情報メディア 1－4
（1991. 5. 24）

専門用語辞典のハイパーテキストシステム

黒橋 禎夫　長尾 眞　佐藤 理史　村上 雅彦
京都大学工学部 電気工学第二教室

要旨

本稿では、専門用語辞典の自動的ハイパーテキスト化の方法について述べる．大量のテキストデータをハイパーテキストに変換する場合には，計算機による自動的リンク付けが必要となる．そこで，重要な用語が辞典の中のどの場所で説明されているかという情報，および用語間の意味関係の情報をリンクとすることを考え，辞典文章中の定義文をとり出して解析することによってこれらの情報を自動抽出した．さらに，抽出した情報を利用して辞典のハイパーテキストシステムを作成した．このシステムは，リンクをスムーズにたどることができるユーザインターフェイスを備え，また辞典内の情報に様々な方法でアクセスすることのできる検索機能を備えている．

Hypertext System of an Encyclopedic Dictionary of a Specific Field

Sadao Kurohashi　Matoko Nagao　Satoshi Sato　Masahiko Murakami
Department of Electrical Engineering, Kyoto University

Abstract

This paper describes a method of automatic hypertext construction from an Encyclopedic Dictionary of Computer Science, which was originally published in a book form. By checking varieties of sentential styles, sentences defining the meaning of head words and other important words are first extracted from the dictionary. Then the keywords and their relations are extracted automatically from these sentences. These extracted terms are connected each other according to the extracted relations. The whole text data of 2 mega characters were stored in a workstation. We constructed a user-friendly interface on X-window system, which realizes varieties of accesses to required information.

1 はじめに

ワープロやパソコンの普及，出版過程の電子化，光ディスクなどの大容量記憶装置の進歩などによって世の中には大量のオンライン・テキストデータが溢れている．今後は，これらの膨大な情報を管理し，その中から必要な情報だけを効率良くとり出すことのできるテキスト検索システムが必要となってくる．このようなテキスト検索システムでは，一次情報をどのように加工・管理するか，どのようなユーザインターフェイスを備えるかといったことが問題となる．このうちデータの管理について，近年，ハイパーテキスト[1]と呼ばれる手法が注目されている．これは，テキストを意味上まとまりのある小部分(これをノードと呼ぶ)に分割し，それらの間をリンクにより関係づけたネットワーク構造によって情報を管理する技術，またはその状態をいう．

現在の計算機の能力にとっては，このようなネットワーク構造のデータを管理することはさほど問題ではない．問題となるのは，データの加工，すなわち既存のテキストデータをノードに分解しノード間に関連性や類似性を認めてリンクを張るというハイパーテキストへの変換である．特に大量のテキストデータを扱う場合にはこれをすべて人手で行なうわけにはいかないので，計算機による自動リンク付けを考える必要がある．

本稿では，見出し語数約4,500，テキスト全体で約200万文字の岩波情報科学辞典[2]を題材とし，これを自動的にハイパーテキスト化する方法を示す．また作成したハイパーテキストシステムを紹介する．

2 ハイパーテキスト化のための言語処理

2.1 リンクの種類

辞典のハイパーテキストを考える場合には，ノードとしては「項目」，すなわち一つの見出し語とその説明文を考えることが自然であり，リンクとしてはまず各項目の見出し語のもつ概念間の関係性を扱うものを考える．そのほかに，説明文中に現れる重要な用語に対しても関連情報との間にリンク付けをする必要があるなど，種々の関係のリンクを設定することを考えねばならない．そこで，本研究では項目に対応する項目ノードだけでなく，岩波情報科学辞典の索引にとられている索引語(見出し語と，辞典テキスト中に現れる見出し語でない重要語をあわせた約13,000語)に対応する用語ノードを考え，それらの間に次の2種類のリンクを張ったネットワーク構造を考えた(図1)．

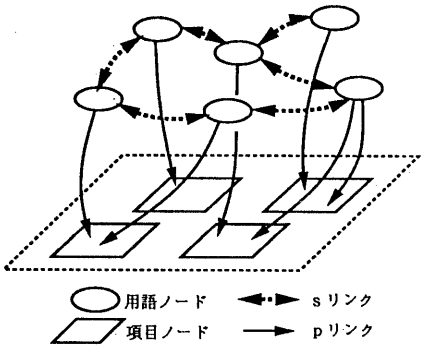

図1: 辞典のハイパーテキスト構造

pリンク: 用語ノードから項目ノードへのリンク．見出し語の用語ノードからはその見出し語の項目ノードへのリンクが張られる．見出し語でない索引語の用語ノードからはその用語の説明文を含む項目ノードへのリンクが張られる．この場合は，その項目内の何文目で説明されているかという情報がリンクに付加される．

sリンク: 用語ノード間のリンクで，用語の意味関係を示す．このリンクには意味関係の種類(上位下位関係，同義語関係など)を示すラベルが付けられている．

辞典に対してこのようなネットワーク構造を考えることにすると，自動ハイパーテキスト化の問題は，1) pリンクとして辞典の中で索引語が説明されている位置，2) sリンクとして索引語間の意味関係，を自動抽出する問題に置き換わる．本研究では以下のような方法によってこれらの抽出を行なった．

2.2 定義文の抽出

まず予備的処理として，辞典のテキストデータをタグによって機械的に項目単位に分割し，さらに日本語解析システムHAPPINESS[3]を用いて分かち書き文に変換した．こうして得られたテキストデータから，次の3種類の定義文を区別して抽出することにした(ここでは複文における節も文と呼ぶことにする)．

同義語文 同義語，略語を示す文．

内包的定義文 本質的内容・性質を説明する文．

外延的定義文 具体例を列挙する文．

岩波情報科学辞典では，次の例に示すように，これらの文がこの順に現れる形で見出し語に対する定義が与えられている．

作図装置　plotter

ブロッターともいう．線画や文字を用紙に描くための図形出力装置をいう．ペンを上下させながら表示すべき線の上を移動するベクトル走査方式で線を描くペンプロッターと，ペン以外の記録方式で上から下へ順次走査するラスター走査方式で線を描くラスタープロッターに分けられる．...

これは，人間が辞典の説明をできるだけ正確に理解するためにとられた表現形式であるが，計算機によって辞典データを処理する場合にも望ましいことである．これらの定義文の言語表現はある種の特徴を持っているので，これらをよく調べパターン化した．これを表1に示す．次にこれらのパターンとの文字列照合によって項目説明文章の中から定義文をとり出すことを行なった．この定義の部分は見出し語を省略した表現が多いが，その省略を前提として処理することにより見出し語について種々の情報を抽出することができた．

岩波情報科学辞典の各項目の説明は見出し語に直接関係することだけでなく，関連する様々な用語についての記述を含んでいる．これらの用語についての定義文も同じ方法で説明文中から抽出した．

2.3　同義語文の解析

同義語文は表1のパターン (a-1), (a-2) によって抽出されるもので，典型的なものは次のような文である．

「A は，B または C ともよばれる．」

ここで A を被定義語，B，C を同義語と呼ぶことにする．パターン照合で抽出されるのは下線部分であるが，その末尾から順に分かち書きされている語をとり出して処理を進める．まず，末尾から「，」，「または」，「あるいは」などでつながる語をとり出し同義語とする[1]．次に，パターン (a-1) の文の場合は「は」または「を」，パターン (a-2) の場合は「は」があるかを調べ，ある場合にはその直前の語を被定義語とする．ない場合には，その文が項目説明の最初の2文目までのものであれば，「は」や「を」で示される語が省略されていて，それは見出し語であると考え，その項目の見出し語を被定義語とする．

この解析の結果から，

- 同義語と同義語文の間に p リンクを張る．
- 被定義語と同義語の間に同義語関係の s リンクを張る．

2.4　内包的定義文の解析

内包的定義文は表1のパターン (b-1) ～ (b-6) によって抽出されるものである．まず，各パターンの文から次のように被定義語 (次に示すパターン中の A) と定義部分 (次に示すパターン中の～) を抽出する．このうち (b-4), (b-5), (b-6) のパターンでは，被定義語がない場合に，その文が各項目の同義語文を除いた最初の2文までであればその項目の見出し語を被定義語とする．

(b-1),(b-2):　「～は A という．」の形の文である．末尾の語 A を被定義語とし，(b-1) の場合は「は」または「を」，(b-2) の場合は「は」から前の部分を定義部分とする．

(b-3):　「～は A であるという．」の形の文である．末尾の語 A を被定義語とする[2]．

(b-4):　「A とは～をいう．」の形の文である．「とは」または「は」の直前の語を被定義語とし，そこから後の部分を定義部分とする．

(b-5):　「～が A である．」と「A とは～である．」の形の文がある．末尾から2番目の語が「が」である場合は前者であるとして末尾の語を被定義語，「が」から前の部分を定義部分とする．そうでない場合は後者であるとして (b-4) と同様の処理を行なう．

(b-6):　体言止めの文については，項目の始めで見出し語が被定義語となる場合のみを扱う．

さらに，これらの処理により抽出した定義部分の末尾の語を被定義語の上位概念を示す上位語としてとり出す．

これらの解析結果から，

- 被定義語がその項目の見出し語でない場合，被定義語と内包的定義文の間に p リンクを張る．
- 被定義語と上位語の間に s リンクを張る．

2.5　用語のカテゴリー分類

ここまでの同義語文と内包的定義文の解析では，文の表層表現だけを手がかりにして用語間の関係などを抽出してきた．しかし外延的定義文を解析したり内包的定義文から上位下位関係以外の意味関係を抽出しようとする場合にはこのような方法では限界がある．そこで，これ

[1] 実際にはとり出した語が索引語である場合のみ同義語とする．このように最終的に索引語のみを扱うのは，被定義語の抽出や，次節以降の処理でも同様である．

[2] このパターンの文は～の部分の性質を定義しているので，他のパターンに対して行なうような定義部分に関する処理は行なわず，A からこの定義文へ p リンクを張ることだけを行なう．

表 1: 定義文を示すパターン

同義語文のパターン	
(a-1)	「と｛いう｜よぶ｜書く｜訳す｝こと も」「とも｛いい｜いう｜よび｜よぶ｜書き｜書く｜訳し｜訳す｝」 「｛と｜とも｝｛略記｜略称｝｛し｜する｝」「｛と｜とも｝略｛し｜す｝」
(a-2)	「と｛いわれる｜よばれる｜書かれる｜訳される｝こと も」 「とも｛いわれ｜いわれる｜よばれ｜よばれる｜書かれ｜書かれる｜訳され｜訳される｝」 「｛と｜とも｝｛略｜略記｜略称｝｛され｜される｝」「の｛略｜略記｜略称｝」

内包的定義文のパターン	
(b-1)	「と｛いい｜いう｜よび｜よぶ｝」
(b-2)	「と｛いわれ｜いわれる｜よばれ｜よばれる｝」
(b-3)	「であると｛いい｜いう｝」
(b-4)	「の｛一つ｜一種｜一形態｜一形式｜一方式｜一分野｜一分科｜一方法｜一手法｜一部門｝」 「の｛呼称｜名称｜総称｜意｜こと｝」「を｛いい｜いう｜指し｜指す｝」「を 意味｛し｜する｝」
(b-5)	「で｛，｜あり，｜あるが，｜ある．｝」
(b-6)	「(体言止め)」

外延的定義文のパターン	
(c-1)	「に｛分けられ｜分けられる｜分れ｜分れる｝」「に｛分類｜大別｝｛し｜する｜され｜される｜でき｜できる｝」
(c-2)	「が｛あり，｜あるが，｜ある．｝」

注) ｜はORの意味で，{}内のいずれの語でもよいことを意味する．(b-5),(c-2)のパターンでは，パターンに照合する文であっても，その末尾が「必要」，「特徴」などの一般的な性質を示す語である文の場合は例外的にとり出さない．

までの抽出結果を利用することにより索引語を半自動的にカテゴリー分類し，以降はこのカテゴリー情報を利用してより詳細な処理を進める．

情報科学の専門用語に対して，

[抽象的実体] 例) 情報科学，巡回セールスマン問題
[性質] 例) 曖昧性，雑音指数
[機能・行為] 例) 構文解析，1次変換
[具体的実体] 例) 通信回路，RAM
[人物・機関] 例) マッカーシー，情報処理学会

の5つのカテゴリーを設定し，以下の手順で索引語12,396語にこれらのカテゴリーを付与した．

1. 複合語である索引語の末尾の語基(たとえば「フーリエ変換」の「変換」)と，内包的定義文から抽出した上位語のうち索引語でない語を集め，その中で頻度の高い463語を人手でカテゴリー分類した．
2. 1の処理でカテゴリー分類された語を末尾に含む索引語や，これを上位語とする索引語に同一のカテゴリーを自動的に付与する．この処理によって8,196語にカテゴリーを付与することができた．
3. 同義語関係にある用語は同じカテゴリーに属し，下位語は上位語のカテゴリーを継承すると考えることができる．そこで，自動抽出した同義語関係，上位下位関係のデータを利用して2の処理で付与したカテゴリー情報を伝搬させた．まず，同義語関係と内包的定義文から抽出した上位下位関係のデータを用いることにより9,208語に，最終的には，外延的定義文から抽出した上位下位関係を加えたデータを用いることにより9,351語にカテゴリーを付与することができた．

2.6 外延的定義文の解析

外延的定義文は表1のパターン(c-1), (c-2)によって抽出される次のような文である．

(c-1): 「A は，～B と～C に分けられる．」
(c-2): 「A には，～B や～C がある．」

ここでAを被定義語，B，Cをその下位語と呼ぶことにする．まず「，」，「と」，「や」などの前の語を下位語の候補とする．次に，パターン(c-1)の文の場合は「は」，パターン(c-2)の場合は「としては」，「として」，「には」，「は」があるかを調べ，ある場合には，その直前の語を被定義語とする．ない場合には，抽出した文の直前の文(節である場合もある)が同義語文

または内包的定義文として抽出・解析されており，そこから被定義語が抽出されていれば，前の文の被定義語をその外延的定義文の被定義語とする．このようにしてとり出した被定義語と下位語の候補の間に次のいずれの関係も成り立たない場合に，その候補を実際に下位語と決定する．

1. 下位語の候補と被定義語のカテゴリーが異なる．
2. 内包的定義文の解析により下位語の候補と被定義語の間に逆の上位下位関係が抽出されている．
3. 下位語の候補が被定義語の末尾の部分文字列となっている．

1の処理によって「自然言語には曖昧性がある．」のような定義文から，また 2，3の処理によって，「幾何歪みには，～に伴う歪みや，～がある．」のように句によって下位概念が示されている定義文から誤った上位下位関係を抽出することを防ぐことができる．

これらの解析結果から，

- 下位語が辞典の見出し語でない場合に，その下位語にとってその外延的定義文が重要な位置であると考え，下位語と外延的定義文の間にｐリンクを張る．
- 被定義語と下位語の間にｓリンクを張る．

2.7 内包的定義文からの用語間の意味関係の抽出

内包的定義文では，被定義語と重要な意味関係を持つ用語によって被定義語の内容・性質が説明されているが，これは被定義語と関連する用語との意味関係を示しているととらえることもできる．そこで先に抽出した被定義語とその定義部分のデータから用語間の意味関係を自動抽出してｓリンクを張ることを考えた．

用語間の意味関係としては様々なものが考えられるが[4] [5]，ここでは情報科学分野の専門用語であることを考慮し順序関係，因果関係，論理的関係などよりも具体的な関係として表2に示す関係を考えることにした．

これらの意味関係とその関係を示す表層表現との対応付けを行い，定義部分中にそのような表現が現れた場合にその直前の用語と被定義語の間に対応付けておいた意味関係があることを抽出する．たとえば，「最尤スペクトル推定」の定義部分「音声波形のスペクトル包絡を求める方法」の中の「を求める」という表現から，

「最尤スペクトル推定」 <deal> 「スペクトル包絡」

という関係を抽出する．

また，ある一つの表現が複数の異なった関係を表すのに使われる場合には，用語のカテゴリー情報を用いることによってそれらを区別する．たとえば，「完全ハッシュ関数」の定義部分「ハッシュ法において，全く衝突が起こらないようなハッシュ関数」からは「において」という表現によって，「完全ハッシュ関数がハッシュ法において扱われている」ことがわかるから，

「完全ハッシュ関数」 <by> 「ハッシュ法」

という関係を抽出する．しかし，「可塑性」の定義部分「脳において神経回路網がその結合を自動的に変える可塑的能力」では同じ「において」という表現が使われているが，「可塑性」のカテゴリーが[性質]であることから「可塑性が脳において扱われている」というよりはむしろ「脳が可塑性という性質をもつ」と考え，

「脳」 <hprop> 「可塑性」

という関係を抽出する．

これらの処理によって得られた関係はｓリンクとして登録する．

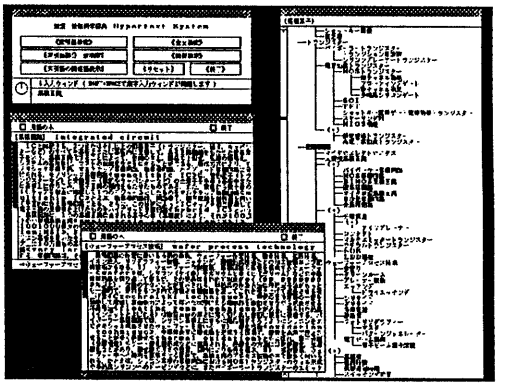

図 2: システムのディスプレイ・イメージ

3 専門用語辞典のハイパーテキストシステムの作成

前章の処理結果を用いてワークステーション上で利用することができる岩波情報科学辞典のハイパーテキストシステムを作成した (図 2)．

3.1 辞典データの構造

すでに 2.1 節で述べたように，システム内での辞典のデータ構造は約 4,500 の項目ノードと約 13,000 の索引語に対応する用語ノード，およびそれらの間のリンクからなるネットワーク構造である (図 1)．リンクとしては，２章で説明した処理によって付与されたｓリンク，ｐリンクに加えて，辞典データから機械的にとり出した次のような情報を用いる．

表 2: 用語間の意味関係

関係	例
A ＜isa＞ B：AがBの上位概念である	「論理回路」 ＜isa＞ 「順序回路」
A <syn> B：AとBが同義語である	「大規模集積回路」 <syn> 「LSI」
A <anti> B：AとBが反義語である	「アップロード」 <anti> 「ダウンロード」
A<hcomp>B：Aの構成要素としてBがある	「ハミルトン・グラフ」<hcomp>「ハミルトン閉路」
A<hprop>B：AがBという性質を持つ	「分散システム」<hprop>「ネットワーク透明性」
A<hfuni>B：Aが内部機能としてBを持つ	「ダイナミックRAM」<hfuni>「リフレッシュ」
A<hfuno>B：Aが外部機能としてBを持つ	「電子ビーム露光装置」<hfuno>「電子ビーム描画」
A <deal> B：AがBを処理する，扱う	「間順走査」 <deal> 「二分木」
A <purp> B：Aの目的がBである	「Ｌｉｓｐ」 <purp> 「記号処理」
A <used> B：AがBで利用，応用される	「差分」 <used> 「ガウスの補間公式」
A <set> B：AがBを発明，設定した	「ＩＳＯ」 <set> 「ＩＳＯ規格」
A <by> B：AがBという状況で扱われる	「ベクトル量子化」 <by> 「信号空間」

- 岩波情報科学辞典には，見出し語相互間の概念関係を一つの木構造で表わした用語の木が示されている．この用語の木での関係をｓリンクとする (ただし用語の木では用語間の意味関係は示されてない).
- 各項目にはそれに関連する参照語が挙げられている．これらも見出し語間の関連性としてｓリンクとする.
- 岩波情報科学辞典では巻末の KWIC 索引によって各索引語の説明されている項目を参照することができるようになっている．このデータをｐリンクとする.

3.2 リンクによる検索

システムはユーザインターフェイスを通してリンクをたどるための機能を提供している．各リンクに対してそれぞれ次のようなインターフェイスを備えている.

- ｐリンク: リンクによる検索あるいは高次検索の結果，ある索引語が選ばれると，システムはその索引語の用語ノードから張られているｐリンクを自動的にたどる．ｐリンクが一つの場合は，そのリンクの先の項目ノードの内容をウィンドウを開いて表示する[3] (これを項目ウィンドウと呼ぶことにする)．ｐリンクが複数の場合は，それらのリンクの先の項目名 (見出し語名) とその索引語が説明されている文の一覧を表示する (図3)．ここで項目名をマウスで指定するとその項目の項目ウィンドウが表示される.

[3] この時，項目の中でその索引語が説明されている文を反転表示する．索引語が見出し語でありその見出し語の項目を表示する場合は反転表示はない.

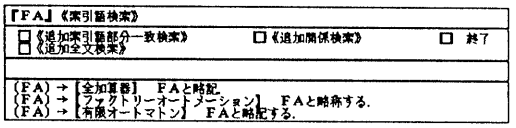

図 3: 「FA」に対する索引語検索

- 自動抽出したｓリンク: 索引語を入力ウィンドウに与えてメニューの「索引語の関連語表示」を選択することにより，その索引語とｓリンクでつながっている索引語およびその間の意味関係の一覧を表示する．この一覧からマウスで索引語を選択すると，上で述べたようにシステムがその索引語の用語ノードからのｐリンクを自動的にたどる.
- 用語の木のｓリンク: 用語の木はその一部が常に一枚のウィンドウで表示されており，その表示は項目ウィンドウの「用語の木」のボタンによってその項目の見出し語を含むように切り替わる．ある見出し語から用語の木のｓリンクをたどることは，用語の木の表示ウィンドウでその見出し語の周りの用語を調べ，適当な用語をマウスで選択することによって行なわれる.
- 参照語のｓリンク: 見出し語の参照語は項目ウィンドウで説明文とは別に表示されており，これらをマウスで指定することにより参照語のｓリンクをたどることができる.

3.3 高次検索

テキストデータに対してどれだけ詳細なリンクを張っておいても，ユーザはリンクによる検索だけで満足する

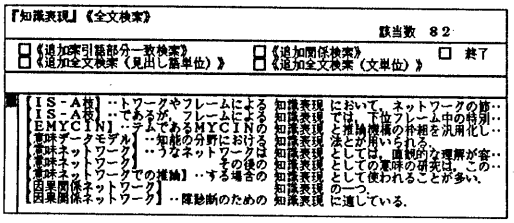

図 4: 「知識表現」に対する全文検索

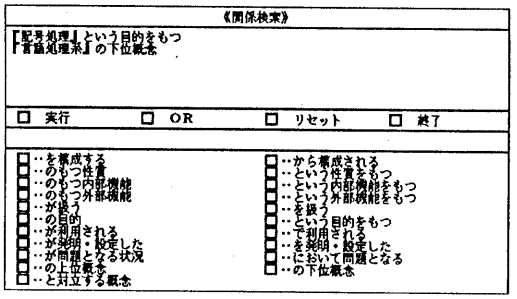

図 5: 関係検索での検索条件を指定するウィンドウ

ことはできない．そのために以下の 4 つの高次の検索法を用意した．

- 索引語検索: ユーザの指定したキーワードと索引語との照合を行なう．キーワードが索引語であれば，先に述べたようにその索引語からの p リンクを自動的にたどって適切な表示を行なう．キーワードが索引語でない場合は，自動的に索引語との部分一致検索を行ないそのキーワードを含む索引語の一覧を示す．
- 全文検索: ユーザの指定したキーワードと辞典の全データとの文字列照合を行ない，そのキーワードを含むすべての文を KWIC 形式になおしその文のある項目名とともに表示する (図 4).
- 関係検索: 索引語 A から関係 <R> (A <R> X あるいは X <R> A というという条件. <R> は表 2の関係) によって得られる索引語を，A と <R> を与える[4] ことによってシステムに推論させる．すなわち，条件を満たす索引語を s リンクのデータから求める．このタイプのいくつかの条件を論理和，論理積の形で重ねることができる．たとえば，「(「記号処理」という目的をもつ) AND (「言語処理系」の下位概念)」という検索条件から「Lisp」という検索結果が得られる．

[4] 条件の指定は図 5のウィンドウでマウスによって行なえる

表 3: 自動抽出による p リンクの位置と辞典に与えられた索引位置の比較

	自動抽出した p リンクの数	辞典の索引位置と一致する数
同義語文からの抽出	1335	1271 (95%)
内包的定義文からの抽出	3411	2710 (79%)
外延的定義文からの抽出	1163	829 (71%)
合計	5909	4810 (81%)

- 追加検索: 一旦得られた検索の結果に対して，さらに追加検索を行なうことができる．これは，索引語検索，全文検索，関係検索の任意の組合せで何度でも行なうことができる．

3.4 システムの規模と能力

システムは SPARCserver390(メインメモリー 56MB) 上で作成した．処理速度，移植性などを考慮して，すべて C 言語で記述し，ユーザのフロントエンドとして X ウィンドウシステムを用いた．

辞典のテキストデータは約 4MB であり，全体を一つのファイルとしそのファイルへのポインターを管理している．それ以外のインデックスデータは約 700KB であり，すべてメモリー上で管理している．

メモリー上のデータに対する検索である索引語検索，関係検索，およびリンクによる検索は約 1 〜 2 秒で行なえる．また，ファイルからデータを読み出す必要のある全文検索は約 5 〜 6 秒で行なうことができる[5]．全文検索も含めて現在の検索処理時間はユーザにとってほとんど問題のない時間であるといえる．

4 考察

前章で述べたハイパーテキストシステムの作成という試用の経験から，2 章で示した自動抽出の精度とその方法の有用性，および 3 章で示したような辞典のハイパーテキストシステムの有用性は以下の通りである．

まず自動抽出した p リンクと辞典に与えられていた索引とを比較した結果を表 3に示す．辞典の索引からの参照場所は，見出し語を参照しているものを除くと合計 11,814ヶ所であり，これらは各索引語に関する重要な位

[5] 文字列照合のアルゴリズムとしては Boyer-Moore アルゴリズムを用いているが，処理時間のほとんどがディスクアクセスの時間であるため，このアルゴリズムを用いることによる処理時間の短縮はほとんどみられない．

表 4: 自動抽出したsリンク

関係	抽出数	関係	抽出数	関係	抽出数
<$\overline{\text{isa}}$>	3362	<hprop>	26	<purp>	141
<syn>	1562	<hfuni>	13	<used>	605
<anti>	57	<hfuno>	116	<set>	10
<hcomp>	101	<deal>	590	<by>	750
合計					7333

置を人手でとり出したものである．自動抽出の結果はこの辞典の索引の参照場所とかなり一致しており，抽出した場所の重要度という点ではこの抽出方法の精度はかなり高いといえる．索引で参照されていない場所で自動抽出されたものについても重要なものが多く，これらは人手による索引付け作業で落ちたものを補っている．

sリンクの自動抽出の結果を表4に示す．このうち，同義語関係，上位下位関係を示すリンクについては意味的に誤りであるようなリンクはほとんどない．しかし上位下位関係については，ある語の上位概念を示す語として概念的に一つ上ではなく数段上のレベルの用語がとり出されていることが多い．これは，辞典の定義文において直接の上位語ではなくさらに上のレベルのより広い概念の用語が用いられているためである．

同義語関係，上位下位関係以外の意味関係のsリンクは内包的定義文から抽出したものであるが，このうち無作為に400組を選び人手でその関係の妥当性を調べた．その結果，関係付けが誤りであると判断したものは34組であり，92%は妥当な関係であった．誤りのほとんどは，句としての意味が重要であるにも関わらずその中の一単語との間に関係付けを行なったものである．

最後に専門用語辞典をこのように計算機上のハイパーテキストシステムとして実現することの有用性を考察する．そこではまず，専門用語の示す概念は孤立したものでなく関連する様々な概念との関係のなかで成り立っているということが重要である．そのため専門用語の意味を調べる作業はいくつかの関連用語との相互参照的な複雑な作業となり，これを冊子体の辞典で行なうことには限界がある．ハイパーテキストシステムでは計算機パワーと柔軟なユーザインタフェイスによってこれらの問題を解決し，リンクによるスムーズな検索を実現することができる．本稿で示したハイパーテキストシステムでは，リンクによる検索機能と高次検索機能をうまく組み合わせて利用することにより，辞典に存在する情報については，何らかのアクセスルートによってほとんどすべて確実にとり出すことが可能となった．

5 おわりに

本稿では，専門用語辞典をハイパーテキストに自動的に変換するために，リンク，すなわち用語間の意味関係や用語の説明されている位置などを自動的に抽出する方法論を示した．これは，辞典データの文章の中の特徴的な言語表現を利用して，パターン照合によって情報を抽出するという方法である．まず信頼度の高い情報を抽出しその情報を利用してインクリメンタルに処理を進め，必要な情報の抽出度を高めるという方式を用いた．また，抽出したリンクを用いて実際に作成したハイパーテキストシステムを紹介し，その有用性を示した(7)．

本稿で示した用語間の意味関係の自動抽出の方法は，最近話題となっている大規模知識ベースの自動作成のための1ステップであるとも考えられる．今後は，同義語文や内包的・外延的定義文以外の文も対象とし，構文情報などを扱うことによって，情報科学辞典の全体の内容を自動的に知識ベースに変換することを考えたい．

謝辞

本研究はFRIEND21の支援を受けた．本研究を進めるにあたり御援助くださいました大日本印刷の斉藤雅氏，岩波書店の宮内久男氏に感謝致します．

参考文献

(1) Smith,J. and Weiss,S. : Hypertext, *Comm. ACM*, Vol.31, No.7, pp.816-819 (1988).

(2) 長尾 眞ほか, 編 : 岩波情報科学辞典, 岩波書店 (1990).

(3) HAPPINESS/2(V02L20) BASE 使用手引書 第一版 , 平和情報処理センター (1986).

(4) 長尾 眞 : 知識と推論, 岩波書店 (1988).

(5) International Organization for Standardization. ISO 2788: Guidelines for the establishment and development of monolingual thesauri, 2nd ed. Geneva:ISO (1986).

(6) 長尾 眞 : 辞典形式での専門分野の知識の体系的構成法, (人工知能学会論文投稿中).

(7) 黒橋, 長尾, 佐藤, 村上 : 専門用語辞典の自動的ハイパーテキスト化の方法, (人工知能学会論文投稿中).