データベースシステム 110－7
（1996. 10. 24）

用語木を用いたドキュメントの概念構造の可視化
-用語木の概念構造と組織化機能-

福永真美　古賀真澄　下井文彦　井上弘隆　岸本令子　伊藤佐智子
（株）学習情報通信システム研究所
〒069　江別市西野幌45
E-mail : {fukunaga, masumi, shimoi, inoue, kishi, itos}@srl.co.jp

概　要

データベース探索によって得られる大量の情報検索結果（ドキュメントの集合）を学習者に効果的に提示するために，対象領域の概念体系を表す用語木を用いてドキュメントの概念構造を分析し，可視的に表現する．本研究では，ドキュメントを構成する要素用語とその出現頻度を用いてドキュメントの索引を動的に生成して「索引マップ」として表示する．利用者は可視化された索引マップを参照することでドキュメントの概念的な位置付けや情報の詳細度を知ることができる．ここではドキュメントが表現している対象領域の分析に重点を置き，非構造ドキュメントから概念構造を抽出するための基準として用語木を利用する．

Visualization of Knowledge Structure using Tree Structures of Terms
-Knowledge Structure and Organizing Functions of Terms-

Mami Fukunaga Masumi Koga Fumihiko Shimoi Hirotaka Inoue Reiko Kishimoto Sachiko Ito
Software Research Laboratory
45 Nishinopporo Ebetsu Hokkaido 069 Japn
E-mail : {fukunaga, masumi, shimoi, inoue, kishi, itos}@srl.co.jp

abstract

As a user's support to handle a large quantity of retrieval results by searching data bases, it is very effective way that analyzes knowledge structures of documents and displays them visually. In this paper we discuss how we represent the knowledge structures of unstructured documents. Although there are some researches about semantics or context level's information retrievals from unstructured documents, we focus on the analyzing domains that documents are contained, and use term trees as a standard to retrieve information. We make indexes of documents by using keywords that appear in documents, and display them as an 'index-map'. Users can get not only the conceptual position but also the detailed information of documents by referring visualized index-maps.

1．はじめに

ネットワークに分散した大量の情報を対象とした情報処理では，人間とコンピュータ間の相互作用として，情報の提供と可視化が重要な問題である[1]．情報の可視化に関する研究として，ドキュメントの内容の分析と可視化に関する研究が行われている．ドキュメントを概念的に分析するためには基本となる領域知識が必要である[2]．分析の基準として辞書や特定の文法がルールとして用いられている[3][4]．意味・文脈レベルの情報の抽出に関する研究として[5][6]などがある．[6]はセマンティックルールを用いることで技術文書の抄録を構成する各文章の役割を分類している．表層的ルールを用いているため，ドメインに依存せずに処理することができるが，反面，領域内容に関する判断は人間が行う必要がある．本研究ではドキュメントが表現している対象領域の分析に重点を置き，非構造ドキュメントの概念構造分析の基準として，対象領域の概念体系を表現した用語木を用いる．ドキュメントを構成する要素用語をその出現頻度とともに抽出し，ドキュメントの索引を動的に生成して「索引マップ」として表示する．利用者は可視化された索引マップを参照することでドキュメントの概念的な位置付けや情報の詳細度を知ることができる．

2．定義

2.1　用語，要素用語，用語木

ここでは「用語」と「要素用語」を区別して用いる．対象とする領域に関する専門的な語を「用語」，特定の領域又は特定の視点のもとで選択された用語を「要素用語」と呼ぶ．要素用語は領域，視点ごとに固有の概念構造を持つ．要素用語の概念構造を木構造で表したものを用語木と呼ぶ．用語木の詳細機能については次章で述べる．

2.2　索引マップ

一般に，ドキュメントの標題（タイトル）はその内容を程度の差こそあれ記述しているか，あるいは少なくとも案じしている[7]．しかし，WWWを用いた情報検索結果のように，類似したタイトルが多数列挙された検索結果の中から自分の必要とする情報を選び出すことは容易ではない．タイトルの他にキーワードやその出現位置，頻度情報を提示する機能を持つものもある[8]が，キーワードや数字の羅列で表現された従来のインデックスではドキュメントの内容を十分に把握することは困難である．

索引マップは，従来のインデックスでは表現できなかったドキュメントの概念構造を，人間が利用しやすい可視化情報として表現したものである．索引マップには，ドキュメントの対象領域，その詳細情報，及びその他の関連情報が表現される．利用者は実際に一次情報を見ることなしに，ドキュメントが表現している領域に関する情報を獲得することができる．

3　用語木

3.1　概念構造

用語木は，各要素用語間の概念関係を木構造で表したものである．要素用語の概念関係に焦点をあてて要素用語を分類し，上位語，下位語などの概念が付与されている．要素用語間の概念関係をそのまま表現するとネットワーク構造になるが，利用者に可視化表示することに重点をおき，木構造で表現した．要素用語が複数の概念を持つ場合は，概念ごとにノードを設け個々の概念ごとに構造を別管理する．概念の追加・構造の変更を用語木の部分木単位で行うことが出来るため，概念構造の管理も容易である．

3.2　情報の粒度

ドキュメントの概念分析の基準となる用語木は，基本的な用語で領域全体を広く覆う必要がある．現在は，教科書，参考書，辞書などから抽出した地球環境の制限された領域に関する基本的な名詞約1,000単語を「大分類」「中分類」「小分類」に分けて管理している．利用者に可視化表示することに重点を置くため，階層が必要以上に深くならないように注意した．

3.3　データモデルと組織化機能

用語木を用いてドキュメントの概念構造を分析するためには，ドキュメントを構成する全ての要素用語の概念を把握し，それを分析・統合

する機能が必要である．すなわち，

(1)ドキュメントを構成する全ての要素用語の上位概念を復元する機能

(2)上位概念が復元された要素用語を概念ごとに統合する機能

が必要である．次節でこれらの機能を実現するための用語木のデータモデル及び組織化機能について述べる．

3.3.1 データモデル

用語木はハイパー情報モデル[9][10]で表現されている．ハイパー情報モデルは，ある粒度の情報をフレームとして表現し，関連するノードをリンクで相互に結ぶもので，オブジェクト指向データモデルの抽象化の概念に基づき，情報の組織化機能を付加したデータモデルである．モデルは，ノードの意味的情報を表わす属性情報，概念構造全体における概念的位置情報を表わす組織情報，及び概念の組織化のための手続きで構成されている（図1）．

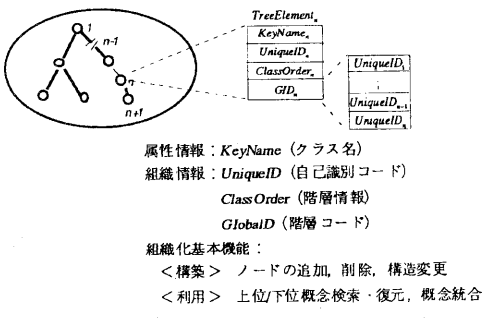

図1 ハイパー情報モデルのデータモデル

3.3.2 組織化機能

各ノードは，要素用語名を属性情報として持つ．また，組織情報として各ノード固有の自己識別コード，及び階層内における自己の位置情報を表わす階層情報と階層コードを持つ．階層コードは，各ノードの上位概念に関する情報（上位ノードの自己識別コード群）をビット情報として持つ．階層コードを利用することで，ハイパー情報モデルで表現された各ノード間には，次のような推移関係が成り立つ．

あるノード O_i,O_j,O_k の間に $O_i \prec O_j, O_j \prec O_k$ の関係があるならば

$$O_i \prec O_k$$

が成立する．ただし，「$\prec$」はノードの上位下位関係を表し，ノードO_bがO_aの上位ノードであるとき

$$O_a \prec O_b$$

と表す．従って，用語木を構成する各ノードは以下の性質を有している．すなわち，

任意のノードO_i,O_jに対し，オブジェクトO_jがO_iの上位関係（$O_i \prec O_j$）にあるならば，階層コードr_i,r_jに対し

$$r_i \otimes r_j = r_j$$

が成立する．ここで，$\otimes$はn次元ベクトルの各要素のAND演算を表す．一般に，複雑な階層構造を有するデータ構造上の推論は，あるノードの直接の上位又は直接の下位ノードを求め，さらにそのノードの上位又は下位ノードを逐次的に探索することにより行われ，データの規模が増大するにつれ非効率的である．しかし，各ノードの構造情報を階層コードを用いて表現することにより，そのノードの上位又は下位関係にある全てのノードを同時並列的に求めることが可能になる．

3.4 概念継承

用語木を構成する各ノードは，ある概念の特殊なインスタンスであり，属性情報と組織情報によって概念構造全体における位置付けがラベル付けされている．各下位のノードはより特殊なインスタンスであり，各上位のノードはより一般的なインスタンスである．各ノードをフレームとして表現することによって，各インスタンスの性質はそのインスタンスに局所的に蓄えられ，階層を降下してその子孫で伝播することができる．ノード間の重要な意味的関係が保存され，概念探索に用いられる．例えば，

要素用語名：オゾン
自己識別コード：5
階層情報：6
階層コード：0001·····

とフレーム表現されたノード「オゾン」は，階層コードの性質を用いて「地球の科学／気候・気象学／温暖化／地球温暖化／温室効果ガス」

という上位概念情報を獲得することができる．すなわち，「温暖化」領域においては，「オゾン」は「温室効果ガス」の一種であり，また「温室効果ガス」は「地球温暖化」の原因の一つであることがわかる．従って，「オゾン」が地球温暖化の原因としての性質を保持し，さらにより制限された性質を持つと推測することができる．

3.5 用語木管理

用語木の概念構造はルールではなく，人間の主観的な概念構造に基づいて構成されており，以下の手順で概念構造の管理を行う（図2）．

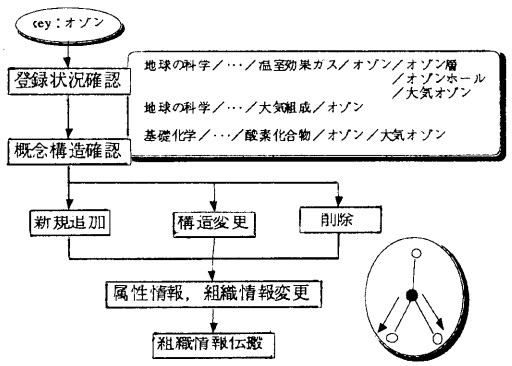

図2 用語木管理処理の流れ

(1)登録状況及び概念構造確認

用語木に新たに概念を追加しようとする際に，用語木の既存データを用いて登録の有無，登録されている場合はその概念構造を確認する．

(2)概念構造編集

確認した概念構造に基づいて，概念構造の新規追加，もしくは既存の概念構造の変更，削除を行う．

(3)組織情報伝搬

概念構造が変更されたノードの下位概念に，変更された組織情報を伝搬する．用語木編集画面を図3に示す．

4．索引マップ生成手順

索引マップは(1)要素用語抽出，(2)概念構造分析，(3)フィルタリング，の順で行う（図4）．

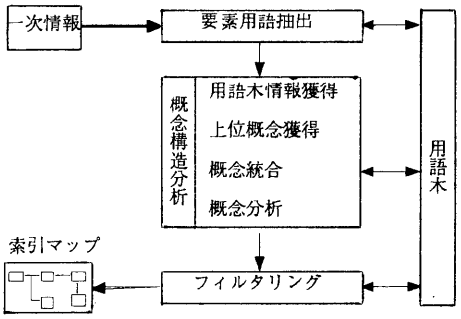

図4 索引マップ作成手順

4.1 要素用語抽出

ドキュメントのタイトル，ドキュメント作成者によってあらかじめ付与された要素用語，及びドキュメント本文から要素用語を抽出し，要素用語と出現頻度の二項関係で構成されるタグリストを生成する．出現頻度は要素用語の現れる場所（タイトル，本文など）によって重み付けを行なう．

4.2 概念構造分析

(1)用語木情報獲得：要素用語抽出機能によって抽出された要素用語を用語木上にマッピングし，対応する用語木情報（属性情報及び組織情報）を獲得する．

(2)上位概念獲得：用語木から獲得した組織情報（階層コード及び階層情報）を用いて要素用語の上位概念を復元する（図5）．

(3)概念統合：ドキュメントを構成する全ての要素用語の上位概念情報を復元し，概念構造を統合する（図5）．

4.3 フィルタリング

要素用語が複数の上位概念を持つ場合，ドキュメントが注目している概念がそのどれに当たるかを選択する機能が必要である．要素用語「オゾン」を例にとって考えてみる．現在，用語木には「オゾン」に関する3種類の概念が登録されている．「上位概念獲得」過程では，「オゾン」という属性情報のみで上位概念を復元するため，

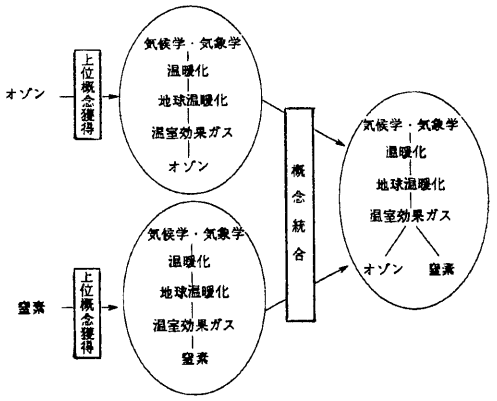

図５　上位概念獲得及び概念統合

要素用語：オゾン
自己識別コード：5
上位概念：地球の科学／気候・気象学／温暖化／地球温暖化／温室効果ガス
下位概念：オゾン層，オゾンホール，大気オゾン

要素用語：オゾン
自己識別コード：36
上位概念：地球の科学／気候・気象学／大気／組成／大気組成
下位概念：なし

要素用語：オゾン
自己識別コード：251
上位概念：基礎科学／無機化合物／酸素属化合物／酸素化合物
下位概念：大気オゾン

という，「オゾン」に関する登録された全ての上位概念が復元されてしまう（図６）．この段階ではドキュメントが注目しているのがどの概念かをまだ判断することはできない．しかし，ドキュメ
だ判断することはできない．しかし，ドキュメントを構成する全ての要素用語の上位概念の占める割合を利用することで，注目すべき概念を選択することが可能になる．

５．具体例

具体例として地球環境に関する教材「例１」について作成した索引マップ（図７）で考える．要素用語「オゾン」は索引マップ上で先の３種類の概念が復元されているが，上位概念の出現頻度情報を用いることによって，自己識別コード251，「大気組成」を上位概念に持つ「オゾン」がドキュメントの概念に最も近いものとして選択されている．強調表示されているのが頻度情報に基づいてフィルタリングを行なった結果である．これにより例１が対象としている領域は「（大分類）地球の科学／（中分類）気候学・気象学／（小分類）大気」であると解釈することができる．小分類「大気」以下のノードはドキュメントの詳細情報を示しており「ドキュメントは大気の組成としての水蒸気，窒素，酸素，オゾンについて述べている」と読み取ることができる．キーワードの羅列として表現されていた従来のインデックス（図７）では表現されなかった，ドキュメントに関するより詳細な情報が表現されていると解釈することができる．

６．評価

索引マップの評価として「ドキュメントが表現している領域」「ドキュメントの詳細情報」「関連情報」についての評価ルールを制定し，上記評価ルールに従って，地球環境に関する教材約２０件）を用いて評価を行った．「領域」は６０％をしきい値として評価を行った結果，ほぼ全ての教材の領域を判断することができた．「関連情報」として，先に述べた「オゾン」のように，複数の概念をもつ要素用語の，ドキュメントの対象領域として選択されなかった概念情報の有効性を評価した．「ドキュメントを学習に用いる場合に，他にどのような視点があるかということを常に利用者に喚起できる」という点で有効であると評価できた．「詳細情報」については「ドキュメントの詳細情報を的確に表現している」という評価ができたのは約７割であった．用語木が1,000単語という粗い概念構造であったため，「詳細情報提示」に関してはまだまだ検討する必要があるが，「概念の可視化」に関する一つの方法の可能性を示唆することができたと考える．

7．索引マップの利用

索引マップには情報検索支援の他に次のような利用方法がある．

(1)理解支援

ドキュメントの対象領域における位置付けや関連する領域との相互関係を可視的に表現する．これにより，自由探索型学習の場合，どの領域を学習しており，どのように進めば良いのかを判断できる．またシステム主導型学習の場合，関連する学習領域に関する情報を提供することができる．

(2)インデックス生成支援

索引マップを用いてシステムと会話的に索引付けを行うことにより，人手の索引付けより高速で効率的であり，自動索引付けよりは柔軟で文脈依存的なインデックス作成を行うことができる．まずシステムは，ドキュメントから自動的に要素用語を抽出し，索引マップを作成する．しかし，要素用語として登録されていない用語，例えば

「地表付近の大気の組成」→「対流圏」

といった解釈をすることが出来ない．この時に人間が，「対流圏」を索引マップ構成要素として加えることで，システムは「対流圏」という用語とその上位概念「大気圏」「中性大気」を索引マップに追加することができる．このように，システムが抽出できなかった概念を人間が補完し，また，その情報をシステムがさらに拡張（上位概念復元）することで，ドキュメント中に陽に表れていない情報を含んだインデックス付けを行うことができる．

8．おわりに

用語木を用いたドキュメントの概念構造の可視化について述べた．非構造ドキュメントの対象領域の分析において，ある用語木を基準として情報抽出を行なう場合，本稿で用いた概念構造の補完と統合化機能が重要である．用語木は情報抽出の基本となる概念構造を保ちながらも状況に応じた概念構造が提示できる機能が重要である．情報の範囲や粒度及び視点表示機能の充実に加え，より柔軟な概念表示機能の向上が今後の課題である．現段階は索引マップを作成し，その概念構造の表現を評価している段階であり，検索支援としての機能を評価するには至っていない．今後は，検索システムとの連携を行い，検索支援システムとしての評価を行う予定である．

参考文献

[1] Digital Libraries, Communications of the ACM, Vol.38, No.4, 1995.

[2] Parsaye K., Chinell M, Khoshafian S. and Wong H. : INTELLIGENT DATABASES, Johon Wiley & Sons, Inc, 1989.

[3] Cohen, R. : Analyzing the Structure of Argumentative Discourse, Computational Linguistics, Vol.13, pp.11-24, 1987.

[4] Grosz, B. and Sidner, C. L. : Attention, Intentions and the Structure of Discourse, Computational Linguistics, Vol.12, pp.175-204 1986.

[5] Liddy, E.D. and E.D. and Myaeng, S.h. : DR-LINK : Document Retrieval Using Linguistic Knowledge, ACM SIGIR Forum, Vol.26, No.2, pp.39-43, 1992.

[6] Miike S., Ito E., Ono K. and Sumita K. : A Full-Text Retrieval System with a Dynamic Abstract Generation Function, Proc. 17th ACM SIGIR, pp.152-161, 1994.

[7] Borko, H, and Bernier, Ch. L. : Indexing Concepts and Methods, Academic Press, 1978.

[8] Rao R. et.al. : Rich Interaction in the Digital Library, Communicatioons of the ACM, Vol.38, No.4, 1995.

[9] 生天目 章：ハイパー情報の自己組織化モデル，システム制御情報学会誌，Vol.7,No.7,pp.319-327, 1993.

[10] Fukunaga, M. and Namatame, A. : Self-Organizing Connectionist Hypernetworks, International Symposium on Neural Information Processing as a part of International Symposia on Information Sciences (ISKIT'92), 1992.

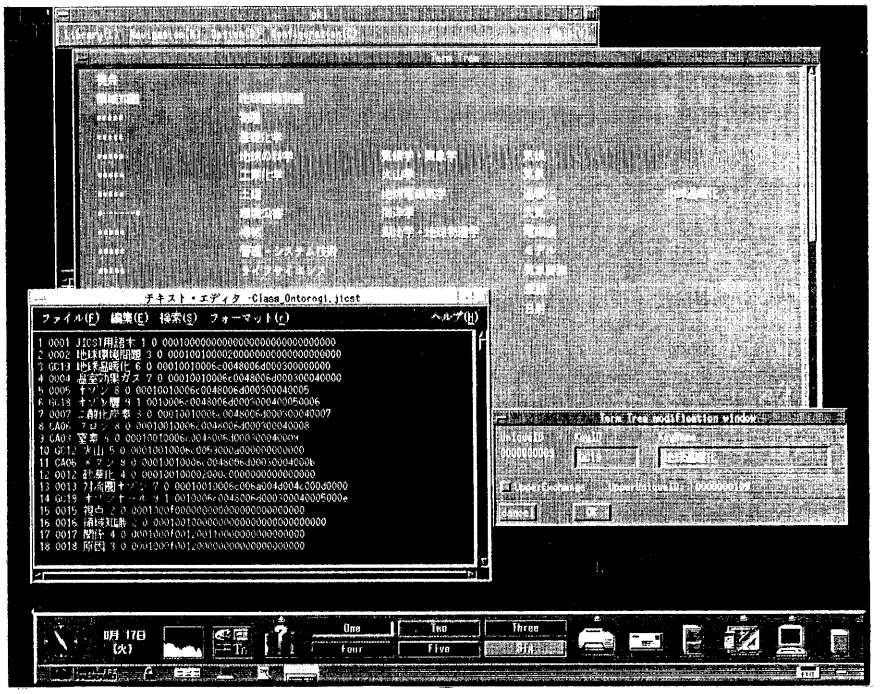

図３　用語木編集画面

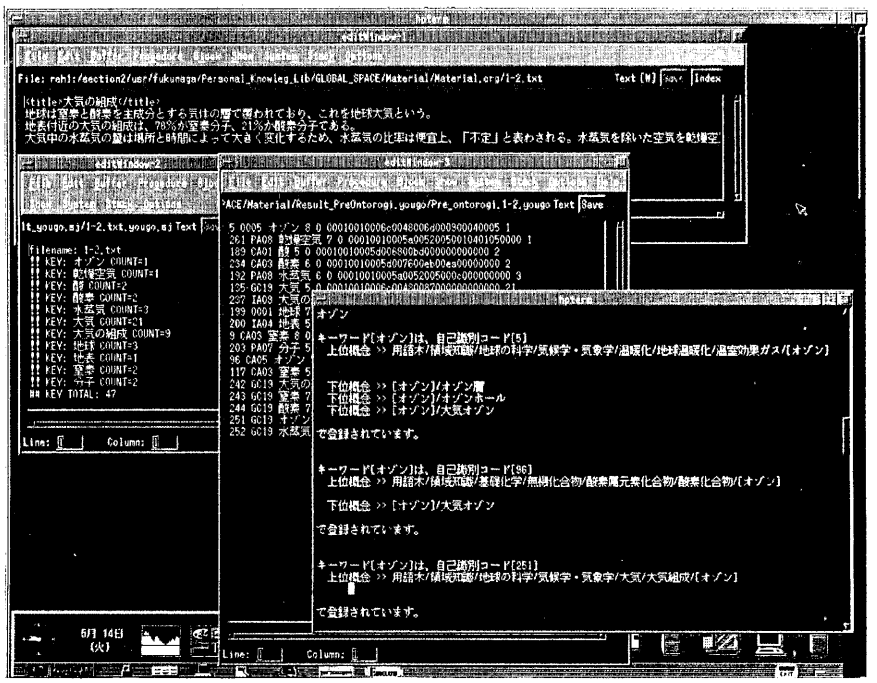

図６　索引マップ作成過程のデータ

（ドキュメントデータ，抽出された要素用語，要素用語が獲得した用語木情報，複数概念情報）

タイトル：大気の組成

タグ：大気圏、中性大気、対流圏、空気、オゾン

地球は窒素と酸素を主成分とする気体の層で覆われており、これ
大気という。地表付近の大気の組成は、78％が窒素分子、21％が酸
である。大気中の水蒸気の量は場所と時間によって大きく変化する
水蒸気の比率は便宜上、「不定」と表わされる。水蒸気を除いた空
燥空気とよぶ。地球大気の温度や圧力、数密度などの物理量の高度
る分布を標準大気モデルで見てみよう。標準大気モデルは、気圧の
ら高度を求めるときに、標準として用いるために定められた、代表
気の鉛直構造である。地球大気にはオゾンがごく僅か含まれている

オゾン	count=1
圧力	count=1
温度	count=1
鉛	count=1
乾燥空気	count=1
高度	count=2
酸	count=2
酸素	count=2
水蒸気	count=3
大気	count=27
大気の組成	count=6
地球	count=4
地表	count=1
窒素	count=2

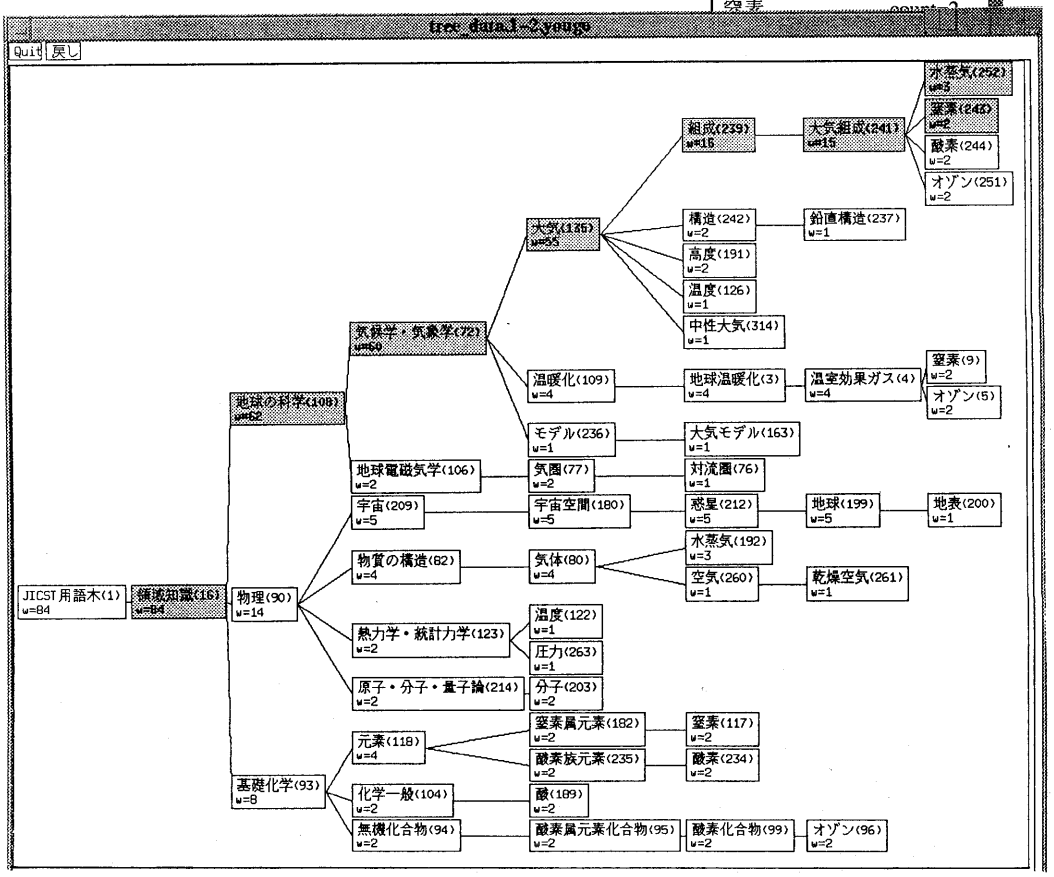

図7　索引マップ：例1「大気の組成」