2.4 概念を用いた学術文献の検索

2.4.1 はじめに

本学のような大学附属図書館の役割は、所属する研究者が研究活動を行う上で必要になる学術情報を研究者が利用しやすい形で迅速に提供することである。現在のディジタルライブラリシステムでは、利用者が莫大な学術情報から望むものを探し出すための手法として全文検索を基本とした検索方式を提供してきている。しかし、ディジタルライブラリが保有する学術情報は増大する一方であり、決して減少することはない。このため現在有効に機能していると思われる全文検索にも、次のような欠点がある。

- ユーザから示された文字列が含まれているかを検索しているため、不適切な検索語に対する検索結果が膨大なものになる。

この問題は、現在の検索技術では一致する文字列が存在するか否かを調べているために発生するものであり、本質的には全文検索の欠点ではない。しかし、全文検索では不適切な検索語に対する結果の増加の割合は通常の検索の場合に比べてはるかに大きくなり、利用者にとって検索結果を評価することが全く不可能になる可能性が非常に高い。さらに、文字列の一致による検索では、検索語の同意語は当然のことながら検索結果に含まれない。これは利用者の立場から見れば望ましいことではない。図書館利用者が司書と対面で検索を行う場合には、司書との会話の中でこの問題を解決することができる。しかし、ディジタルライブラリでは生身の人間である司書との対話は行われずに、計算機と利用者との対話で解決する必要がある。このために、これまでにも様々な試みがなされている。米国議会図書館（Library of Congress）では、文献から二次情報を作成する際に用いる単語を制限し、限られた語彙（control words）で文献の二次情報を作成する方法を提案している。これは図書館情報学の分野ではよく知られた方法である。これにより、文献に関するデータベースに出現する語彙が限られたものとなるため、検索の際の曖昧性を抑制することができる。他にも、検索の際にシソーラス（同義語辞書）を併用し、検索語と同じ意味を持つ言葉を検索語の候補として提示し、利用者に対する支援を行う方法も考案されている[1)]。さらに、単語の持つ意味についての関係を意味ネットワークとして表現し、これをシソーラス辞書と同じように用いることも提案されている[2)]。また最近では、「概念検索」を実現する商用システム[3)]も見られるようになってきた。このシステムでは単語の生起確率をニューラルネットワークで学習し、擬似的に「概念検索」を実現しているが、真の意味で単語あるいは言語の概念をとり扱ったものではない。これに似た考え方はUniversity of Illinois[4)]でも研究されている。

我々は、このような検索における問題点を解決するために、自然言語処理的な方法

を用い、検索語、検索対象の論文からそれらが表現しようとしている概念を抽出し、検索に用いることを考え、研究を行った。ここでは、学術論文は研究により得られた知見を伝えるために作成されるものと考える。このため、学術論文は他の一般的な文章（例えば小説）とは異なり、明確に読者に伝えたい事実あるいは考え方、すなわち概念があり、しかも主題（伝えたい概念）は少ないものであると仮定する。また、単語の概念を記述した辞書が利用可能であるとする。実際、自然言語処理の分野では、単語の概念を記述した辞書[5]が存在している。この辞書では日本語と英語の単語について、その概念が記述されているだけでなく、概念間の関係も記述され、しかも日本語と英語で共通の概念を用いている。このため、日本語のキーワードを用いて英語学術論文の検索や、英語キーワードによる日本語文献の検索といった言語を横断した検索を実現することができる可能性を秘めている。

以上の考え方のもとに、日本語で記述された学術論文を対象とした概念検索の研究をこれまで行っている。以下にこれまでの研究の進捗状況と今後の研究方針、計画を示す。なお、これまでの研究成果の詳細については発表された文献[6]、[7]を参照されたい。

2.4.2 基本的な考え方

単語の概念を用いた検索を行うためには、検索語からの概念の抽出と検索対象となる学術論文から主題となる概念を抽出する必要がある。検索語から概念を抽出することは、概念辞書を検索して概念を求めることであり、特に難しいものではない。しかし、学術論文には多数の単語が含まれており、これらから概念を抽出する必要がある。以下、日本語で記述された学術論文を対象として議論を進める。図 2.4.1 に、学術論文が表現しようとしている概念を抽出する処理の大まかな流れを示す。また、ディジタルライブラリでの応用を前提とするため、学術論文のテキストデータはスキャンした論文の画像から光学的文字読取り（OCR）処理を用いて論文の画像から抽出されたものを仮定する。OCR 処理では、画像すなわち論文から文字情報を誤りなく抽出する

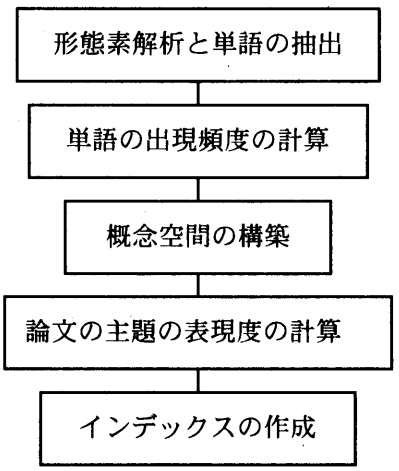

図 2.4.1　処理の流れ

ことは不可能であり、必ず認識誤りを含む。本学のディジタルライブラリでは OCR 結果を用いた全文検索を実施しているため、実用性を考慮して OCR 結果を用いることとした。

2.4.3 形態素解析と単語の抽出

学術論文では、名詞、動詞、形容詞などいくつかの品詞が出現している。学術論文はなんらかの事実、考え方を伝えるために作成されていることを考え、名詞のみを学術論文から抽出し、概念を求めることとする。英語では単語ごとにスペースが挿入されており、単語の分かち書きがなされている。これに対して日本語では、単語の区切りは文章中に存在しない。このため、文章から単語を抽出する必要がある。これは自然言語処理の分野では形態素解析と呼ばれ、単語の品詞解析も同時に行われる。ここでは、その研究成果である「茶筅」を用いることにした [5]。「茶筅」は本学情報科学研究科松本裕治教授の研究グループが開発した形態素解析ソフトウェアであり、日本語文章から単語の抽出および品詞解析を行うもので、インターネット上で一定の条件のもとでソースコードが配付されている。

2.4.4 単語の出現頻度の計算

先の処理で学術論文に含まれる名詞の抽出ができた。こののち、抽出された単語から概念を求め、最終的には学術論文が表現しようとしている概念を求める。学術論文では、伝えたい概念を表現する単語がくり返し用いられていると仮定する。実際、何か伝えたいことがあれば、それを表す単語が多数出現している。そこで、論文が伝えようとしている主題を表現する概念を示す単語は、その論文中に多数出現すると考える。この考えのもとに、次の２つの出現頻度分布を指標として用いて論文の特徴抽出を行う。

- 出現単語の文字列についての出現頻度分布
- 出現単語の概念についての出現頻度分布

文字列についての出現頻度分布から「単語の出現の重み」を論文の主題の主張度として求め、概念の出現頻度分布から「単語の概念の重み」を論文の主題が表す内容の表現度として求める。今回対象としている学術論文、特に科学技術論文のように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索引語を求めると、対象とする分野で頻繁に使用され、かつ重要である単語を索引語から取り除いてしまう。この問題を避けるために、論文ごとに表現度を明らかにし、主題に対する主張度から主題を論じている強さを求め、これらを用いて論文の特徴ベクトルの特徴量を決定する。

出現単語の文字列についての出現頻度分布は、おおまかには次のようにして求める。処理対象とする論文データベースに含まれる論文ごとに、重複しない単語の出現頻度をその論文でもっとも頻繁に出現する単語の出現頻度で正規化して求める。また、出現単語の概念についての出現頻度分布を求めるには概念空間の構築が必要となるが、それについては次節で述べる。

2.4.5 概念空間の構築

2.4.5.1 EDRにおける概念辞書

今回単語から概念を抽出するためにEDRと呼ばれる電子化辞書を用いた。EDRは11のサブ辞書から構成され、ここではそのうちの日本語単語辞書、概念体系辞書、概念見出し辞書の3つを用いた。日本語単語辞書では単語と概念記述子の関係が、概念体系辞書では概念記述子間の上位・下位関係が、概念見出し辞書では各概念記述子が示す概念の見出しと説明がそれぞれ記述されている。この3つの辞書を利用することにより、単語から概念の抽出を行って概念空間を構成し、これを用いることによって学術論文が伝えたい主題となる概念を抽出する。概念見出し辞書は、抽出した概念（ここでは概念識別子で記述されている）を人間が分かる形で提示するために利用される。また、EDRのサブ辞書には英語単語辞書も存在し、用いられている概念識別子は日本語単語辞書のそれと同一なものであるので、言語間での概念の比較を行うことも可能である。

EDRの概念体系辞書では「概念」を最上位の概念とし、その下位概念である基本概念を5つ定義している。これらは「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」である。概念は上位－下位関係を階層構造として体系化されている。単語はこの階層構造の中で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。さらに多重継承が存在し、1つの概念が2つ以上の上位概念を持つ場合があるため正確には木構造にはならない。大まかには次のような手順で概念空間を構築する。

- 日本語単語辞書を用いて単語から対応する概念識別子を求め、これを概念Aとする。
- 概念体系辞書を用いて概念Aの上位概念を全て求める。これを概念Bとする。
- 概念Bのさらに上位概念を求めるために、概念Bを概念Aと置き換えて前項の操作を繰り返す。

例えば、日本語の「ベース」という単語について上記の操作を行うと図2.4.2に示すような概念の階層構造が求められる。図の例では3回の繰り返しを行って上位概念を求めている。この繰り返しの数は単語の抽象度により最適な数が変化し、実際に用

いる際には経験的に決定するが、おおむね３回から５回程度で十分である。この繰り返しの回数が多すぎると、単語によっては最上位の概念に到達してしまったり、必要以上に取得される概念が抽象化されてしまう。最上位の概念は先にも述べたように非常に抽象的であるので、検索を行うときに意味をなさなくなる。逆に繰り返しの回数が少なすぎると、抽出した概念が具象的になってしまい、ある程度は抽象的であるべき論文の主題を反映しなくなってしまう。

後に、文書中に出現した単語が共通に示す概念を発見する際、部分概念空間を用いる。これは図 2.4.2 に示した概念空間の中で、節を根とする階層構造である。これはある単語が示す一つの概念を反映していると考えることができる。

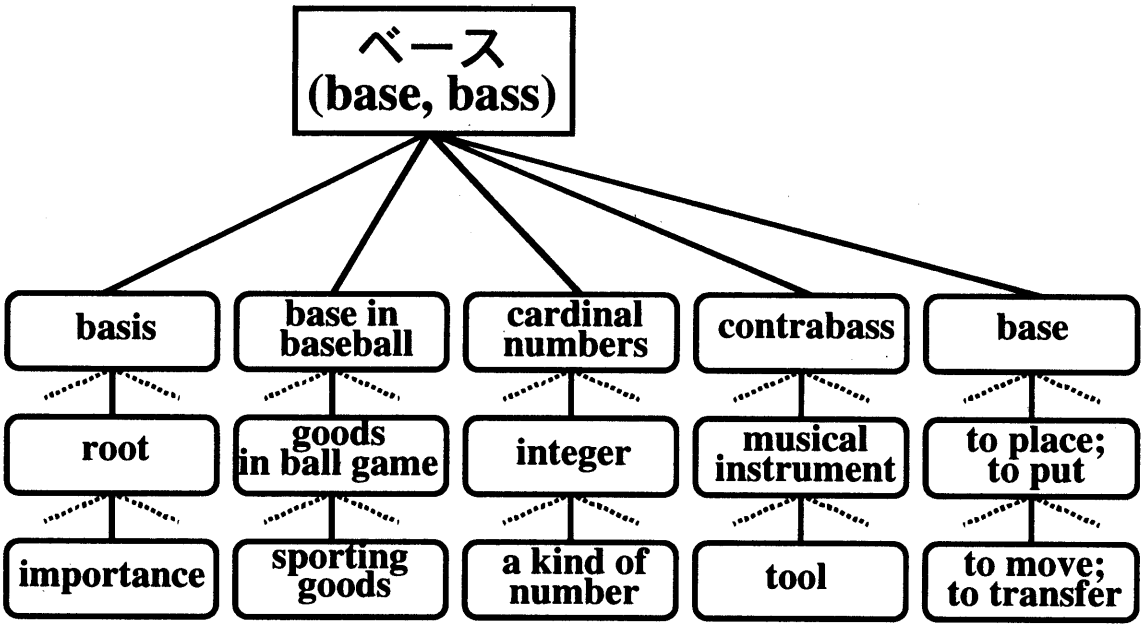

図 2.4.2　単語「ベース」の概念空

2.4.5.2　概念空間の照合と部分概念空間の出現頻度

一般に、一つの単語が示す概念は複数存在する。このため、ある伝えたい概念を表現するために、同じ概念を持つ言葉をいくつか用いて文章を作成する。そこで、文書に含まれる単語から抽出した概念空間を照合することによって、文書が伝えようとしている概念を推定する。このとき、照合の単位として先ほど述べた部分概念空間を用いる。そして、どの概念を強調しているかを調べるために部分概念空間の出現頻度を調べる。

概念空間の照合とは、文書から重複して出現する単語を取り除き、各単語の概念空間に存在する部分概念空間で共通して存在する部分概念空間を見つけだすことである。大まかには以下のような手順で照合を行う。同時に部分概念空間の出現頻度も求める。

- 文書から重複語を取り除いた単語それぞれについて概念空間を構築する。
- 求めた概念空間に含まれる部分概念空間をすべて求める。
- 各部分概念空間の集合において、共通要素となる部分概念空間を抽出する。これが照合により絞り込まれた部分概念空間である。
- 絞り込まれた部分概念空間を要素とする部分概念空間の集合の数を出現頻度とする。

2.4.5.3 連結数と概念パス

概念空間は概念を節とした木構造のような構造をなしており、連結関係を含んだ、節に相当する概念（正確には概念記述子）によって構成されている。これは、ある概念を根とした木構造の集合として表現することができる。また部分概念空間は単語の概念空間の真部分集合であり、概念空間の場合と同様に概念を根とした木構造の集合として表現される。ここで、概念空間、部分概念空間（これらは集合である）を構成する要素として、概念パスを定義する。概念パスはある概念を根として、そこから外部節（子を持たない節）までの連結を示すパスとする。そして、概念パスの根から外部節までの相対レベル（根から外部節までの節の数）を連結数とする。この連結数が少ない概念パスは、内容が抽象的な表現になり過ぎていると考え、ここでは有効性が低いものとして考慮しない。

2.4.5.4 文書内容の表現度

文書内容の表現度として、文書に出現する単語の部分概念空間の出現頻度を用いる。文書から重複語を取り除いた単語について、単語の上位概念の検索を再帰的に繰り返して求められる概念空間から構築された部分概念空間の表現度を次のように計算する。文書から重複語を取り除いた単語から構築された部分概念空間の出現頻度を、文書内で最も多く出現した部分空間の出現頻度で正規化する。このとき上位の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になる。

2.4.6 論文の主題の表現度の計算とインデックス

検索に用いることを前提とした文書データベースを構築する際には、文書の主題についての内容を抽出するだけでなく、主題の主張の強さについても考慮する必要がある。これを示すために、文書の特徴量を、出現単語の主張度の和に文書内容の表現度を加えたものとする。この特徴量を索引語として概念を用いた検索を実現する。

2.4.7 実装

これまでに提案した手法を実装するためには、検索式を概念で表現する検索式部、

学術論文の主題を選定する索引語作成部、検索式と索引語を照合する照合部からなる。以下、それぞれについて簡単に述べる。

2.4.7.1 検索式部

検索の制度を向上するためには、検索語として入力された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る検索意図と検索質問の違いを推測しなければならない。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ことによって、この問題を解決する。検索式を概念で獲得する手順は以下の通りである。

- 利用者が入力した検索語を概念記述子に変換する。
- 各概念記述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。
- 概念空間を絞り込み、利用者の要求にそった概念パスを獲得する。
- 絞り込まれた概念パスとその内容を利用者に提示する。

提示した概念パスが妥当なものであるかどうかを利用者に確認したのち、最終的な検索式とする。

2.4.7.2 学術論文からの概念の獲得

検索時の適合性を向上させるためには、文書の主題となる検索語を正確に抽出する必要がある。以下の手順で論文からの主題を獲得する。

- OCR 結果としての論文テキストを形態素解析し、単語の抽出とその文字列の出現頻度を求める。
- 各単語を概念記述子に変換し、概念空間を構築する。
- 各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。
- 単語の文字列の出現頻度と概念の出現頻度の分布から論文の特徴量を求める。
- 論文の主題を概念パスの集合として保存する。

2.4.7.3 概念レベルでの照合

概念パスで表現された検索式を概念パスの集合として表現されている学術論文と照合し、その結果を利用者に提示する。

2.4.8 評価

提案した手法がどの程度正しく論文の主題を抽出しているかを調べるため、システムが抽出した論文の主題の評価を行った。情報科学の知識をもつ５人に、システムで

抽出した論文の主題を提示し、システムが抽出した概念がどの程度正しく論文の主題を表現しているかの評価を行った。その結果、おおむね正しく論文の主題を抽出していることが確認された。正しく主題を表現していないと判定された論文には次のような特徴があった。

1. 形態素解析に用いた「茶筅」の辞書や概念抽出の際に用いた EDR の辞書には存在しない学術用語が使用され、それが論文の主題に密接な関係があった。
2. 形態素解析の際に、不適切な単語分解が行われた。(「３次元」が「３＋次元」に分解されたなど)

最初の例は、学術用語を含んだ辞書を用いることにより解決することができる。ただ、学術用語は学問の進歩とともに新しい用語が創造される性格があるため、新しい学術用語を自動的に抽出し、辞書に登録するメカニズムが求められる。２つめの例では、形態素解析の際に単語の区切り方を指定できるので、これを最適なものとすることが必要である。

さらに、今回提案した手法を用いた検索システムを WWW から行えるように実装した。図 2.4.3 にその検索語入力、図 2.4.4 に検索語から抽出された概念パスを利用者に提示している様子を示す。図 2.4.5 に検索結果を、図 2.4.6 に論文を閲覧している画面を示す。

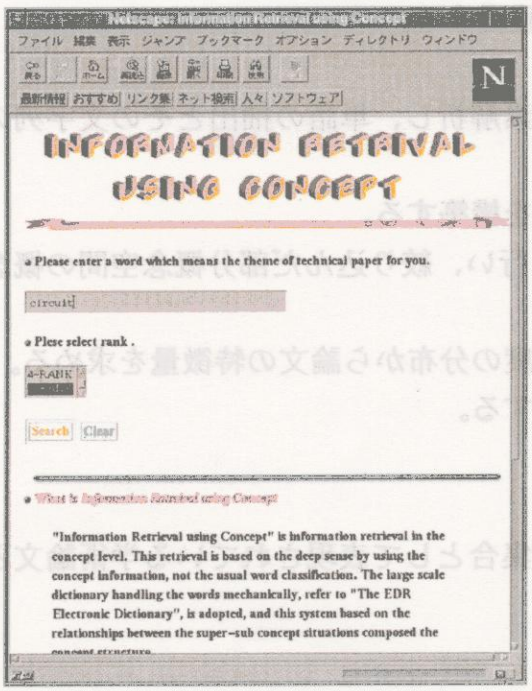

図 2.4.3　試作システムの検索語入力画面

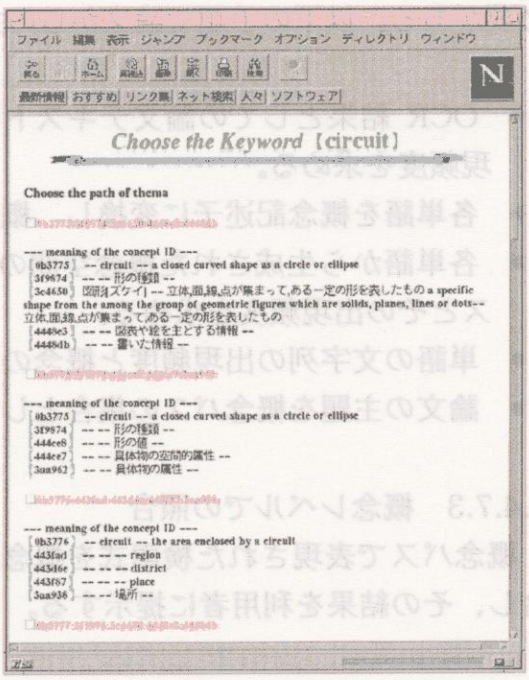

図 2.4.4　試作システムの検索語の概念表示画面

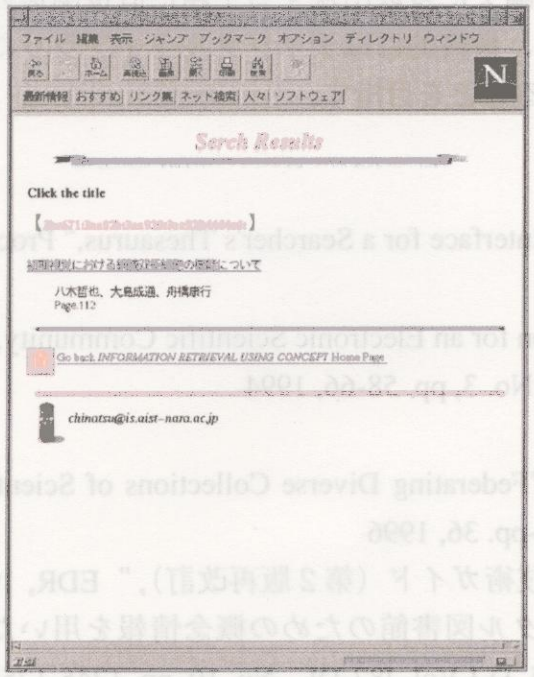

図 2.4.5　試作システムによる検索結果

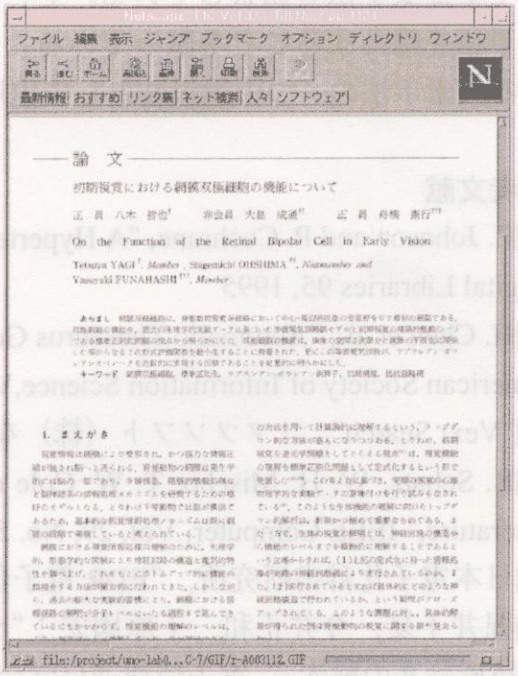

図 2.4.6　試作システムによる論文の閲覧

2.4.9　これまでの問題点と今後の計画

これまでに、日本語で記述された学術論文を対象にして概念の抽出とそれを用いた検索方法の提案と検証を行ってきた。ここで提案する手法は、EDR という電子化辞書を用いることによって文字列としての単語を取り扱うのではなく、単語が示す概念を用いて論文の主題といった意味内容を取り扱う。この特徴は、単語から概念を求めることが可能であれば、論文を記述している言語には関係なく検索を実現可能であるところにある。もちろん、単語から概念を求めるには、適当な電子化辞書の存在が必要不可欠となるが、幸いなことに本研究で用いた EDR は日本語だけでなく、英語についても単語の概念を記述した辞書が用意されている。さらに、EDR の中で概念は日本語と英語で共通のものが用いられているので、言語間での比較が可能である。これは現在問題となっている多言語間での検索を可能にする。そのためには、単語の概念を記述した電子化辞書の整備が前提となるが、これは自然言語処理分野でも必要なものであり、潜在的な需要が高いと考えられ、時間とともに整備されていくものと考える。

本研究での問題点をいくつかあげると次のようになる。

1. 現在の処理対象が日本語だけである。
2. 処理手法の提案に留まっている。
3. 論文には主題とは関係ないが多数出現する単語が存在し、論文の主題抽出に悪影響を与えている。
4. 論文の構造を利用した概念の抽出がなされていない。

今後はこれらの問題点を順次解決していく予定である。現在、英語の学術論文について本提案手法を拡張している。そこでは、論文に多数出現する主題に直接関係のない単語を除外することを試みている。そして、最終的にはディジタルライブラリにおける検索手法の一つの選択肢として実用化することを目指す。

参考文献

[1]E. Johnson and P. Cochrane, "A Hypertextural Interface for a Searcher's Thesaurus," Proc. of Digital Libraries 95, 1995

[2]H. Chen et al., "Automatic Thesaurus Generation for an Electronic Scientific Community," J. American Society of Information Science,Vol. 46, No. 3, pp. 58-66, 1994

[3]"Vext Search", コマツソフト（株）など

[4]B. Schatz, W. H. Mischo, T. W. Cole et. al., "Federating Diverse Collections of Scientific Literature," IEEE Computer, Vol. 29, No. 5,pp. 28-pp. 36, 1996

[5]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

[6]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 D-I,Vol. J82-DI，No. 10, pp. 1245-1255, 1999 年 10 月

[7]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.