知識型情報検索システム NIRS の設計およびソフトウェア構造†

伊藤 秀昭††

本論文では，知識型システムとして実現された情報検索システムNIRSの設計，概念構造および実現のためのソフトウェア構造について述べる．本システムの設計および開発の目標は，利用者の有する検索に係わる知識を定義することの可能な環境を提供し，さらに検索要求に知識を用い応えるための知識型情報検索システムを開発することにある．本システムの主たる構成モジュールは，情報検索に係わる知識を格納する知識ベース，これに格納された事実および手続き的知識を解釈し実行するための推論機構，および検索対象を格納したデータベースオブジェクトベースである．知識ベースは，シソーラスの機能を果たす記述子グラフ（DG），およびDGの構造に基づき問い合わせ要求を解釈する手続きの集合などより成る．このとき，手続きは知識ベースの構成要素となる．これを可能とする機能を提供することにより利用者はDGだけではなく，推論メカニズムをも利用者の有する目的および必要に応じ変更することが可能となる．ここで，DGは重み付きセマンティクネットワーク，手続きはプロダクションルールにより，それぞれ，表現されている．

1. はじめに

情報検索システム（以後，IRSと記す）は，利用者の記述する問い合わせ文に記述されている要求（検索要求）を満たす計算機に格納されたドキュメントまたは情報を得る情報システムの一種である[9]．既存の多くのIRSでは，利用者が有する文献に対する知識および検索のために利用することの可能な語彙をシステムに設定し利用すること，ある情報に対する知見を利用した検索を行うことなどは困難である．このような問題点を解決し，検索される情報の質を高めるための一つの方法に情報検索に係わる知識を計算機に格納しシステムを構築する方法がある[10),13]．この種のシステムは，一般に，知識型IRS（KIRSと記す）と呼ばれている．さらに，近年具体的にKIRSを実現することにニーズが高まっている[6]．

KIRSの一つとしてNIRS（Network-based Information Retrieval System）の開発を進めてきた[4]．本システムの検索対象は，収録分野が工学であり，市場に提供されている文献データベース（eg, INSPEC, JOIS, CONFなど）を説明するデータベース台帳総覧[14]に記載されている内容の一部を編集し，再構成したものである．なお，本稿では，データベースを，DBと略記することにする．

本システム開発の目標は，次のようなことである．
—利用者固有の情報検索に係わる知識を利用した情報検索システムの実現．
—これを達成するために必要となるソフトウェア構造の提供．
—情報検索のために対象とする分野の概念構造に基づく推論メカニズムの実現と種々の実験を行うための実験システムの実現．

本システムの特徴は，次のようなことである．
—取り扱う対象が論文などの直接的な文献ではなく，文献を検索するために必要な情報を得るための情報である．したがって，広い学問分野をカバーするためのシソーラスが定義される．本システムでは，この機能を果たすための構成要素を記述子グラフ（DG）と呼んでいる．
—問い合わせ言語（第3.2節）の構成要素である論理結合子 and および or の解釈は，DGの構造を反映している．
—プロダクションシステム（以下PSと記す）PKBUS[16]により機能的に分割される推論手続きの一部およびそれに必要となる情報（例えば，検索範囲を限定するための閾値，関連度の計算方法，など）が定義されている．

PKBUSは，我々がNIRSを実現するために設計および実現したPSの一つである．これはルールベースのモジュール化を可能としており，個々のモジュールをルールセット（RS）と呼んでいる．

現在までに，いくつかのKIRSが既に提案されて

† Design and Software Structure of the Knowledge-based Information Retrieval System NIRS by HIDEAKI ITO (Department of Computer Science, School of Computer and Cognitive Sciences, Chukyo University).

†† 中京大学情報科学部情報科学科

いる．ここでは，シソーラスを知識ベースとみなし開発されたシステムについて述べる．従来から利用されているシソーラスでは階層の上位方向のリンクは利用されていない[2]．しかしながら，本システムでは上位方向のリンクを利用し，検索される記述子の範囲を拡張している．これはできる限り多くの記述子を得ることにより，答集合を多くするためである．ただし，この方法は先の方法に比較して，得られる答集合は多くなるが，精度は悪くなることが知られている[2),9]．NIRS では，再現率の向上を目的としている（第2章参照）ため，むしろこの方法が望ましい．

RUBRIC[7),12] では，記述子を得るために語彙の展開を重み付きプロダクションルールとして記述する．推論は問い合わせ文に出現する語彙を書き換えることに相当する．したがって，推論の結果は分野や用語をノードとする AND/OR グラフとなる．しかしながら，この方法により，分野の構造を反映するようなルールの集合を記述するのみでは，直観的な理解の容易性は低いであろう．

NIRS と同様に，知識ベースとしてグラフを利用した KIRS の一つに，Biswas[3] らにより提案されたものがあり，グラフに基づき分野間の関連性の重みが計算される．この重みの計算方法は，いくつかの理論（統計的手法，Dempster-Shafer の理論，など）に基づき変更することが可能であるが，柔軟性はこの種の計算の多様性に限定されている．しかしながら，NIRS では，ノード間の関連性を求めるための計算方法だけではなく，推論メカニズムを要求に応じ変更することを可能とするようなソフトウェア構造を提供している．さらに，NIRS に組み込まれている推論メカニズムは，概念に基づく推論を達成するためにグラフにより記述される概念構造を利用している．

本稿では，NIRS の設計方針，概念構造およびソフトウェア構造などについて述べる．さらに，本システムにおいて試作，実現した基本的な知識表現およびメカニズムについて示し，その評価は，動作および基本的機能の明確化，ソフトウェア構造の柔軟性，および NIRS に組み込まれている手法の適応可能性を確認するという観点から行った．

2. 動機および設計

2.1 動　機

本システムは，既存の情報検索システムの一つであるデータベースクリアリングシステム（DBCS と記す)[6] の有する問題点の一部を解決することを目的に設計および実現された．

一般に，DBCS と同様にインバーテッドファイルに基づくシステムには，次のような課題がある．

（1） 問い合わせに用いることが可能なキーワードは利用するシステムにより固定されている．

（2） 検索対象として単なるテキスト（文字列）のみを対象としている．これから，次の二つの問題が生じる．まず，問い合わせ言語を用い記述される語彙が検索対象に含まれていないならば，IRS から得られる情報がない，または量が少ない場合がある．さらに，利用者は提供されたキーワードにより検索要求を記述する．この種の問題を解決するためにシソーラスを利用することはよく知られている．このとき，検索要求を記述するための語彙を拡張または利用者固有のものに変更することを容易とすることが必要である．

（3） 利用者の有する知識を反映し，IRS がこれを利用することが困難である．例えば，利用者個人の有する語彙の種類と効果，他の語彙との間に存在する関連性の利用，など．したがって，個人の有する検索のための知識を定義することが可能なシステム構成であることが必要となる．

（4） インバーテッドファイルに基づく論理結合子 and および or の処理は，集合演算の集合積および集合和に対応している．この方法では，分野に存在する概念構造を反映することは，困難である．

NIRS を実現するために次のようなことを仮定した．

—自然言語により記述された概要から抽出される単語（記述子，インデックス・ターム，と呼ばれている[1)]）が，個々の検索対象である DB を特徴づける．

—分野を表すための概念を記述するノード間に付加された重みは，相互の関連度を表し，それらは事前に与えられる．

—主題分野を表す個々の概念は，研究対象，問題，性質などの属性を有する．これは，階層の下位に位置する概念に受け継がれる．ただし，このような属性を個々に記述することは困難であるため，このような属性は存在するものとする．

—概念間の関係と関連度は利用者が，また，推論手続きは管理者または利用者が記述する．このとき推論のための管理機構は，本システムが固有に提供する．

2.2 設計の方針

設計および実現の方針として考えたことは，次のよ

うなことである．

—情報検索を行う利用者の固有の概念に基づき定義される検索のための知識を利用すること．

—直感的な理解の容易な知識表現および推論メカニズムを提供すること．

—推論は概念に基づくこと．これは問い合わせ文の解釈に際し，グラフの有する構造をいかに反映するのかということに相当する．

—知識表現は，PS，重み付きセマンティクネットワーク（WSN と記す）とすること．

—IRS としてのみ機能するシステムではなく，利用者固有の知識を定義するためのインタフェースを備えること．

—ケーススタディを通じ，システムの拡張を容易とすること．

—拡張および利用者固有の知識の一部を記述を可能とするために，知識表現システムを開発すること．

本システムを実現するための知識表現として，WSN および PS を利用した理由は，次のようなことである．一般に，IRS の利用者は，個々の学問分野やその体系などについて，種々の知見を有するであろう．これらは，分野の階層の上位-下位関係を中心として，他の分野との関係に基づき定義されることになるであろう．さらに，関連の強さを表す関連度は，均一ではなく，個々の状況により異なると考えられる．WSN では，ノードは学問分野などを記述する（後述）概念対象に，リンクはこれらの間の関連のタイプに，それぞれ，相当する．リンクにはノード間の関連度が重みとして付加されている．ノードとリンクによる記述は，従来のシソーラス[1),8),10)]と同様であるが，NIRS ではリンクのタイプに5種あること（第3.2節B），テキストがシソーラスのノードに相当しないこと，リンクに重みが付加されていること，などが異なる．

上に述べたような構造および問い合わせ要求を解釈するための手続きが必要である．本システムでは先に述べた目的および方針より柔軟な構造および記述方法を提供することが要求される．

PS は，固有のインタプリタによりルールが実行され，インタプリタとパタンマッチングを提供していること，実行単位のモジュール性が高いこと，などの利点を有している[15)]．PS では，本システムの記述言語である Lisp（第3章参照）により直接記述することに比較して，問題の性質およびプログラムの目的に依存することになるが，少ない記述により WSN および問い合わせ文の解釈のための手続きを記述することが可能であると考えられる．ただし，このような推論メカニズムの実現方法は，実行効率という観点から効率の悪いものとなる．しかしながら，我々は，実行効率というよりむしろ，記述の負担の軽減とモジュール単位による記述の容易性を目標としている．本システムの開発に際して PS を実現した理由は，手続きを記述するための言語として柔軟であるということよりもむしろ，それを用い実現されるある機能を果たすタスク単位の再構成が記述言語である Lisp に比較して容易であるという柔軟性による．

3. システム構成

3.1 概　　略

NIRS の概念的な構造を図1に示す．本システムは，三種の構成要素から成り，それらは，推論機構，知識ベースおよび検索対象を格納したデータベースオブジェクトベース（DBOB と記す）である．

知識ベースは，情報検索を行うための手続きおよび DG より成る．手続には，DG を解釈し，問い合わせに答えるための手続きがある．すなわち，論理結合子（第3.3節参照）を解釈するために DG の構造を反映する推論メカニズムを記述した論理結合子処理知識，および求められた結果のランク付を行うランク付知識である．

推論機構は，利用者から与えられた問い合わせ文を解釈し，答となる情報の集合を得るための推論を行う．これは，問い合わせ解析部，問い合わせ文処理部

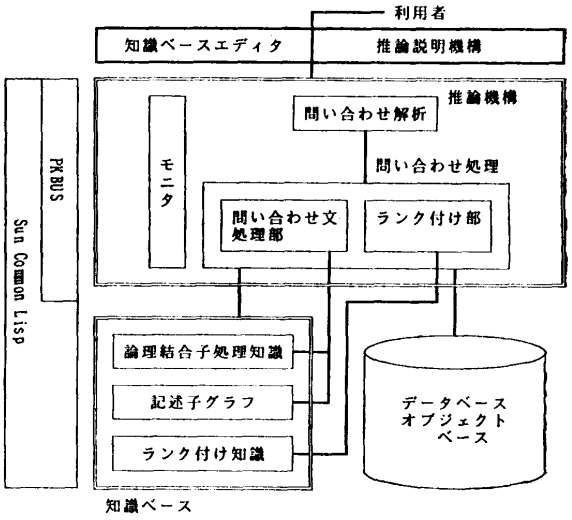


図1 NIRS の概念的構造
Fig. 1 An overview of conceptual structure of NIRS.

およびランク付部より成る．問い合わせ解析部は，利用者が記述する検索要求をシステムが処理する形式に変換する．問い合わせ文処理部は，解析部から得られた結果に基づき検索を行う．さらに，ランク付部は，結果となる答集合をある順序に従って並べるためのモジュールである．このような一連の処理の流れは，モニタにより管理される．このように本システムにおける推論機構は，知識ベースに定義された手続き的知識の起動を行うための枠組みを提供している．なお，問い合わせ文処理部およびランク付部において用いられるプログラムは，先に述べた知識ベースの構成要素である論理結合子処理知識およびランク付知識に，それぞれ，格納されている．

利用者のために設けられたインタフェースは，上に述べた知識ベースの構成要素を更新するための知識ベースエディタ，および推論の過程および結果を表示，推論において利用されたDGのノードの表示，などの機能を備える推論説明機構より成る．

また，本システムは，PS PKBUS および Sun Common Lisp[11] により記述され，S-4 ワークステーションにより稼働する．

3.2 システムの構成要素

A．検索対象および問い合わせ文

図2に，文献 DB である 'CONF' の記述例を示す．個々の DB を記述するために，次のようないくつかの属性がある．データベースオブジェクト (DBO) の識別子を格納するための @id，DB の正式名称およびよく知られた通称などを記入するための @name，アブストラクトを構成する単語列から成る @abstract，および収録地域を示す @area など，がある．

属性 @abstract には，各々の DB の特徴，例えば，主たる収集分野または主題，収録のための情報源（雑誌，特許情報など），編集学会および関連ファイルなどを表す名詞の列が記入される．

```
*dbo
    @name "CONF"
    @id   d00008
    @abstract  "物理学" "数学" "コンピュータ学"
          "天文学" "天体物理学" "航空"
          "宇宙工学" "エネルギー" "自然科学"
    @dbp  78
    @area 世界
    @items  会議名 主催機関名 会議開催地
          会議開催日 問い合わせ先 関連出版物
          キーワード
    @sour 会議資料
    @field 自然科学 技術
    @keyword  学会 会議 科学技術
    @language  eng
```

図2 データベースオブジェクト記述例
Fig. 2 An example of description of a database-object.

問い合わせ文は，次の形式により書かれる．

$(Q_1\ op_1\ Q_2\ op_2 \cdots op_n\ Q_{n+1})$

ここで，op_i は and または or である．一つの Q_j は，'(P c_i)' と表され，P を述語，c_i を引き数と，それぞれ呼ぶことにする．述語Pは，検索要求がいかなる種類の情報であるものかということを示す．**表1**に，代表的な述語の種類を示す．述語には大別し二種あり，一つは DG を用いる問い合わせに用いられるものであり，他方は DBO を記述するために設けた属性の値を参照するためのものである．例えば，"機械工学および鉱山工学に関する文献" は，"(subj 機械工学) and (subj 鉱山工学)" と表現される．ここで，述語 subj のための引き数 c_i は DG に定義されたノードが対応し，また，述語が DBO を定義するための属性に相当するならばその属性値の集合に属する一つの要素が c_i に記入される．

B．記述子グラフ

ノードとなる記述子は，主題，一般語および固有名詞の三種のクラスに分類されている．それらには，分野における学問領域 ("機械工学", "電子工学", など) を表す名詞が，"手法", "理論" などの比較的多くの分野においても出現すると考えられる名詞が，学会名称，関連ファイルなどの固有名詞が，それぞれ属する．これら三種のクラスを，それぞれ subj, gene および prop と呼んでいる．これは，先に述べた問い合わせ言語に属する述語に相当する．このようなクラスを設けた理由は，次のようなことである．検索対象とする各々の DB の記述における概要には，個々の DB がどのような分野の文献，データなどを収録しているのか，そしてその特徴としていかなる観点より構築した文献 DB であるのかという話題 (主題) が中心に記述されている．このとき，名詞が DBO の特徴を表すために中心的な役割を果たすと考えられる．したがって，これらを整理し，記述子を分類することにより DG の構造が明確となる．現段階では上記の三種によ

表1 述語の種類
Table 1 Some predicates and their examples.

述語		内容	例
記述子グラフ	subj	主題，分野を表す	(subj 機械工学)
	gene	一般的な用語	(gene 理論)
	prop	固有名詞	(prop IEEE)
属性	area	収録地域	(area アメリカ)
	sour	収録情報源	(sour 特許)
	lang	文献記述言語	(lang eng)

って十分であると考えている．

二つの記述子を結合するためのリンクには，次に示す5種がある．それぞれのリンクには0から1の範囲にある実数値による重みが付加される．説明のために，ここでは，あるリンクにより結合される二つのノードおよび重み間の関係を三項関係 'R(a b w)' と書くことにする．ここで，Rはリンクのタイプ，aおよびbは記述子，wはリンク付加された重みである．記述子aはリンクの始点，bは終点となるノードである．重みはある二つの記述子間の関連度の強さを表すとともに，問い合わせ要求に対し推論の結果として得られた DBO を順序づけるために利用されている．

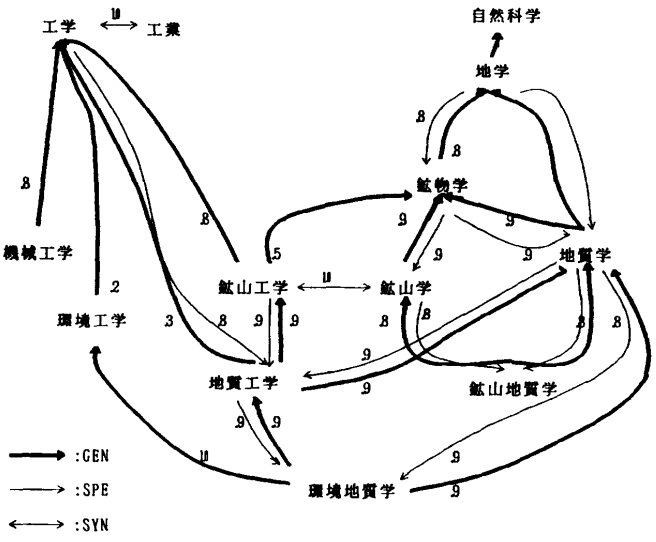


図3 記述子グラフの一部
Fig. 3 A part of the descriptor graph.

（1） Gen (a, b, w): 記述子aの上位概念または上位語は b. 上位語を複数定義することも可能である．

（2） Spe (a, b, w): a の下位概念または下位語は b. gen と spe は記述子間の階層構造を定義するリンクである．

（3） Syn(a, b, w): aおよびbは同義語．このリンクは双方向であり，bからaへのリンクの重みは，同一である．

（4） Ass(a, b, w): bはaより連想される記述子．例えば，Ass (機械工学，設備，0.4)．

（5） Sea(a, b, w): See also の略であり，関連のある分野に対する参照を表すためのリンク．

図3に，NIRS において実現した DG の一部を示す．

リンク "syn" を除き，他のリンクは単方向の参照である．これは，重みの計算において，検索範囲を限定するために用いられている．例えば，比較的検索要求が明確であり，その上位へのパスのアクセスの範囲を制限するためには，階層の上位方向の範囲は数値により制限される．

これらのリンクを設定した理由は，次のようなことである．一般に，IR において記述子間を結合するリンクの種類には，大別し，3種の基本的な関係，すなわち，等価または同義であることを示す等価関係，上位-下位概念を示す階層関係，およびこれら以外に関連を記述するための連想関係がある[1]ことが指摘されている．

DG を構築するためには，概念の有する階層構造にしたがい整理することが見通しを高める方法の一つであろう．例えば，工学-環境工学-環境地質学，自然科学-地質学-環境地質学．この種の階層構造を記述するために "gen" および "spe" を設けた．むろん，一方のリンクによりこの構造を記述することは可能であるが，上に述べた理由により，この2種を設けている．

等価関係を示すリンクは，"syn" であり，その DG において同義とみなされる概念を結合する．ある二つの記述子がこのリンクにより結合される多くの場合は，"コンピュータ" と "電子計算機" のように同一な概念に対する別称を記述することに用いられている．しかしながら，この種の事例のみではなく，鉱山工学および鉱山学のようにクラス subj に属する記述子間に設定される場合もある．これは，主題分野が狭いならば概念が異なるが，DG の扱う主題分野が広いために同義語と考えられるからである．

リンク "ass" および "sea" は，連想関係に相当する．これらは，前者のリンクに比較して明確な関係ではないが，ある概念から連想されるノードを記述するためのものである．このとき，同じクラスに属するノード間および異なるクラスの属するノード間は，それぞれ "sea" および "ass" リンクにより結合される．この両者を区別することにより，DG の構造と，その結果得られる他分野との関係がより明確となる．例えば，ある主題に関する問い合わせを発したとき，その結果の評価は DG の検討を伴うことになるであろう．このとき，リンクの種類が区別されていることによ

り，実際に検索された記述子が主題分野を記述するものであるかどうかの判定が容易となる．

3.3 記述子グラフの利用

ここでは，NIRS に組み込まれている DG の構造を反映し，主題に関する概念（クラス subj に属する）に基づく推論を達成するための DG の利用方法について述べる．他の種類の述語およびそれらの組み合わせについては，機会を別に述べたいと考えている．現在我々は，主題の検索を対象としており，これが IRS においても最も重要な課題の一つとなっている．

NIRS の提供する推論機構では，階層を記述する “spe” および “gen” リンクが，“and” および “or” に対応づけられ問い合わせ文が解釈される．一方，この2種のリンクを含め5種のリンクは，ある記述子と関連する記述子を求めるために用いられている．むろん，この解釈を実現するためには，階層の記述のためのリンク，および関連を記述するためのリンクを設けることにより実現することが可能である．しかしながら，この方法では，リンクの数が少ないことにより記述子間の関連が不明確となり混乱したものとなるであろう．また，開発されるシステムの目的および要求される機能に依存することになるが，リンクがより細分化されているならば，細かな指定が可能となる．しかしながら，それらを使い分けることが困難な状況も生じることがあるであろう．これらの理由により，5種のリンクを用いることにより，現時点では十分である．例えば，物理的な全体-部分関係は，“sea” または “ass” を用い記述される．また，ある主題分野の記述において，それが他の主題分野の一部に注目するものであるならば，下位概念として記述することが適切である．

本節では，まず，主題に関して関連するとみなされる記述子の計算方法を示し，次に，論理結合子の解釈について述べる．

述語 subj を用い記述される問い合わせ文に応えるために求められるすべてのノードの重みは，**図4** に示す二つのタイプの計算方法により計算される．図4.a は二つのノードがある関係により直列に結合されていることを，図4.b はある一つのノードからあるノードに対して複数のノードを経由して入力があることを，それぞれ表している．ここで，w_m および k_n は，それぞれノード c_m の重み，リンクに付加された重みである．図4.a および b に示されるノード a の重み w_a は，次式により計算される．

（a） $w_a = w_1 k_1 k_2$,

（b） $w_a = \mathrm{Max}(w_1 k_1, w_2 k_2, w_3 k_3)$

このようにして求められた w_a とシステムに設定された閾値（*Th*）の比較が行われ，w_a が閾値よりも大きい値（$Th \leqq w_a$）ならば，ノード a は記述子 b に関連する記述子（概念）とみなされる．このとき，閾値は利用者が目的に応じ設定する．

問い合わせ文が一つの要素からなる文であるならば，その近隣点（後述）が求められる．一方，結合子 “and” または “or” を用いた文であるならば，本節において述べるような複合分野または包含分野を表す記述子を中心にして近隣点が求められる．なお，この処理の流れは，第3.4節に示す．

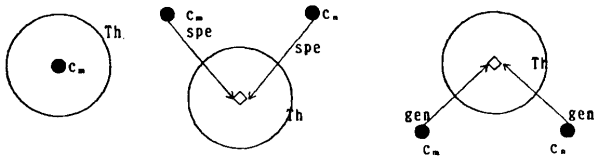
問い合わせ文の構成要素である一つの（P c_n）の処理では，ノード c_n の重みを1に設定し，上に示した計算により関連する記述子が求められる（**図 5.** a）．この操作により利用者があいまいな語彙により問い合わせ文を記述することを可能としている．このようにして求められた記述子を近隣点と，ここでは呼ぶことにする．

あるノード c_1 の近隣点の集合は，次の式により記述され，リンクにしたがい適応されるノード c_3 の集合である．

$$\forall c_1, c_2, c_3, k_1, k_2, \exists w_1, w_3$$

（a） 単一経路　（b） 多重経路
（a） Single path.　（b） Multiple paths.

図 4 ノード間の関係
Fig. 4 An overview of relation among nodes for weighting calculation.



（a） (subj c_m).　（b） (subj c_m) and (subj c_n).　（c） (subj c_m) or (subj c_n).

図 5 問い合わせ処理のための検索ノード
Fig. 5 Retrieving neighborhoods for processing primitive queries.

$$(R_m \times R_n(c_1, c_3, w_3) \wedge Th \leqq w_3) \rightarrow$$
$$(R_m(c_1, c_2, k_1) \wedge R_n(c_2, c_3, k_2)$$
$$\wedge Th \leqq F(w_1, k_1, k_2))$$

ただし，c_i はノード，R_i は第 3.2 節Bに示したすべてのタイプのリンク，w_i はリンクに付加された重み，Th はシステムに設定されている閾値である．このとき，w_3 はノード c_3 の計算された重みであり，w_1 は初期値として用いられる値である．なお，×は関係の合成であり，近隣点の集合はリンクに関して繰り返し求められる．また，Fは前記(a)の方法による計算である．

問い合わせ "(subj c_m) and (subj c_n)" は，次のように解釈される．これは，ある概念を表すノード c_n および c_n により示される分野の複合領域となる分野に関するデータおよび文献を収録する DBO を求めることに相当する．例えば，"地質学" および "鉱山工学" の複合分野として "地質工学" および "環境地質学" がある（図3参照）．したがって，このようにして得られる概念が問い合わせに示される概念における主題において共通の問題を取り扱うまたはより特殊化された問題を扱う分野であると考えられる．ここで，複合分野は DG における二つのノード (c_m, c_n) から spe リンクにより結合されているノードが相当する（図5. b)．

一方，結合子 or を用いた問い合わせ "(subj c_m) or (subj c_n)" は，c_m および c_n に示される分野を結合子 "and" とは逆に包含する領域であると解釈される．例えば，"地質学" と "鉱山工学" の二つの分野を包含する上位概念は，"鉱物学" である．これは，二つのノードから gen リンクを辿ることにより，双方からのパスが一致するノードに相当する（図5. c)．

このようにして得られたノードを中心に近隣点が求められる．したがって，"and" および "or" の解釈により，複合分野または包含分野とみなされる記述子とそれに関連するとみなされる記述子を含む DBO が求められる．

概念的には，"and" および "or" の解釈は，上に述べたものである．より形式的には，階層構造において次のようなことに相当する．ここで，二項述語 spe* および gen* の引き数は共にノードであり，それぞれの述語において第一引き数は，第二引き数の下位または上位ノードであることを示す．なお，ここでは混乱が生じないと思われるので，簡略化のために重みは記述しないことにする．また，spe1, gen1 は，第 3.2 節Bに示した spe および gen リンクに相当する．これらの述語を用い階層のすべての階層の下位ノードは，次のように表される．

$$\forall c_1, c_2 \ \mathrm{spe1}(c_1, c_2) \rightarrow \mathrm{spe*}(c_1, c_2)$$
$$\forall c_1, c_2, c_3 \ \mathrm{spe*}(c_1, c_2) \wedge \mathrm{spe1}(c_2, c_3) \rightarrow \mathrm{spe*}(c_1, c_3)$$

同様に spe* および spe1 を，それぞれ，gen*，gen1 に書き換えることにより，すべてのあるノードの上位ノードが表される．

結合子 "and" および "or" は，一般には複数の引き数が与えられるが，ここでは，二つの引き数が与えられる場合について示す．例えば，問い合わせ "(subj c_1) and (subj c_2)" が与えられたとき，c_1 および c_2 の下位ノードが上の式により示される．ノード c_i について得られるノードの集合を，$N(\mathrm{spe*}, c_i)$ と書くならば，この問い合わせの解釈されるノードの集合は，$N(\mathrm{spe*}, c_1) \cap N(\mathrm{spe*}, c_2)$ であり，また，結合子 "or" は，$N(\mathrm{gen*}, c_1) \cap N(\mathrm{gen*}, c_2)$ である．

これは，形式的にではないが，第 2.1 節において示した第3項目の仮定より，次のように解釈されることに相当する．あるノードにより示される概念は，本システムでは記述されていないが，その分野で取り扱われる問題，性質，研究対象などの属性を備えていると考えられるであろう．ここで，あるノード c_i の属性の集まりを A_i とし，問い合わせ "(subj c_1) and (subj c_2)" により，複合分野がいくつか得られたとする．このとき，複合分野に相当する個々のノード n_i は，$A_1 \cap A_2$ なる属性を持つであろう．一方，or を用いた問い合わせにより得られるノードは，属性に関し $A_1 \cup A_2$ と考えられる属性を備えることになるであろう．これらの解釈は，"and" および "or" により示される分野が，それぞれ，複合分野および包含分野に相当すると考えている．

このとき，複合分野または包含分野が存在しない問い合わせ文が与えられる場合がある．このシステムでは，これは例外として扱われている．このような場合には，それぞれのノードを中心に近隣点を求め，それを含む DBO を答集合としている．むろんこれは，これまでに述べてきた二つの論理結合子の解釈とは異なる．さらに，この処理の結果得られる答集合の精度は低下すると考えられる．しかしながら，再現率を向上するために関連があると評価された記述子を含む DBO を可能な限り得る，および出力がないということを可能な限り防ぐという観点から，この処理を行っている．

また，結合子 "and" および "or" の処理において，リンク "syn" の処理により，DG の解釈が混乱することがある．これは，問い合わせ文を構成する c_1 または c_2 に同義語が存在したときである．このとき，二つの記述子が同値なものとして処理されるならば，複数の階層にまたがりリンクが設定されることになるため論理結合子の解釈が相異なるパスに基づき適応されることになる．したがって，NIRS では，現在これを反映していない．これは，本展開手法が階層構造に基づき結合子を解釈するため，"syn" によりその構造が混乱することを避けるためである．このような DG の状態は本システムの備える知識ベースエディタ，推論説明機能を用い確認することが可能である．この問題は，DG の表現とその利用に関して，次のように考えている．上に述べたように推論において混乱が生じる場合があるため，および，同値な関係があるような記述子を用いる場合には，DG においてそれを確認することにより，より問題分野の構造を反映する問い合わせ文を記述することが可能となるであろうと考えられるため，リンク "syn" は，他のタイプのリンクと同様に近隣点を求めるためにのみ用いられている．

3.4 問い合わせ処理の流れ

問い合わせ処理を実現するための PKBUS による一連の処理の概要を，RS の制御に基づき，**図 6** に示す．ここに示した制御の流れは，活性化される RS の順序を表している．ここで，ノードは RS の名前，矢線は RS 間の制御の流れ，二重矢線はデータの流れを，それぞれ表す．また，四角で囲まれている(I)，(II)および(III)はそれぞれ，モニタ，論理結合子処理知識およびランク付知識に格納された RS である．

例えば，work-area は問い合わせ文を処理可能な形式に変換する問い合わせ解析部である．また，node-proc，and-operation および or-operation は，それぞれ，一つの Q_i (第 3.2 節A)，論理結合子である and および or を解釈するための RS である．それらは，論理結合子処理知識ベースに格納されている．また，result-weight は，得られた記述子の重みを基に，個々の DBO の重みを計算している．この重みは，記述子の実現値であるテキストに付加される重みを合計したものである．この計算方法は，result-weight に定義されたルールを変更することにより，例えば最大値を得るというような手続きに書き換えることができる．これらのほかに，近隣点を求めるための RS および DBO を検索するための RS など，がある．

4. 実 験 例

小規模な DBO の集合を対象として実験を行った．現在，我々が検索対象として，DBO は 60 件あり，DBCS の有する文献 DB の約 10 分の 1 に相当する．また，DBO の数に比してノードとリンクの数は多く，リンクおよびノードの数は，それぞれ約 1000 本および 350 個である．これは，今後の DBO の数の拡張に備えるため，および抽象度の低い概念を表す記述子のために必要となる中間的な記述子を定義したためである．

図 7 に，問い合わせ "((subj 地質学) and (subj 鉱山工学))" に対する実行結果の一部を示す．ここで，問い合わせは，QUERY (USER) に示されている．この問い合わせに対する答集合の出力において，第一カラムは順位であり，第二カラムは DB 名，次に本システ

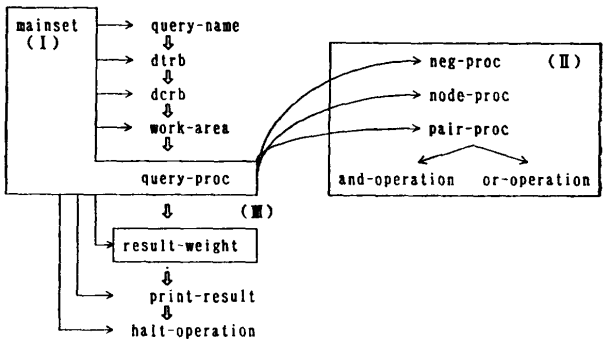


図 6 ルールセットによる問い合わせ処理の概要
Fig. 6 A flow of query processing represented in rule sets.

```
query(user)   =  ((subj 地質学) and (subj 鉱山工学))
query(system) =  ((subj 地質学) and (subj 鉱山工学))
1  Ei Engineering Meetings (d00004)      <  7.84      0.00>
2  JICST科学技術文献ファイル (d00039)    <  6.23      0.00>
3  GEODE (d00011)                        <  6.03      0.00>
4  System for Information on Grey Literature in Europe (d00032)
                                         <  5.06      0.00>
5  Science Citation Index (d00029)       <  3.24      0.00>
6  PASCAL (d00024)                       <  2.99      0.00>
7  GEOMECHANICS ABSTRACTS (d00053)       <  2.80      0.00>
8  TELEGEN (d00037)                      <  2.25      0.00>
9  Engineering Microsoftware Review (d00051)
                                         <  2.16      0.00>
10 Oceanic Abstracts (d00021)            <  1.80      0.00>
11 CORP (d00009)                         <  1.70      0.00>
12 AGRICOLA (d00044)                     <  1.54      0.00>
```

図 7 システムの実行例の一部
Fig. 7 An experimental answer set for a query.

ムがDBOを管理するために付加している識別子が表示されている．これは，“(”および“)”により囲まれている．さらに，“〈”および“〉”により，個々のDBOの計算された重みが表示されている．

本稿において述べたように，この問い合わせでは，二つの分野の複合領域である地質工学および環境地質学を中心に近隣点が求められる（図3参照）．本システムから得られる答集合の大きさを決定する近隣点となるノードの個数は，閾値の値を変更することにより変えることができる．

本システムの展開方法の有効性を示す一例として，**図8**に，次に示す二つの問い合わせ文，

① ((subj 地質学) and (subj 鉱山工学))

② ((subj 地質学) or (subj 鉱山工学))

に対する閾値の変化による再現率と精度の結果を示す．

問い合わせ文①では，従来のインバーテッドファイルを用いた演算により処理を行うならば，答集合は得られなかった．また，②に対する答集合の精度および再現率は，我々の行った評価によるならば，それぞれ，66.7% および 20% であった．

NIRSのキーワード展開の方法では，①および②に対して閾値の変化により得られるDBOの件数は異なる．

②では，閾値が 0.7〜0.8 の間において再現率が変化することなく精度の向上がみられる．したがって，閾値が0.8のとき，最も良好な結果が得られると考えられる．このとき，精度および再現率は，85.7% および 60% である．これは，従来の方法に比較して，19% および 40%，それぞれ，向上している．ただし，この閾値は，常に，0.8 が最適ということではない．

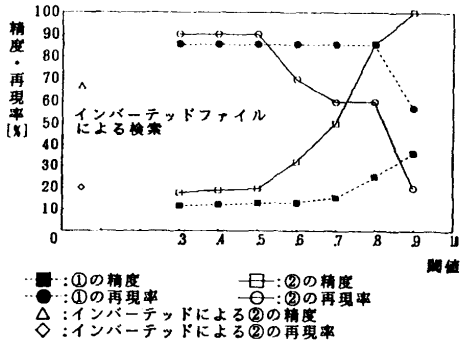


図 8 問い合わせ文①および②を用いた閾値の変化による精度と再現率の推移

Fig. 8 Precision and recall for the queries ① and ② with varying a threshold.

このように最適である閾値を決定することは試行錯誤を伴うであろう．本システムにおいて閾値の変更は，エディタを用い行われる．現段階では，この機能を用い検索の目的および問い合わせ文の構造にも依存することになるが閾値を高く設定し，答集合の大きさを確認しながら，順次その値を小さくすることが効率よく答集合を得る方法であると考えている．

本システムの動作とシステムの提供する柔軟性の確認，インバーテッドファイルを用いた従来の手法と実験システムとしてNIRSが提供するキーワードの展開手法を比較することによりその有効性の確認，および，実験例を通してDGをより整備するために設けた13件の文例を用い得られる答集合を再現率および精度を求めることにより評価した．これらの文例は，これまでに述べてきた主題に関する問い合わせ文である．ただし，文例の数は少ないが，DGの構造と“and”または“or”の解釈において相当するノードが存在する場合，存在しない場合を含んでおり，その動作を確認することが可能である．取り扱う文献の数を増やし，より一般的なIRSとしての評価は機会を別に述べたいと考えている．さらに，これらの文例に評価を通じて，DGがより充実したものとなると考えられる．

従来の手法では，13件のうち7件の問い合わせについて答集合が得られた．この7件の再現率と精度は，それぞれ，33.2% および 77.4% である．

一方，本手法では閾値を調整することにより，すべての問い合わせに対し答集合が得られる．閾値の変化による再現率および精度，また，実際に答集合の得られた文例の数の変化を，**表2**に示す．例えば，閾値が0.9のとき，再現率は 51.4%，精度は 54.6% および答集合の得られた問い合わせの文例の件数は8件であることを示す．この表に示すように，再現率の向上がみられた．この結果および上述の①および②の文例に示したように本展開手法は問い合わせがDGの構造と適合したものであるならば有効であったと考えている．

表 2 閾値の変化による精度と再現率の推移

Table 2 Precision and recall for 13 queries with varying a threshold.

閾　値	0.9	0.8	0.7	0.6	0.5	0.4	0.3
再現率	51.4	51.9	70.8	76.3	86.7	81.6	85.7
精　度	54.6	36.8	27.5	24.6	18.8	17.2	15.2
件　数	8	10	10	10	11	13	13

また，例外処理および従来の手法による検索を実現するためのプログラムは，本システムの提供するソフトウェア構造により実現することが可能であった．これは，本システムの柔軟性を示唆すると考えられる．

5. おわりに

本システムの開発の長期的な目標は柔軟なソフトウェア構造を備え，概念に基づく KIRS を供することにある．その一段階の目標は，検索システムとして再現率の向上を図るということであった．この点において第一段階は終了したと考えている．

NIRS は情報検索のために必要となる知識において，記述子間の関係およびその関係を用いた推論のための知識を有している．これは，本システムでは DG を通して問い合わせの検索要求が解釈され，この結果として答集合を求めるということである．これらの知識には，DG とそれを解釈するための手続き的知識がある．解釈を行うための手続き的知識には，DG のリンクと論理結合子間の関係を用いた問い合わせ文を解釈する手続き，およびあいまい性を含む情報（数値により表されている）に基づく推論を達成するための手続きがある．

本システムには，次の二つの利用方法がある．これは，利用者により定義される知識の種類についてのものである．一つは，閾値および記述子グラフの値または構造を利用者固有に構成することである．この種の知識は，比較的容易にエンドユーザがエディタを用い，現在定義されている DG を修正および更新することが可能である．これにより検索に利用される記述子の範囲を変更することが可能である．一方，本システムの提供する推論のためのメカニズムを利用し，提供されている記述子検索のための推論方法を PKBUS を用い記述することにより固有に変更することが可能である．ただし，このときモニタの提供する機能を変更する機能は提供されていない．むろん，利用者の負担は，前者の種類の知識の定義に比較して，大きいものとなる．

我々の経験に基づくならば DG を初期状態より混乱することなく定義することは容易ではない．この定義を支援するための一貫性管理機能を備えたインタフェースおよび説明機能の拡充などが今後システムとして必要である．さらに，DG の構造を利用者の用いた問い合わせ文から得られる答集合に対する応答を DG に柔軟に反映するためのメカニズムが必要であると考えている．この問題は知識ベースの半自動的な更新と深い係わりがあるであろう．従来の IR において提案されているメカニズムはドキュメントに付加された重みの変更により利用者からのフィードバックを反映している[9]．しかしながら，本システムでは DG の構造とリンクに付加された重みの変更に相当することになるであろう．

現在，NIRS 充実のために既存の関係型 DB を利用する方法を検討を進めている．また，本システムを情報検索システムとして評価することは今後の課題の一つである．これらについては機会を別に述べたいと考えている．

謝辞　日頃よりご指導くださる中京大学　福村晃夫教授に深謝いたします．また，様々なコメントをいただいた中京大学　嶋田晋講師に深謝いたします．さらに，本研究は，筆者が(財)日本情報処理開発協会在職中に始められたものです．本システムの設計開発に協力していただいた米田順美氏に深謝いたします．本研究の継続の機会および許可を与えていただきました山本欣子常務理事，市川隆開発研究室室長に深謝いたします．最後に有益なご意見をいただいた査読者の方々に深謝いたします．

参考文献

1) Aitchison, J. and Gilchrist, A.: *Thesaurus Construction*, 内藤，中倉ほか(訳)：シソーラス構築法，丸善 (1989).
2) Arikawa, S. and Kitagawa, T.: Multistage Information Retrieval System Based upon Researcher Files, *Proc. 2nd USA-Japan Comp. Conf.*, pp. 149-153 (1975).
3) Biswas, G. et al.: Knowledge-Assisted Document Retrieval: II. The Retrieval Process, *J. Am. Soc. Inf. Sci.*, Vol. 38, No. 2, pp. 97-110 (1987).
4) 伊藤，米田：知識型情報検索システム NIRS の構造について，「人工知能システムの枠組み」シンポジウム，pp. 151-160 (1987).
5) 伊藤，上野：フレーム型知識表現システム ZERO における付加手続きとしての Prolog，人工知能学会誌，Vol. 3, No. 3, pp. 337-349 (1988).
6) 機械システム振興協会：高度データベースシステムの開発に関するフィージビリティスタディ報告書(3) (1985).
7) McCune, B. P., Tong, R. M. et al.: RUBRIC: A System for Rule-based Information Retrieval, *IEEE Trans. Softw. Eng.*, Vol. SE-11, No. 9, pp. 939-945 (1985).
8) 日本科学技術情報センター：タームチャート

(1982).
9) Salton, G. and McGill, M. J.: *Introduction to Modern Information Retrieval*, McGraw-Hill (1983).
10) Salton, G.: Another Look at Automatic Text-Retrieval Systems, *Comm. ACM*, Vol. 29, No. 7, pp. 648-656 (1986).
11) Sun: Sun Common Lisp Manual, ver. 3.0, Sun Micro (1988).
12) Tong, R. M., Shapiro, D. G. et al.: A Comparison of Uncertainty Calculi in an Expert System for Information Retrieval, *Proc. IJCAI-83*, pp. 194-197 (1983).
13) Tong, R. M. (ed.): *International Journal of Intelligent Systems*, Special Issue on Knowledge-based Techniques for Information Retrieval, Vol. 4, No. 3 (1989).
14) 通商産業省：データベース台帳総覧 (1983).
15) 上野：知識工学入門，オーム社 (1989).
16) 米田ほか：プロダクションシステム RKBUS の構造について，第 36 回情報処理学会全国大会論文集，pp. 1417-1418 (1987).

(平成2年11月22日受付)
(平成3年 6 月13日採録)

伊藤 秀昭 (正会員)

昭和57年東京電機大学工学部経営工学科卒業．昭和59年同大学大学院修士課程修了．同年博士課程進学，昭和60年中退．(財)日本情報処理開発協会を経て，現在，中京大学情報科学部情報科学科講師．知識表現システム，知識型情報検索システム，データベースと知識ベースの統合化の研究に従事．人工知能学会，電子情報通信学会各会員．