に規則性がある文書を構造に基づいて分割し，これを最小単位とした情報の提示を実現する．提案した文書主題の抽出法を文書全体ではなく各構造部分に適用することで文書構造に基づいた内容の抽出を行う．ただし，ここではディジタル図書館に保管している文書を対象としているため，構造形式に規則性のある文書には技術論文を用いるものとする．

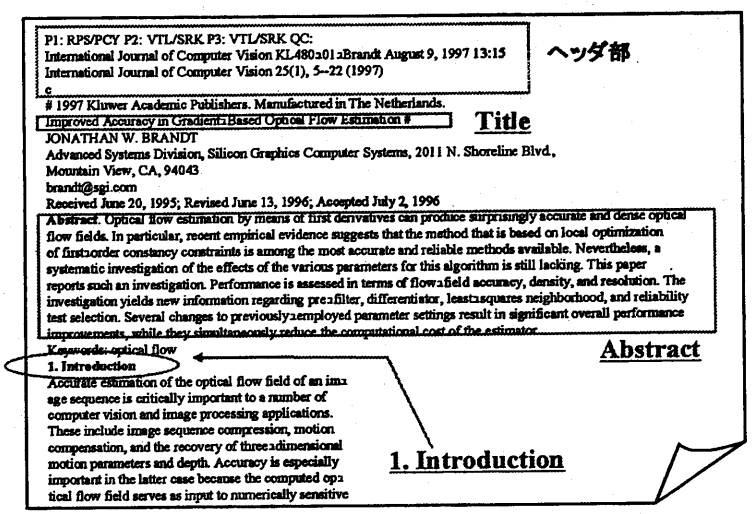


図 4.15 PDF 形式の電子ファイルから抽出したテキストと文書構造

4.2.1 文書構造とその内容

技術論文は主に研究目的で利用され，多くの利用者は具体的な内容について記述した論文を探す．例えば 利用者が “動的輪郭モデルを用いた胃領域の抽出に関する論文” を入手したい場合，そのアルゴリズムについては “動的輪郭モデル” についての論文を探す必要があり，“胃領域の抽出” については胃に対する実験結果を述べている論文を探すべきである．本論文では，このような利用者の具体的な要求に対応するために，技術論文の文書構造に基づいて文書を分割し，各構造部分が表す内容を利用者に提供する．技術論文の構造は図 4.14に示すような六つの構造に分割する．現在のディジタル図書館では画像と PDF 形式の文書ファイルを保管しているが，文書構造に基づいて分割するには文書レイアウトに忠実なテキストの抽出が前提条件である．そのため，文字コードをあらかじめ持っている PDF 形式の文書ファイルを対象として用いる．図 4.15に PDF 形式のファイルから抽出したテキストとその文書構造の例を示す．

また，文書の構造毎に内容を提示することは，多読支援に役立てることができる．日本語で書かれた技術論文は 8 頁前後と短いが，英語の技術論文は数十頁と

非常に長い．そのため，必要な文献だけを手早く入手したい利用者にとって，その論文が自分の必要とする論文か否かを確認することは非常に手間がかかる．文書全体を対象とした要約の利用も考えられるが，手法や実験結果といった論文の詳細な内容について調べることはできない．論文の構造に基づいた内容抽出はこのような場合に非常に有効である．

以下に文書構造に基づいた内容抽出処理の流れを示す．

（1）PDF 形式の電子ファイルを入力とし，2 値文書画像（BPM 形式）とテキストデータを作成する．
（2）2 値文書画像に対して文書構造の解析を行い，文書を六つの領域に分割する．
（3）文書構造の解析結果に基づいて文書のテキストを分割する．
（4）各構造のテキストデータに対して形態素解析を行い，その結果から単語を抽出し，文字列に関する出現頻度を求める．
（5）単語を概念識別子に変換し，概念空間を構築する．
（6）単語から生成された概念空間の照合を行い，絞り込んだ部分概念空間の概念パスとその出現頻度を求める．
（7）単語の文字列および概念に関する出現頻度分布から各構造についての特徴量を獲得する．
（8）各構造の特徴量をその内容として保管する．

上記の処理における各構造の内容抽出には 3 章で提案した文書の主題抽出法を適用する．また，論文画像からの文書構造解析については付録で述べる．

4.2.2 文書構造における内容の提示

文書の各構造部分における内容の抽出は，3 章で提案した主題抽出法の抽出対象を文書全体から各構造部分に変更したものである．そのため，内容の抽出方法は同じものであり，その有効性も同等であると考えられる．しかし，抽出対象を文書全体から各構造部分へと詳細にすることで，同じ文書でも構造の位置によって論じる内容に違いがあることがわかる．図 4.18から図 4.23の概念リストは図 4.17に示す論文を“abstract”，“introduction”，“theory”，“result”，“conclusion”，“reference”の六つの構造部分に分割し，それぞれの内容を抽出した結果である．ただし，抽出結果の提示形式は図 4.16に示す通りである．文書中の単語の概念

■■■概念識別子 (A):概念識別子 (B):概念識別子 (C):　　文書中の単語■■■
概念識別子 (A)　　概念識別子 (A) の見出し語
概念識別子 (B)　　概念識別子 (B) の見出し語
概念識別子 (C)　　概念識別子 (C) の見出し語

図 4.16 文書構造に基づいた内容抽出結果の提示形式

が概念識別子 (A) であり，上位概念が概念識別子 (B)，その上位概念が概念識別子 (C) を表す．また，記載内容は一部抜粋であり，抽出結果のリストは構造の内容を強く表現している順に並べた．図 4.17に示した論文から推測できるように，この論文は“表情の認識”に関する論文であるが，内容の抽出結果より構造の位置によってその特徴が異なっていることがわかる．一般的に“abstract”や“conclusion”は，その名の通りまとめ的な位置づけとなるが，文書の量が少なく重要な個所とそうでない個所との違いがないため，図 4.18や図 4.22の抽出結果のように文書の特徴はあまり表現されない．これに対し，図 4.19や図 4.20に示すような“introduction”と“theory”の抽出結果では，“computer vision”や“optical flow”など手法についての内容が強く現れている．また，図 4.21に示す“result”の抽出結果からは，手法ではなく実験の題材について強く表現している．ここでは，顔の表情についての実験を行っており，内容には“smile”や“brow”，“face”などが抽出された．“reference”では，本文とは直接関係のない著者名や出版社名などが多く含まれ，その内容に一貫性がない．図 4.23の抽出結果においても“Springer”などの固有名詞が抽出されている．しかし，“reference”にある題名の情報から“vision”や“pattern recognize”などの本文に関連のある内容も抽出されており，題名だけを取り出して内容を抽出することが必要だと考えられる．

上記のように論文一つを取り出しても，その構造位置で記述している内容に違いが見られる．文書主題の抽出は文書全体を把握する上で非常に有効であるが，利用者の検索要望が具体的な場合には，このような構造別の内容提示が効果的である．

Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion

MICHAEL J. BLACK
Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304
black@parc.xerox.com

YASER YACOOB
Computer Vision Laboratory, University of Maryland, College Park, MD 20742
yaser@cs.umd.edu

Received March 27, 1995; Revised June 4, 1996; Accepted August 9, 1996

Abstract. This paper explores the use of local parameterized models of image motion for recovering and recognizing the non-rigid and articulated motion of human faces. Parametric flow models (for example affine) are popular for estimating motion in rigid scenes. We observe that within local regions in space and time, such models not only accurately model non-rigid facial motions but also provide a concise description of the motions in terms of a small number of parameters. These parameters are intuitively related to the motion of facial features during facial expressions and we show how expressions such as anger, happiness, surprise, fear, disgust, and sadness can be recognized from the local parametric motions in the presence of significant head motion. The motion tracking and expression recognition approach performed with high accuracy in extensive laboratory experiments involving 40 subjects as well as in television and movie sequences.

Keywords: facial expression recognition, optical flow, parametric models of image motion, robust estimation, non-rigid motion, image sequences

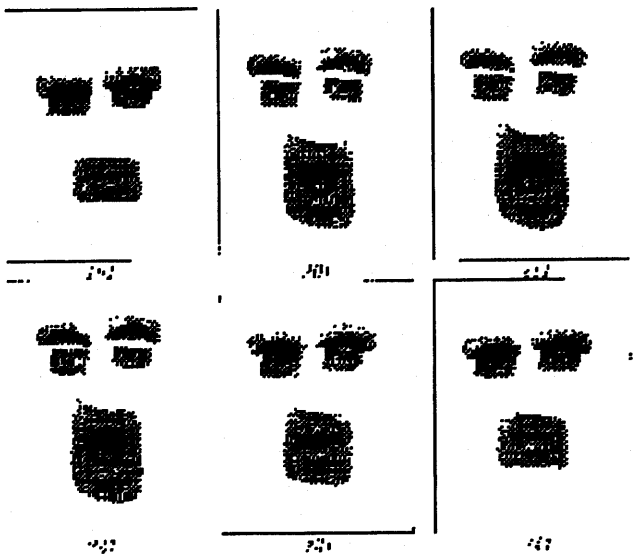


Figure 10. Sample experiment: facial expression tracking. Features every 10 frames.

図 4.17 文書構造の内容抽出に用いた技術論文

■■■ 2f1aea:2f30e2:2f30de: recognition ■■■
2f1aea 認識する [ニンシキ・スル] perceive 物事を認識する to perceive something
2f30e2 − − [情報処理関係のその他の (動作・行為)] −
2f30de − − [情報処理関係の「動作・行為」] −

■■■ 3f96b6:444c8c:3aa94b: flow ■■■
3f96b6 − − 外への移動 −
444c8c − − 始点から終点への移動 −
3aa94b − − 方向の決まっている移動 −

■■■ 0e2fbe:30f83f:30f83e: motion ■■■
0e2fbe 動作 [ドウサ] movement 体の動き movement of the body
30f83f − − 身体的活動 −
30f83e − − 行為 −

■■■ 3cfd49:3ce76f:0faa7a: model ■■■
3cfd49 モデル [モデル] model 見本や模型 a sample or a model
3ce76f 模範 [モハン] exemplar 手本となるもの a typical example which represents the character or state of a particular class, or group
0faa7a 相互関係 [ソウゴカンケイ] interrelationship 二つ以上の事物が対等の資格で形成する関係 the relationship on equal terms between two or more entities, called interrelationship

■■■ 0c2be4:30f785:10625c: image ■■■
0c2be4 − image − a drawing, painting, or carving portraying a person or object's likeness
30f785 − − 絵画 −
10625c 美術品 [ビジュツヒン] work of art 美術の作品 a work of art

■■■ 0bbf1a:30f86f:3f96e6: expression ■■■
0bbf1a − expression − the act of drawing forth the essential aesthetic element of a musical work, passage of text, or work of art
30f86f − − 表現する −
3f96e6 − − 情報の発信 −

図 4.18 “abstract” の内容抽出結果

■■■ 0cd3b5:30f7ab:30f78d: parametric ■■■
0cd3b5 – parametric – of a parameter
30f7ab – – 数の種類 –
30f78d – – 数学用語 –

■■■ 3c9ad2:2e298d:3c7dc1: computer vision ■■■
3c9ad2 コンピュータービジョン [コンピュータービジョン] computer vision コンピュータビジョンという,ビジョン vision named computer vision
2e298d 表示装置 [ヒョウジソウチ] – 表示装置 display;
3c7dc1 周辺装置 [シュウヘンソウチ] peripheral 補助記憶装置と入出力装置の総称 in computing, a
generic name for a subsidiary memory unit and input-output unit

■■■ 2f1aea:2f30e2:2f30de: recognition ■■■
2f1aea 認識する [ニンシキ・スル] perceive 物事を認識する to perceive something
2f30e2 – – [情報処理関係のその他の (動作・行為)] –
2f30de – – [情報処理関係の「動作・行為」] –

■■■ 0d3100:443f66:30f6d8: region ■■■
0d3100 – region – a part of the human anatomy
443f66 – – – part of an animal body
30f6d8 – – 動物の部分 –

■■■ 3bc7ce:3f961f:30f6d8: mouth ■■■
3bc7ce 口 [クチ] mouth 動物の器官としての口 an organ of an animal, called mouth
3f961f – – 身体 –
30f6d8 – – 動物の部分 –

■■■ 443913:443965:4438e7: eye ■■■
443913 – – – to see; to look
443965 – – – to receive information (general)
4438e7 – – – reception of information

■■■ 0ef0d0:444dae:444e40: shape ■■■
0ef0d0 形式 [ケイシキ] form ある一群のものに共通している,外に表れた特徴的な形態 the way of ex
istence of a thing that shows its essential nature
444dae – – 物事の構造や内容 –
444e40 – – 抽象物 –

■■■ 0e2fbe:30f83f:30f83e: motion ■■■
0e2fbe 動作 [ドウサ] movement 体の動き movement of the body
30f83f – – 身体的活動 –
30f83e – – 行為 –

■■■ 2e6a69:2dd230:2f3374: frame ■■■
2e6a69 枠組構造 [ワクグミコウゾウ] – 枠組構造 –
2dd230 構造 [コウゾウ] – 構造 structure
2f3374 – – [計算機の構造属性] –

■■■ 3cd14b:2f2e5d:2f2e59: optical flow ■■■
3cd14b オプティカルフロー [オプティカルフロー] optical flow 光通信におけるオプティカルフロー an optical flow in optical communication
2f2e5d – – [フロー (流れ)] –
2f2e59 – – [現象物全般] –

■■■ 2deb50:2f2d38:2f2d03: template ■■■
2deb50 テンプレート [テンプレート] – テンプレート template;
2f2d38 – – [対象] –
2f2d03 – – [情報処理関係の抽象物] –

■■■ 3cf5c4:3cf6b2:30f7d3: scheme ■■■
3cf5c4 構造 [コウゾウ] organization 物事を構成している仕組み the structure or framework of something
3cf6b2 構造 [コウゾウ] framework 物事の構造や仕組み the system or structure of something
30f7d3 – – 構造 –

■■■ 3be074:4445de:0eeed4: eyebrow ■■■
3be074 眉 [マユ] eyebrow まゆという,顔の部分 hair just above the eye, called eyebrow
4445de – – 顔に生えている毛 –
0eeed4 毛 [ケ] – 哺乳類の皮膚に生ずる毛 the hair that grows on the skin of mammals

図 4.19 “introduction” の内容抽出結果

■■■ 3ceda5:30f7bf:444707: level ■■■
3ceda5 グレード [グレード] level ある集団における価値, 能力あるいは重要性の程度 a degree of value, ability or importance in a group
30f7bf – – 身分や経済状態や能力などの階級的な位置 –
444707 – – 階級, 等級など, 上中下や数字で表わした, 物事や能力のレベル –

■■■ 2f19d9:2df2c4:2f32cf: curvature ■■■
2f19d9 曲率 [キョクリツ] curvature 曲率という, 曲線面上のまがり程度を示す値 curvature which in dicates the slope of a curve
2df2c4 率 [リツ] – 率 rate
2f32cf – – [定量属性全般] –

■■■ 3cd14b:2f2e5d:2f2e59: optical flow ■■■
3cd14b オプティカルフロー [オプティカルフロー] optical flow 光通信におけるオプティカルフロー an optical flow in optical communication
2f2e5d – – [フロー (流れ)] –
2f2e59 – – [現象物全般] –

■■■ 3c9ad2:2e298d:3c7dc1: computer vision ■■■
3c9ad2 コンピュータービジョン [コンピュータービジョン] computer vision コンピュータビジョンという, ビジョン vision named computer vision
2e298d 表示装置 [ヒョウジソウチ] – 表示装置 display;
3c7dc1 周辺装置 [シュウヘンソウチ] peripheral 補助記憶装置と入出力装置の総称 in computing, a generic name for a subsidiary memory unit and input-output unit

■■■ 0d67d9:30f78d:444e41: sequence ■■■
0d67d9 – sequence – a set having its members arranged in numerical order, in mathematics
30f78d – – 数学用語 –
444e41 – – 思考活動の道具 –

■■■ 443913:443965:4438e7: eye ■■■
443913 – – – to see; to look
443965 – – – to receive information (general)
4438e7 – – – reception of information

■■■ 3cf02f:444b93:3f96e6: mask ■■■
3cf02f 覆い隠す [オオイカク・ス] hide 覆い隠す to cover or hide something
444b93 – – 物事を隠す –
3f96e6 – – 情報の発信 –

■■■ 2f1896:2f2ce9:2f2cd6: image plane ■■■
2f1896 図形 [ズケイ] – 立体, 面, 線, 点が集まって, ある一定の形を表したもの a specific shape from the among the group of geometric figures which are solids, planes, lines or dots
2f2ce9 – – [グラフ、図] –
2f2cd6 – – [知的活動の抽象的道具] –

■■■ 3c965f:2f3409:2f3408: approximation ■■■
3c965f 近似 [キンジ] approximation 近似という, 似かよっていること something which is very close to the actual figure named an approximation
2f3409 – – [情報処理関係の属性値全般] –
2f3408 – – [情報処理関係の属性値] –

■■■ 3cc329:2f1718:2dcde0: partial derivative ■■■
3cc329 偏導関数 [ヘンドウカンスウ] partial derivative 偏導関数という, 関数 a derivative named a partial derivative
2f1718 導関数 [ドウカンスウ] derivative 導関数という関数 the derivative of a mathematical function
2dcde0 関数 [カンスウ] – 関数という機能 function

■■■ 3cbea3:2de364:2f32cf: invariant ■■■
3cbea3 不変量 [フヘンリョウ] invariant 物理における不変量 a physical invariant
2de364 量 [リョウ] quantity 量 quantify
2f32cf – – [定量属性全般] –

■■■ 0bcf23:4438f2:4438f3: filter ■■■
0bcf23 – filter – to extract (something) by passing a fluid through a barrier with small holes
4438f2 – – – to pass by, over, through
4438f3 – – – to carry

図 4.20 “theory” の内容抽出結果

■■■ 3bdc87:444d95:30f7ba: smile ■■■
3bdc87 微笑み [ホホエミ] smile 微笑 a smile
444d95 – – 笑い –
30f7ba – – 表情 –

■■■ 3cfd53:4445dc:3f961f: brow ■■■
3cfd53 額 [ヒタイ] forehead 人の額 a forehead
4445dc – – 顔の部分としての額 –
3f961f – – 身体 –

■■■ 0e81af:3f961f:30f6d8: face ■■■
0e81af 顔 [カオ] face 人の顔 a face
3f961f – – 身体 –
30f6d8 – – 動物の部分 –

■■■ 0bc8c6:30f957:3f9616: feature ■■■
0bc8c6 – feature – any one of the conspicuous parts of the face
30f957 – – 身体的特徴の値 –
3f9616 – – 人の外見的印象の値 –

■■■ 0f5035:3f9666:3d1ae7: line ■■■
0f5035 境界線 [キョウカイセン] – 境界線 a boundary line
3f9666 – – 形でみると線になる部分や,線で示す物の部分 –
3d1ae7 線 [セン] line 線 a drawn line

■■■ 3c9ad2:2e298d:2de885: computer vision ■■■
3c9ad2 コンピュータービジョン [コンピュータービジョン] computer vision コンピュータビジョンという,ビジョン vision named computer vision
2e298d 表示装置 [ヒョウジソウチ] – 表示装置 display;
2de885 装置 [ソウチ] – 装置 device;

■■■ 3bf5e4:444d18:3f9874: line ■■■
3bf5e4 軌跡 [キセキ] locus 幾何学で,点の移動した軌跡 in geometry, the path traced by a moving point
444d18 – – いろいろな形 –
3f9874 – – 形の種類 –

■■■ 3cf06f:30f80d:3aa95e: tracking ■■■
3cf06f 追いかける [オイカケ・ル] pursue 先に進んでいるものに後から近づこうとする to approach an advanced thing from the behind
30f80d – – 近づく –
3aa95e – – 接近 –

■■■ 3ce641:444d9b:30f863: anger ■■■
3ce641 立腹する [リップク・スル] anger 腹を立てる to get angry
444d9b – – 対象によって引き起こされる感情活動 –
30f863 – – 感情活動 –

■■■ 0ecff7:3cf86b:30f88f: clip ■■■
0ecff7 切除する [セツジョ・スル] lop 一部分を切って除く to cut off a part of a thing
3cf86b 切り離す [キリハナ・ス] lop 切断する to cut away something
30f88f – – 切る –

図 4.21 “result” の内容抽出結果

■■■ 0e2fbe:30f83f:30f83e: motion ■■■
0e2fbe 動作[ドウサ] movement 体の動き movement of the body
30f83f – – 身体的活動 –
30f83e – – 行為 –

■■■ 0bbf1a:30f86f:3f96e6: expression ■■■
0bbf1a – expression – the act of drawing forth the essential aesthetic element of a musical work, passage of text, or work of art
30f86f – – 表現する –
3f96e6 – – 情報の発信 –

■■■ 0bc8c6:30f957:3f9616: feature ■■■
0bc8c6 – feature – any one of the conspicuous parts of the face
30f957 – – 身体的特徴の値 –
3f9616 – – 人の外見的印象の値 –

■■■ 3cd14b:2f2e5d:2f2e59: optical flow ■■■
3cd14b オプティカルフロー[オプティカルフロー] optical flow 光通信におけるオプテイカルフロー an optical flow in optical communication
2f2e5d – – [フロー(流れ)] –
2f2e59 – – [現象物全般] –
2f2e58 – – [現象物] –

■■■ 2f1aea:2f30e2:2f30de: recognition ■■■
2f1aea 認識する[ニンシキ・スル] perceive 物事を認識する to perceive something
2f30e2 – – [情報処理関係のその他の(動作・行為)] –
2f30de – – [情報処理関係の「動作・行為」] –

図 4.22 “conclusion” の内容抽出結果

■■■ 3c2fb9:4449dd:444634: press ■■■
3c2fb9 印刷所 [インサッジョ] print shop 印刷の作業所 a shop where printing is done
4449dd – – 作業所 –
444634 – – 機能で捉えた建物または建物群 –

■■■ 3cd14b:2f2e5d:2f2e59: optical flow ■■■
3cd14b オプティカルフロー [オプティカルフロー] optical flow 光通信におけるオプテイカルフロー an optical flow in optical communication
2f2e5d – – [フロー (流れ)] –
2f2e59 – – [現象物全般] –

■■■ 0e14c7:30f85d:444e5b: vision ■■■
0e14c7 – vision – the operation or reality of seeing
30f85d – – 見る –
444e5b – – 認識のための行為 –

■■■ 0ae046:30f888:3f96e6: ash ■■■
0ae046 – ash – to become or take the form of ashes
30f888 – – 発行する –
3f96e6 – – 情報の発信 –

■■ 2deb50:2f2d38:2f2d03: template ■■■
2deb50 テンプレート [テンプレート] – テンプレート template;
2f2d38 – – [対象] –
2f2d03 – – [情報処理関係の抽象物] –

■■■ 3c8e75:3c83b2:2f30e2: pattern recognition ■■■
3c8e75 パターン認識 [パターンニンシキ] pattern recognition 自動的な手段によって, 形・輪郭又は構成を認識すること an act of recognizing the form, shape or structure automatically
3c83b2 認識 [ニンシキ] – 人間の能力としての認識 recognition as a human ability
2f30e2 – – [情報処理関係のその他の (動作・

■■■ 3c9ad2:2e298d:3c7dc1: computer vision ■■■
3c9ad2 コンピュータービジョン [コンピュータービジョン] computer vision コンピュータビジョンという, ビジョン vision named computer vision
2e298d 表示装置 [ヒョウジソウチ] – 表示装置 display;
3c7dc1 周辺装置 [シュウヘンソウチ] peripheral 補助記憶装置と入出力装置の総称 in computing, a generic name for a subsidiary memory unit and input-output unit

■■■ 443db8:443cd2:30f6bf: springer ■■■
443db8 – – – fish
443cd2 – – – animal
30f6bf – – 動物 –

■■■ 0ddefc:3f9758:3f9704: transaction ■■■
0ddefc – transaction – something dealt; a negotiation or matter
3f9758 – – 処理する –
3f9704 – – 物事の状態をよくするための行為 –

■■■ 0cd3b5:30f7ab:30f78d: parametric ■■■
0cd3b5 – parametric – of a parameter
30f7ab – – 数の種類 –
30f78d – – 数学用語 –

■■■ 0ed558:3bc690:444d75: conference ■■■
0ed558 議会 [ギカイ] – 議員で組織された議決機関 a decision making organization that consists of members of the Japanese Diet
3bc690 会議 [カイギ] – 機関としての会議 a council as an organization
444d75 – – 組織化された, 検討協議のための会 –

■■■ 0c2be4:30f785:10625c: image ■■■
0c2be4 – image – a drawing, painting, or carving portraying a person or object's likeness
30f785 – – 絵画 –
10625c 美術品 [ビジュッヒン] work of art 美術の作品 a work of art

図 4.23 “reference” の内容抽出結果

第5章

結論

本論文では，ネットワーク上やディジタル図書館などの大量文書データから利用者が適切な情報を入手するために必要な文書主題の抽出法を提案した．文書中の単語を単なる文字列の羅列として解釈するのではなく，その概念情報を用いることで単語の意味がもつ曖昧性を解消し，文書中で使用されている単語の意味を求めた．本手法はこの意味に基づいて文書の主題を抽出した点に新規性がある．単語の概念情報はEDR電子化辞書より取得した．概念情報から概念空間を構築し，単語がもつ意味のルーツを探索した．文書中の全ての単語に対して意味のルーツを照合することで概念空間を文書内容に沿って絞り込み，この結果を文書主題として抽出した．

第1章では研究のモチベーションを明らかにし，提案手法の概要と本論文の構成について述べた．

第2章ではディジタル図書館における情報検索を取り上げ，その問題から文書主題の抽出についての必要性と関連研究について述べた．

第3章では提案手法である概念に基づいた文書主題の抽出法について述べた．単語情報に加えて概念情報を用いることで文書の特徴を深層的な観点から抽出し，文書内容の表現についての充実を図った．文書の特徴は単語と概念の出現頻度分布から文書の強調度と表現度を求め，両者に基づいた特徴ベクトルを用いて定量的に表現した．提案手法の有効性を検証するために，技術論文の主題を概念で獲得する評価実験を行った．その結果，従来の文字列に着目した抽出法では単語の意味まで絞り込むことは困難であったが，単語の概念情報から筆者が論文中で用いた意味を推測することが出来た．また，アンケート質問“本手法で抽出した論文の主題が妥当であるかどうか”に対する調査の結果，76%の被験者が“excellent”，

"good" と評価しており，本手法で抽出した主題が被験者の想定する主題とほぼ一致していることがわかった．

第4章では提案手法の応用として，意味検索と文書構造に基づいた内容抽出の2例について述べ，それに伴う文書処理について記述した．意味検索では，検索質問を概念で表現する検索式部，科学技術論文の主題を選定する索引語作成部，および検索式と索引語を照合する照合部の三つから成る検索システムの試作について述べた．検索精度は，利用者が入力する検索質問の妥当性に大きく依存するため，検索式の導出において利用者の視点で検索式を選択できるよう支援した．また，文書構造に基づいた内容抽出では文書をその構造から六つに分割し，各特徴を抽出した．文書処理については，ディジタル図書館で保管されている日本語／英語の技術論文を入力として，文書構造解析，テキスト化，形態素解析，名詞の抽出について述べた．

今後の課題としては，知識情報に関するデータベースの充実が挙げられる．情報を効率よく処理する上で知識情報は欠かすことのできないものである．近年，オントロジーやエージェント技術を用いた知識情報の獲得に関する研究が盛んに行われているが，これらの知識は互いに整合性がなく，重複して蓄積されている．広範囲に渡って知識情報を獲得するためには，情報を統合して活用する技術が求められる．また，本論文では，提案した内容抽出法を意味検索と文書構造に基づいた主題抽出に適用したが，この他に文書要約やデータマイニング，情報フィルタリングへの適用とその整備が課題として挙げられる．

ディジタル図書館の歴史は浅く，稼働したばかりの赤ん坊のような存在である．多くの可能性を秘めているが，同時に，多くの問題も抱えている．本研究の取り組みは，この問題の1片を解決するための1手法にすぎないが，情報活用に関する問題の突破口的な存在となることを確信している．

謝辞

本研究を進めるにあたり，終始多大なる御指導を頂いた奈良先端科学技術大学院大学情報科学研究科千原國宏教授に厚く御礼申し上げます．研究者としての考え方，行動の仕方など多くのことを学ばせて頂きました．また，迷いが生じた時，暖かいお言葉を頂いたことは，著者にとって常に心の支えでした．重ねて御礼申し上げます．

情報科学研究科植村俊亮教授には，御厚情に満ちた御指導を頂き，また，研究活動，学生生活においても暖かいお言葉を頂きました．心より御礼申し上げます．

情報科学研究科横矢直和教授には，本研究をまとめるにあたり，数々の有益な御助言を頂き，また，はじめての学会発表では大変暖かく見守って頂きました．厚く御礼申し上げます．

本研究を遂行するにあたり，研究活動，学会発表，論文作成などに懇切丁寧な御指導を頂きました情報科学研究科今井正和助教授に厚く御礼申し上げます．

情報科学センター砂原秀樹助教授には，本研究を遂行するにあたり数々の有益な御助言を頂きました．また，研究活動，学生生活において暖かい励ましを頂きました．心より御礼申し上げます．

佐藤宏介助教授（大阪大学基礎工学研究科）には，直面した問題に対して適切な御助言を頂きました．これらは研究活動を行う上で非常に有益なものでありました．厚く御礼申し上げます．

本研究を遂行するにあたり，茶筅およびEDR電子化辞書を快く提供して頂いた情報科学研究科松本裕治教授に深く感謝致します．また，多くのアドバイスや暖かな励ましを頂いたことに心から御礼申し上げます．

先端科学技術研究調査センター大城理助教授には，日頃より，御指導，御助言を頂きました．厚く御礼申し上げます．

情報科学研究科眞鍋佳嗣助教授には，データ作成において有益な御教示を頂き，

また，研究活動においては暖かな励ましを頂きました．深く御礼申し上げます．

ミーティングなどにおいて，数多くの助言を頂きました情報科学研究科助手土居元紀先生，情報科学研究科助手黒田知宏先生，並びに像情報処理学講座の皆様に深く感謝致します．

最後にこの場を御借りして，著者を理解し，暖かく見守り続けてくださった両親と婚約者に厚く御礼申し上げます．

参考文献

[1] 住田一男，三池誠司，“知的情報検索の動向”，人工知能学会誌，Vol. 11，No. 1，pp.10–16，1996.

[2] Hearst，M. A. and Pederson，J. O，“Revealing collection structure through information access interfaces”，Proc. IJCAI-95，pp.2047–2048，1995.

[3] 前田晴美，糀谷和人，西田豊明，“情報ベースのユーザフレンドリなインタフェースのための連想構造の提案”，ヒューマン・インタフェース研究論文集，Vol. 5，No. 1，pp.49-56，1996.

[4] 武田英明，“ネットワークを利用した知的情報統合”，人工知能学会誌，Vol. 11，No. 5，pp.680-688，1996.

[5] 岩爪道昭，白神謙吾，細谷和右，武田英明，西田豊明，“オントロジーに基づく広域ネットワークからの情報収集・分類・統合化”，情報学論，Vol. 38，No. 3，pp.606-615，1997.

[6] 日本電子化辞書研究所，“EDR 電子化辞書技術ガイド（第2版再改訂）”，EDR，1995. [http://www.iijnet.or.jp/edr]

[7] 奈良先端科学技術大学院大学附属電子図書館編，“NAIST 電子図書館レポート'97”，奈良先端科学技術大学院大学，1997.
[http://dlw3.aist-nara.ac.jp/]

[8] 白尾 隆行，“世界情報インフラの整備およびその実験 －G 7 アプリケーション共同研究－ ”，情報管理，Vol. 38，No. 8，1995.

[9] 長尾真，“電子図書館”，岩波科学ライブラリー 15，岩波書店，1994.

[10] 宮井均，市山俊治，“電子図書館が見えてきた”，NECクリエイティブ，1999.

[11] 青空文庫．[http://www.aozora.gr.jp/]

[12] 国会図書館．[http://210.145.31.35/about/e_library.html]

[13] 次世代電子図書館システム研究開発事業.[http://www.dlib.jipdec.or.jp/]

[14] 堀井光彦，吉田哲三，“次世代電子図書館システム研究開発事業におけるプロトタイプシステムの概要”，ディジタル図書館（ISSN 1340-7287），No. 14，1999．[http://www.dl.ulis.ac.jp/DLjournal/No_14/]

[15] 奈良先端科学技術大学院大学.[http://dlw3.aist-nara.ac.jp/index-j.html]

[16] 今井正和　新麗，羽田久一，西村亨，砂原秀樹，千原國宏，“ある電子図書館の運用と統計”，ディジタル図書館（ISSN 1340-7287），No. 13，1998．[http://www.dl.ulis.ac.jp/DLjournal/No_13/]

[17] 東京大学ディジタルミュージアム．[http://www.um.u-tokyo.ac.jp/]

[18] 九州大学附属図書館．[http://www.lib.kyushu-u.ac.jp/index-j.html]

[19] 学術情報センター．[http://www.nacsis.ac.jp/els/els-j.html]

[20] 安達 淳，“学術情報センターの電子図書館システムの試行実験報告”，ディジタル図書館（ISSN 1340-7287），No. 3，1995.
[http://www.dl.ulis.ac.jp/DLjournal/No_3/]

[21] Project Gutenberg．[http://www.promo.net/pg/]

[22] NSF/ARPA/NASA Digital Libraries Initiative.
[http://dli.grainger.uiuc.edu/national.htm]

[23] Informedia．[http://www.informedia.cs.cmu.edu/]

[24] University of California at Berkeley Digital Library Project.
[http://elib.cs.berkeley.edu/]

[25] University of California at Santa Barbara (Alexandria Digital Library).
[http://alexandria.sdc.ucsb.edu/public-documents/]

[26] University of Michigan. [http://www.si.umich.edu/UMDL/]

[27] The WWW Virtual Library. [http://www.vlib.org/]

[28] Library of Congress. [http://www.loc.gov/]

[29] Vatican Exhibit Main Hall.
[http://metalab.unc.edu/expo/vatican.exhibit/exhibit/Main_Hall.html]

[30] National Library of Canada. [http://www.nlc-bnc.ca/]

[31] British Library. [http://www.bl.uk/]

[32] Bibliotheque National de France. [http://www.culture.fr/]

[33] 中国国家図書館（National Library of China）.
[http://www.lib.tsinghua.edu.cn/chinese/beitu/index.htm]

[34] 済賀宣昭，電子図書館の動向と学術情報流通.
[http://www.lib.ehime-u.ac.jp/user/saiga/saiga96a.html]

[35] 杉本重雄，“ディジタル図書館へのアプローチ”，ディジタル図書館（ISSN 1340-7287），No. 3, 1995. [http://www.dl.ulis.ac.jp/DLjournal/No_3/]

[36] Qiu, Y. H.P.Frei, “Concept Based Query Expansion”, In Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrival, pp.160-169, 1993.

[37] 松本裕治，須藤茂，中山拓也，平尾努，“複数の言語資源からのシソーラスの構築”，情処学基礎研報，FI, No. 42, pp.23–28, Jly. 1996.

[38] How to Create an Ontology.
[http://www-ksl-svc.stanford.edu:5915/doc/frame-editor/how-to-create-an-ontology.html]

[39] Jinxi Xu and W. Bruce Croft, “Query expansion using local and global document analysis”, 19th Annual Int ACM SIGIR Conf on R&D in Information Retrieval, pp.4–11, Aug. 1996.

[40] Harter, S. P, “A probabilistic approach to automatic keyword indexing. Part II: An algorithm for probabilistic indexing”, Journal of the American Society for Information Science, No. 26, pp.280-289, 1975.

[41] 石川徹也，“日本語テキストを対象とした自動索引システムの課題”，情報科学と技術，Vol. 42，No. 11，pp.994-1002，1992.

[42] 諸橋正幸，“自動索引付け研究の動向”，情報処理，Vol. 25，No. 9，pp.918-925，1984.

[43] 中川 裕志，情報検索入門.
[http://www.r.dl.itc.u-tokyo.ac.jp/ nakagawa/infoDB/syllabus.html]

[44] 全文検索システム協議会. [http://www.asahi-net.or.jp/ zc7t-urb/]

[45] P. イングベルセン，“情報検索研究”，トッパン，1995.

[46] 大高利夫，“情報検索の基礎”，情報科学技術協会，1995.

[47] David Ellis，“情報検索論”，丸善株式会社，1994.

[48] 高速文字列検索ライブラリ SUFARY Version2.1b2，1999.
[http://cl.aist-nara.ac.jp/lab/nlt/ss/]

[49] 諸橋正幸，堤泰治郎，丸山宏，野美山浩，“情報検索システムにおける効果的ナビゲーション機能の提案”，ディジタル図書館ワークショップ，No. 2，pp.45–49，Nov. 1994.

[50] Salton, G.，“The Vector Space Model, Automatic Text Processing – The Transformation, Analysis, and Retrieval of Information by Computer”，Addison-Wesley Publishing Company Inc. 1989.

[51] Doszkocs, T., Reggia, J. and Lin, X.，“Connectionist Models and Information Retrieval, Annaul Review of Information Science and Technology”，Vol. 25，pp.209-260，1990.

[52] Bartell, B. T., Cottrell, G. W. and Belew, R. K. “Representation Documents Using an Explict Model of Their Similarities”，American Society for Information Science，Vol. 46，No. 4，pp.254-271，1995.

[53] 武田英明，“知的情報検索の動向”，人工知能学会誌，Vol. 11，No. 1，pp.10-16，1996.1.

[54] Pollitt, S. "CANSEARCH : An Expert Systems Approach to Document Retrieval", Information Processing & Management, Vol. 23, No. 2, pp.119-138, 1987.

[55] Coroft, W. B. and Thompson, R. H. "I^3R : A New Approach to the Design of Document Retrieval Systems", American Society for Information Science, Vol. 38, No. 6, pp.389-404, 1987.

[56] Gorden, M. "Probabilistic and genetic algorithms for document retrieval", ACM Comm., Vol. 31, No. 10, pp.1208–1218, 1988.

[57] Ricchio, J. J. "Relevance Feedback in Information Retrieval, The Smart System – Experiment in Automatic Document Processing", Prentice-Hall Inc., pp.313-323, 1971.

[58] James Allean, "Relevance feedback with too much data", 18th ACM SIGIR Conference on Information Retrieval, pp.337–343, Jul. 1995.

[59] Rao, R., Pedersen, J. O., Hearst, M. A., Machinlay, J. D., Stuart, K. C., Masinter, L., Halvorsen, P. and Rovertson, G. G. "Rich interaction in the Digital Library", ACM Comm., Vol. 38, No. 4, pp.29–39, 1995.

[60] Lin, X., Soergel, D. and Marchionini, G. "A Selforganizing Semantic Map for Information Retrieval", 14th ACM SIGIR, pp.262-263, 1991.

[61] Hearst, M. A. and Pederson, J. O, "Revealing collection structure through information access interfaces", Proc. IJCAI-95, pp.2047–2048, 1995.

[62] Carpineto, C. and Romano, G, "A system for conceptual structuring and hybrid navigation of test databases", 1995 AAAI Fall Sump. on AI Applications in Knowledge Navigation and Retrieval, pp.20–25, 1995.

[63] ConceptBace, 株式会社ジャストシステム.
[http://www.justsystem.co.jp/cb/information/index_tech.html]

[64] VextSearch, コマツソフト.

[http://www.komatsusoft.co.jp/develp/vxtsc/vxtsc_f.html]

[65] 佐藤円, 佐藤理史, 篠田陽一, "電子ニュースのダイジェスト自動生成", 情処学論, Vol. 36, No. 10, pp.2371-2379, 1995.

[66] 松尾利行, 武田英明, 西田豊明, "KP 化による論文内容の効果的提示方法とその応用", 第11回ヒューマンインタフェースシンポジウム論文集, pp.581-588, 1995.

[67] 関根聡, "テキストからの情報抽出－文書から特定の情報を抜き出す－", 情報処理学会誌, Vol. 40, No. 4, pp.370-373, 1999.

[68] 奥村 学, 難波 英嗣, "テキスト自動要約に関する研究動向", 自然言語処理,「テキスト要約のための言語処理」特集号, Vol. 6, No. 6, 1999.

[69] Salton, G. Allen, J. and Buckley, C, "Automatic structuring and retrieval of large text files", Communications of the ACM, No. 37, pp.94-108, Feb. 1994.

[70] Luhn, H., The automatic creation of literature abstracts IBM Journal of Research and Development, Vol. 2, No. 2, pp.159–165, 1958.

[71] 若尾孝博, 江原暉将, 白井克彦, "テレビニュース番組の字幕に見られる要約の手法", 情報処理学会自然言語処理研究会報告, pp.83–89, 1997.

[72] Edmundson, H, "New methods in automatic abstracting" Journal of ACM, Vol. 16, No. 2, pp.264–285, 1969.

[73] Brandow, R., Mitze, K., Rau, L, "Automatic Condensation of Electronic Publications by Sentence Selection", Information Processing and Management, Vol. 31, No. 5, pp.675–685, 1995.

[74] 亀田雅之, "日本語文書読解支援系 QJR の検討", 情報処理学会自然言語処理研究会報告, pp.57–64, 1996.

[75] 仲尾由雄, "見出しを利用した新聞・レポートからのダイジェスト情報の抽出", 情報処理学会自然言語処理研究会報告, pp.121–128, 1997.

[76] 福本淳一, "文の結合度に基づく内容抽出手法", 言語処理学会第3回年次大会発表論文集, pp.321–324, 1997.

[77] Salton, G., Singhal, A., Buckley, C., Mitra, M, "Automatic Text Decomposition Using Text Segments and Text Themes", In Proc. of the 7th ACM Conference on Hypertext, pp.53-65, 1996.

[78] 松本裕治, 北内啓, 山下達雄, 今一修, 今村友明, "日本語形態素解析システム ChaSen 使用説明書 version1.0", 奈良先端科学技術大学院大学, 1997. [NAIST Technical Report 97007]

[79] 山下達雄, "MOZ と LimaTK の説明書", LimaTK パッケージ付属, 1999. [http://cl.aist-nara.ac.jp/ tatuo-y/ma/]

[80] 後藤英昭, 阿曽弘具, "文字行の局所的な曲線を利用した頑健・高速な文字行抽出法ディジタル", 電子情報通信学会論文誌 Vol. J78-D-II, No. 3, pp.465-473.

研究業績

[学術論文(査読付)]

1. 堀井千夏，今井正和，千原國宏："ディジタル図書館のための概念情報を用いた科学技術論文の検索"，電子情報通信学会論文誌 Vol. J82-D-I, No. 10, pp.1245-1255, 1999年10月 (本論文第3章4章に関連).

[国際会議(査読付)]

2. Masakazu IMAI, Chinatsu HORII, Hisakazu HADA, Naokazu YOKOYA and Kunihiro CHIHARA : "Design of a Digital University Library : Mandala Library", Proceedings of International Symposium on Digital Libraries 1995, pp.119-124, Tsukuba, Japan, August, 1995.

3. Masakazu IMAI, Syouhachi KAWATA, Chinatsu HORII, Hisakazu HADA, Naokazu YOKOYA and Kunihiro CHIHARA : "Construction of Mandala Library", Proceedings of 2nd Joint Workshop on Multimedia Communication, pp.8-1-1 -pp.8-1-8, Seika, Kyoto, Japan, October, 1995.

4. Chinatsu HORII, Masakazu IMAI and Takeshi UNO : "Tracking and Collision Avoidance of Mobile Robot with Vision", Proceedings of 1996 Japan-U. S. A. Symposium on Flexible Automation, pp.577-580, Boston, USA, July, 1996.

5. Chinatsu HORII, Masakazu IMAI, and Kunihiro CHIHARA : "An Information Retrieval using Conceptual Index Term for Technical Paper on Digital Library", Proceedings of International Symposium on Research, Develop-

ment & Practice in Digital Libraries 1997 : ISDL '97, pp.205-208, Tsukuba, November, 1997(本論文第 3 章に関連).

6. Chinatsu HORII, Masakazu IMAI and Kunihiro CHIHARA : "Conceptual Information Retrieval of Technical Papers for Digital Libraries", Proceedings of IEEE Advances in Digital Libraries Conference '99, pp. 171-178, Baltimore, Maryland, USA, May, 1999(本論文第 3 章 4 章に関連).

[国内発表(査読付)]

7. 堀井千夏, 今井正和, 千原國宏:"概念情報を用いたディジタル図書館のための情報検索手法", 2000 年情報学シンポジウム –ネットワーク型情報メディアの活用と情報を活かす新技術–, 東京, 2000 年 1 月 (本論文第 3 章 4 章に関連).

[研究会・学会]

8. 堀井千夏, 今井正和, 烏野武:"視覚を持ったロボットによる追跡と衝突回避", 電子情報通信学会パターン認識と理解研究会報告, PRU95-25, 名古屋, 1995 年 5 月.

9. 今井正和, 羽田久一, 堀井千夏, 山口英, 佐藤宏介, 竹村治雄, 横矢直和, 千原國宏, 嵩忠雄:"曼陀羅図書館の構築の試み", 電子情報通信学会パターン認識と理解研究会報告, PRU95-32, 名古屋, 1995 年 5 月.

10. 堀井千夏, 今井正和, 烏野武:"視覚を持ったロボットによる追跡と衝突回避", 日本機械学会ロボティクス・メカトロニクス講演会'95, Vol. B, pp.1102-1105, 1995 年 6 月.

11. 堀井千夏, 今井正和, 千原國宏, 嵩忠雄:"ディジタル図書館のための概念情報を用いた科学技術論文の検索手法", 電子情報通信学会総合大会, SD-3-2(pp.7-445 /7-446) , 東京, 1996 年 3 月 (本論文第 3 章に関連).

12. 堀井千夏, 今井正和, 千原國宏, 嵩忠雄:"ディジタル図書館のための概念情報を用いた科学技術論文の検索手法", 第 39 回自動制御連合講演会, 3055(pp.381-382), 奈良, 1996 年 10 月 (本論文第 3 章に関連).

13. Chinatsu HORII, Masakazu IMAI, Kunihiro CHIHARA and Tadao KASAMI, "An Information Retrieval using Conceptual Index Term for Technical Paper on Digital Library", 2nd NAIST Symposium on Digital Libraries, Dec 4, 1996(本論文第3章4章に関連).

14. 堀井千夏, 今井正和, 千原國宏, 嵩忠雄:"ディジタル図書館のための概念情報を用いた科学技術論文の検索手法", 第9回ディジタル図書館ワークショップ, pp.85-92, 1997年3月 (本論文第3章4章に関連).

付録

A. 論文画像からの文書構造解析

文書構造に基づいた論文の分割には，論文画像のレイアウト解析を参照する．論文の文字行部分を抽出するには，後藤らによる区分直線連結法 [80] を用いる．これは，2 値画像ファイルを対象に図 A.1に示すような基本矩形を l ドットごとに作成し，その矩形内で連続する黒画素領域を抽出する手法である．l ドットには経験的に求めた (最大文字サイズ) $\times$ 0.6 の値を用いる．この文字行に基づいて文書構造を解析する．ただし，図や式中にある文字行は解析のノイズとし，間隔や文字行の位置が文書構造的に不自然なものを対象に除去する．

以下に文書構造の解析処理の流れを示す．

(1) 文書画像の入力を 2 値文書画像（BMP 形式）にする．
(2) 文書画像上に座標系をとり，y 座標について l ドット毎に矩形を作成する (図 A.1).
(3) 矩形内で連続する黒画素領域を抽出する．
(4) 矩形内の中心点を x 軸方向へ写像した分布のピーク付近を文書行として抽出する．
(5) 文字行の先頭と末尾の座標を求める．
(6) 隣接する行の位置関係から構造解析ノイズを除去する．

上記の処理に従って文書構造の解析を行った結果を図 A.2～A.13に示す．図 A.2 に示すような表紙画像から文字を抽出した結果を図 A.3に，文字行の抽出結果を図 A.4に示す．さらに，解析ノイズを取り除いた結果を図 A.5に示す．また，式を含む頁（図 A.6）と図を含む頁（図 A.10）についても同様な結果を示す．

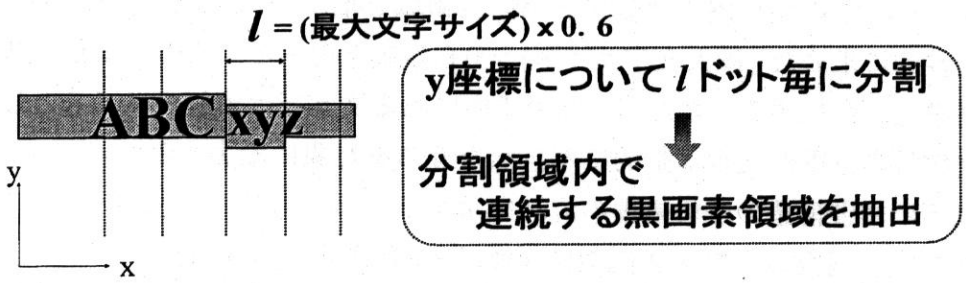
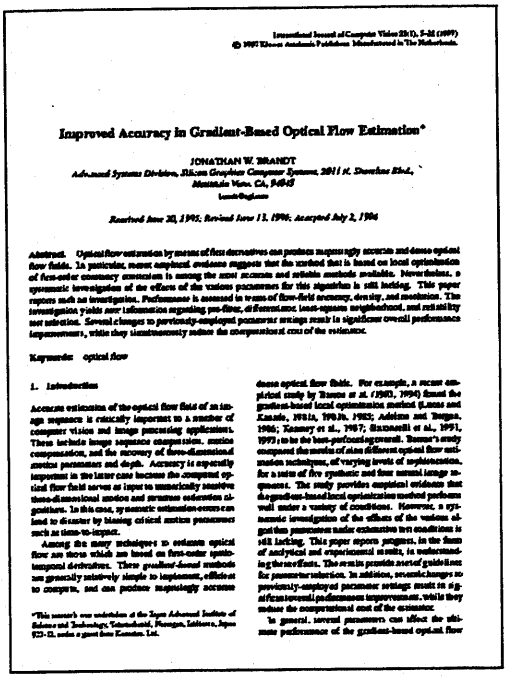


図 A.1 文字抽出のための分割矩形



Improved Accuracy in Gradient-Based Optical Flow Estimation*

JONATHAN W. BRANDT

Abstract. Optical flow estimation by means of first derivatives can produce surprisingly accurate and dense optical flow fields. In particular, recent empirical evidence suggests that the method that is based on local optimization of first-order constancy constraints is among the most accurate and reliable methods available. Nevertheless, a systematic investigation of the effects of the various parameters for this algorithm is still lacking. This paper reports such an investigation. Performance is assessed in terms of flow-field accuracy, density, and resolution. The investigation yields new information regarding pre-filter, differentiator, least-squares neighborhood, and reliability test selection. Several changes to previously-employed parameter settings result in significant overall performance improvements, while they simultaneously reduce the computational cost of the estimator.

Keywords: optical flow

1. Introduction

Accurate estimation of the optical flow field of an image sequence is critically important to a number of computer vision and image processing applications. These include image sequence compression, motion compensation, and the recovery of three-dimensional motion parameters and depth. Accuracy is especially important in the latter case because the computed optical flow field serves as input to numerically sensitive three-dimensional motion and structure estimation algorithms. In this case, systematic estimation errors can lead to disaster by biasing critical motion parameters such as time-to-impact.

Among the many techniques to estimate optical flow are those which are based on first-order spatio-temporal derivatives. These *gradient-based* methods are generally relatively simple to implement, efficient to compute, and can produce surprisingly accurate dense optical flow fields. For example, a recent empirical study by Barron *et al.* (1993, 1994) found the gradient-based local optimization method (Lucas and Kanade, 1981a, 1981b, 1982; Adelson and Bergen, 1986; Kearney *et al.*, 1987; Simoncelli *et al.*, 1991, 1993) to be the best-performing approach. Barron's study compared the results of nine different optical flow estimation techniques, of varying levels of sophistication, for a suite of five synthetic and four natural image sequences. The study provides empirical evidence that the gradient-based local optimization method performs well under a variety of conditions. However, a systematic investigation of the effects of the various algorithm parameters under estimation test conditions is still lacking. This paper reports progress, in the form of analytical and experimental results, in understanding the methods. The results provide a set of guidelines for parameter selection. In addition, systematic changes to previously-employed parameter settings result in significant overall performance improvements, while they reduce the computational cost of the estimator.

In general, several parameters can affect the ultimate performance of the gradient-based optical flow

*This research was undertaken at the Toyo Advanced Institute of Science and Technology, Tatsunokuchi, Ponegun, Ishikawa, Japan 923-12, under a grant from Komatsu Ltd.

図 A.2 表紙の論文画像

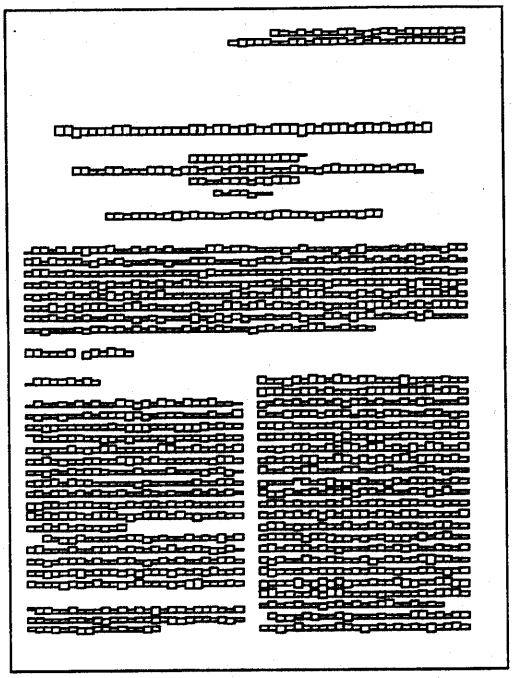


図 A.3 表紙の論文画像から文字を抽出

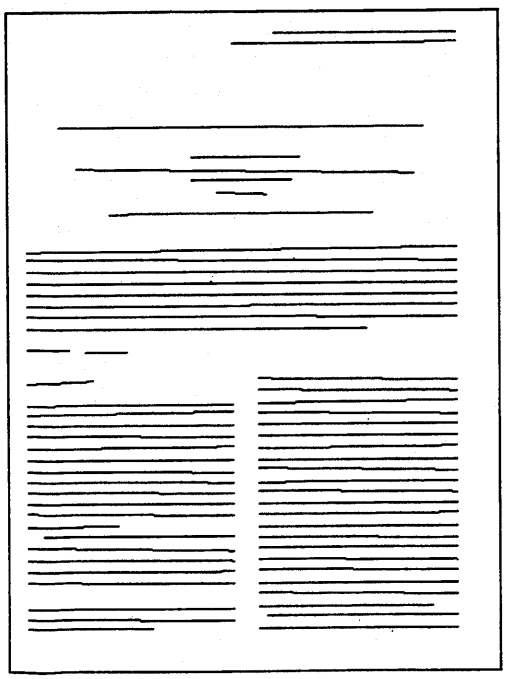


図 A.4 表紙の論文画像から文字行を抽出

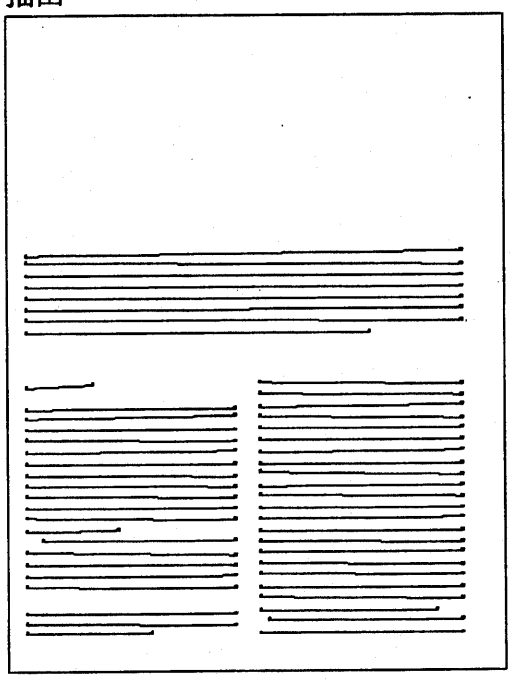


図 A.5 図 A.4 からノイズを除去

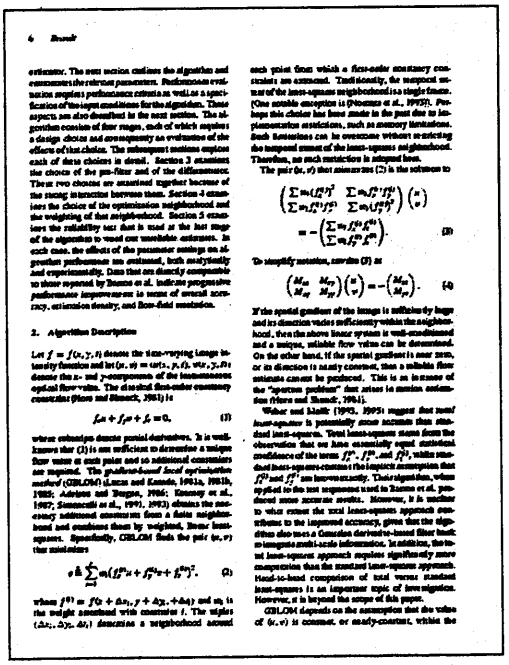


図 A.6 式を含む論文画像

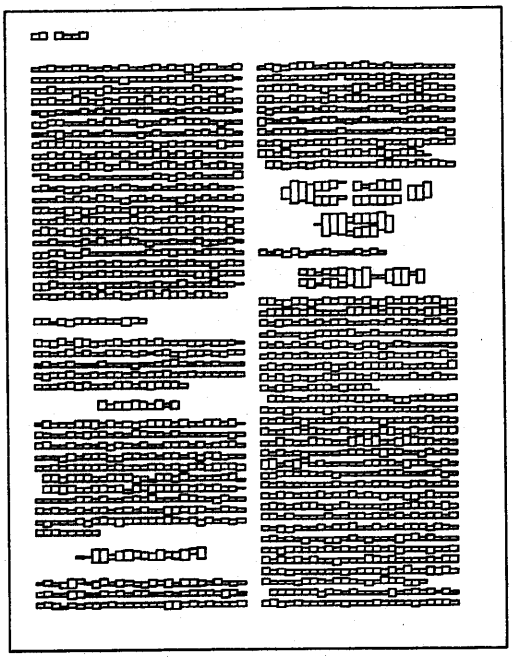


図 A.7 式を含む論文画像から文字を抽出

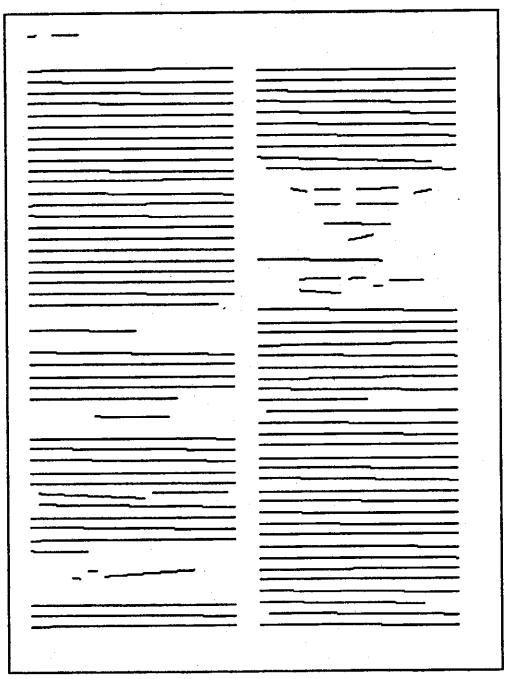


図 A.8 式を含む論文画像から文字行を抽出

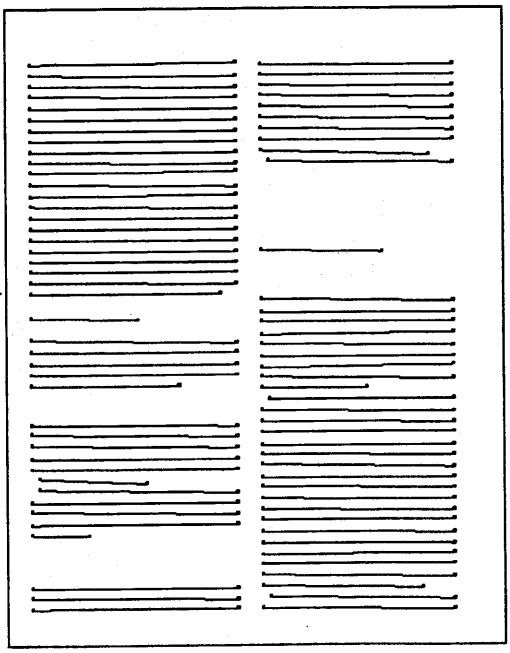


図 A.9 図 A.8 からノイズを除去

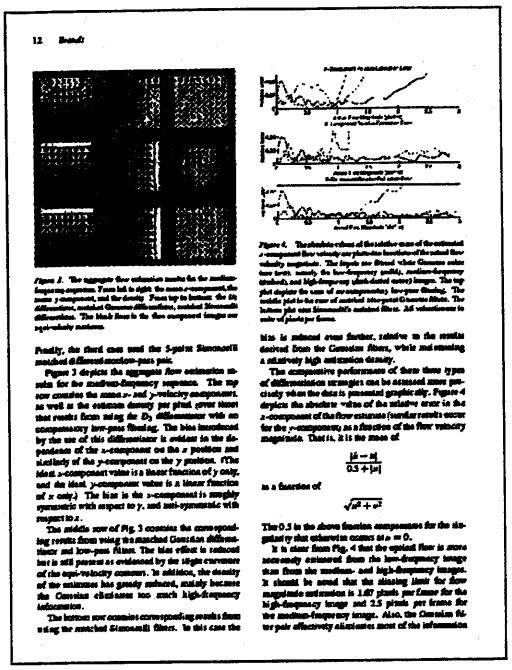


図 A.10 図を含む論文画像

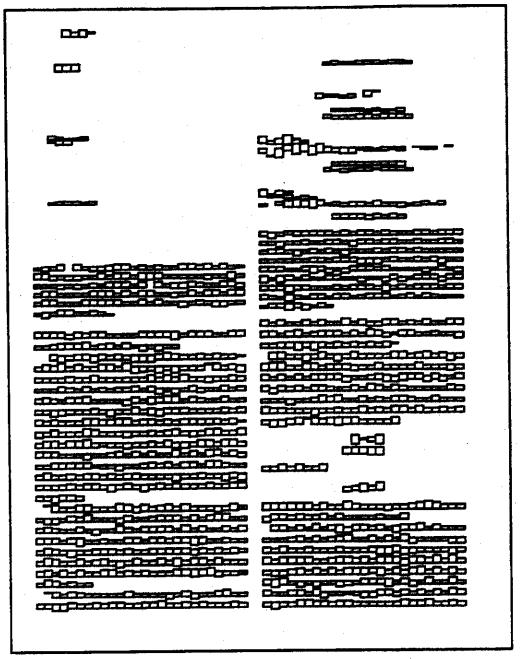


図 A.11 図を含む論文画像から文字を抽出

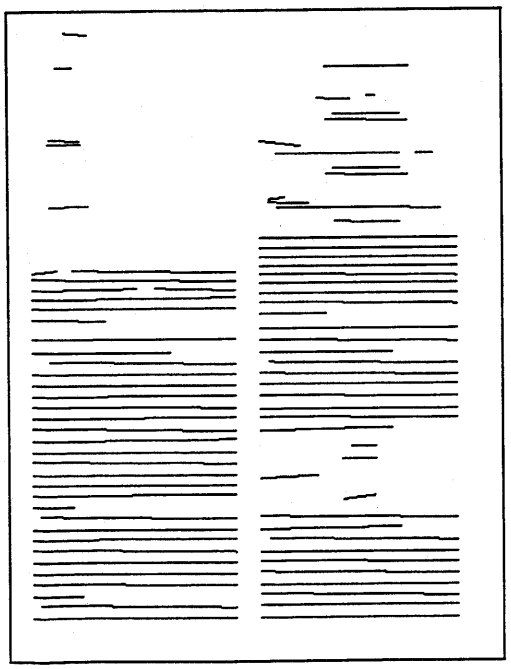


図 A.12 図を含む論文画像から文字行を抽出

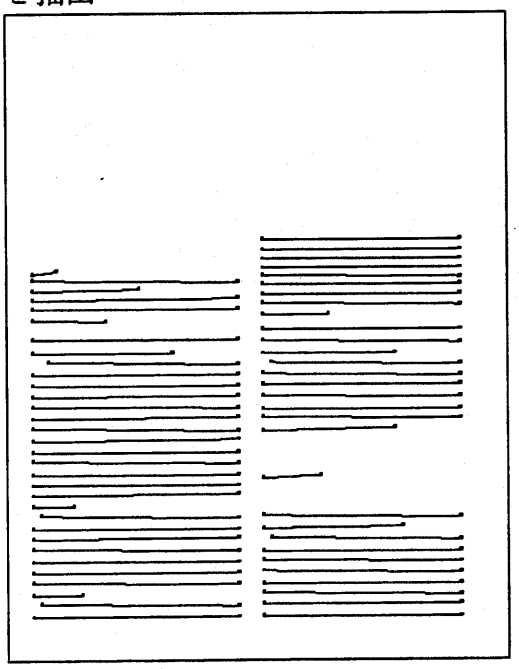


図 A.13 図 A.12 からノイズを除去