

汎用シソーラスを利用した 検索用の索引メニュー構成法

千田恭子, 篠原靖志, 坂内広蔵

電力中央研究所 情報研究所 情報科学部

{senda, sinohara, bannai}@denken.or.jp

検索の初期入力を支援するため、汎用シソーラスを利用して検索対象の文章中に含まれる語を意味別に分類して提示する索引メニューシステムについて報告する。検索に役立てるために、対象領域用のシソーラスを作成し、文書中の用語をそのシソーラスの分類にそって意味別に提示するシステムがある。しかし、検索対象ごとにシソーラスをつくるのはコストがかかる。そこで我々は、「分類語彙表」という汎用的なシソーラスを利用し、ある企業の規定文書を対象に、低コストで簡便な索引メニューシステムを試作した。しかし分類語彙表の体系は索引提示に利用するために作成されていないため、そのまま利用するには、最上位のメニュー提示に適した階層がないなどの問題点があった。そこで、索引項目を二つに分け、分類の視点がより明確である項目を先に提示するなどの修正処置を行なった。本論ではそのような修正を含めた索引メニューの構築方法と、その考察を述べる。

A Construction Method of Index Menu for Information Retrieval Using a General Thesaurus

SENDA Yasuko, and SINOHARA Yasusi, BANNAI kozo

Information Science Department

Communication & Information Research Laboratory

Central Research Institute of Electric Power Industry

11-1, IWADO KITA 2-CHOME, KOMAE-SHI, TOKYO 201 JAPAN

We report a new method for making an index menu of full text database. This index menu and its presentation system supports finding keyword for the users unfamiliar to the target text database. Although, some existing full text database systems have special thesaurus developed to match their contents, it is expensive to develop different thesaurus for each databases by the experts. Hence we developed an index menu by modifying a general thesaurus "Bunrui-Goi-Hyou" to make an index of text databases in general. We discuss on the general thesaurus, modification process, and an index menu presentation system.

1 はじめに

現在、情報資源として文書データベースの構築にのりだす企業が増加している [1]。それらは、日報、企画書、マニュアル、社内規則、法律、新聞情報など、仕事に必要なあらゆる情報を文書データベースに蓄積することを目的としている。蓄積された文書データベ

スが様々な部門からアクセスされるに従い、検索対象の文書や検索に慣れていないユーザでも利用できるような、使いやすい検索機能の必要性は一層高まっている。

なじみのない文書を検索するユーザは、自分の求める情報が、どんな用語で表現されているのか、どんな

個所で言及されているのかわからない。その場合、文書の検索に適切な入力は何であるか悩みがちである。

そのようなユーザには、検索に適切な入力語句を想起、確認させるため、対象文書に含まれる語のリストをユーザの必要に応じて提示し、入力を支援する手法が有効と考えられる。本論では、シソーラスを利用して、文書中の用語を意味別に分類、提示する手法について検討した。

対象文書ごとに領域用のシソーラスを作成し、キーワードを上位語／下位語に展開して検索できるようにしたシステムや、キーワードの分類項目を階層ごとにメニュー表示して選択させるシステムがある[2, 3, 4, 5]。だが、検索対象となる文書情報が急増する昨今、領域や文書ごとにシソーラスを作成しては、大変なコストがかかる。

そこで本論では、低コストで簡便なメニューシステムの構築が期待できるため、汎用的なシソーラスを利用して、文書中の語を意味別に分類し提示する索引メニューシステムの試作について報告する。

汎用シソーラスとして国立国語研究所の分類語彙表[6]を利用した。しかし、分類語彙表など一般のシソーラスは、索引メニュー用に作成されていないため、そのまま使うには不都合があり、幾つかの点を修正して利用した。

文書データベースとしては、企業内に流通する文書の代表例として、ある特定の企業の規定文書を採用した。

本論では、2章で分類語彙表について説明し、3章で分類語彙表を利用するための問題点の修正方法と、それに基づく規定文書の分析、索引メニューの構築手順を述べ、4章で索引メニューの使用例を示し、5章で本論がとった方法について考察する。

2 分類語彙表の分析

この節では、索引メニューシステムに汎用シソーラスとして採用する分類語彙表について、概要を述べた後に問題点を分析する。

2.1 分類語彙表の概要

分類語彙表は、記載語数が約 32600 語で、日本語の汎用的なシソーラスとして記載語数が比較的多く、安価で利用できる数少ないものである。その分類体系は、大きくは「1. 体」(名詞)「2. 用」(動詞)「3. 相」(形容詞・形容動詞)「4. その他」(接統詞・感動詞・副詞)の四つに大別される。本論では、索引メニューを「1. 体」(名詞)の分類を利用した。

体(名詞)の分類は、表1で示すように最上位の第一階層で五つに分類されている。

表 1: 分類語彙表の五つの分類

.1	抽象的關係
.2	人間活動の主体
.3	人間活動 - 精神および行為 -
.4	人間活動の生産物 - 結果および用具 -
.5	自然 - 自然物および自然現象 -

この五つの分類のうち、「1.1 抽象的關係」には「人間や自然のあり方のわく組み」を指す語(例:「現象」「不在」)、「2. 人間活動の主体」には「活動の主体であるもの」を指す語(例:「妻」「銀行」)、「3. 人間活動 - 精神および行為 -」には「人間活動そのものの様相 - 精神と行為 -」を指す語(例:「安眠」「練習」)、「4. 人間活動の生産物」には「人間が直接に活動の結果として作り出した物および作り出すために利用する物」を指す語(例:「鉢」「電動機」)、「5. 自然」には「人間の主体的活動からは比較的自由に、外界として存在するもの」を指す語(例:「カルシウム」「鳥」)が分類されている。

この下の第二階層の分類項目は番号でのみ分けられ、第三階層の分類項目には分類名がついている。(下図参照)

1. 体の類

1.1 抽象的關係

- 1.100 こそあど、1.101 事柄、1.102 事項、…
- 1.1100 類・例、1.1101 等級・系列、1.1110 …
- 1.1200 有無、1.1210 出現、1.1211 復活・前兆…

第一階層の五つの分類項目の下に、第二階層の分類項目は順に 10、9、9、8、7 個あり、末端の第三階層の分類項目数は五つの分類ごとの平均で、順に 16、7、21、10、9 ある。

2.2 索引メニューに利用するための問題点

2.1 章で説明した分類語彙表を索引メニューに利用するには、以下の問題点がある。

1. 第二階層の分類項目には名前がついていないので、索引項目として提示するためには、分類名をつける必要がある。
2. 対象文書中の名詞の形態素で分類語彙表に登録されていない語があった。
3. 最上位のメニューとして提示するのに適した階層がない。その理由は以下の点である。

- (a) 分類語彙表の第一階層の五つの分類項目の名前は抽象的で漠然としすぎていて、どんな語が下位に分類されているのかわかりにくい。

- (b) 第二階層の分類数は43で、一覧するには多すぎる。一般にメニュー項目を提示する場合、項目数は一覧できる数以内におさえることが重要であると考えられる。
- (c) また、例えば第一階層の「1 抽象的関係」と「3 人間活動」はどのような視点から分類されているか明確でない。

以上より、分類語彙表の階層を索引メニューに利用するには、体系の構成を変えて提示する必要があることがわかる。

3 分類語彙表による規定文書の索引メニュー構築

3.1 問題点の回避方法

この章では、索引メニューを作成するために、2.2で述べた問題点をどのように回避するかについて述べる。

2.2の1に関して、第二階層の分類項目の名前は、その下位にある一番目の分類項目の名前からとった。これは、第三階層の0番の項目は総括的な分類とされているからである[6]。但し、それだけではどんな分類かわかりにくい場合があったため、一部、下位の高頻度の語や分類を参考に、作成者の判断で修正したり、他の分類項目名と併記しているものがある。

問題点の2については、表2に示す通り、分類語彙表に登録されている語が規定文書中の語の3/4以上を占めるため、とりあえず登録語から扱うこととする。

表2: 登録のべ語数

登録されている語	登録されていない語
23636(76%)	7284(24%)

以上の1は分類語彙表の問題であり、2は辞書によくある未登録語の問題である。

問題点の3は、分類語彙表のような汎用のシソーラスを索引メニューに適用するための本質的な問題であると考えられる。この問題に対して本論では、以下の処置をとることで改良を試みた。

1. の問題を持つ第一階層を避け、第二階層からを索引メニューとして提示する。
2. 第二階層の分類項目のうち、項目名が具体的、もしくは語を分類する視点が明確でユーザにわかりやすいと筆者が判断した「2 人間活動の主体」「4 生産物および用具」「5 自然物および自然現象」「16 時間(位置・地点・場合)*」「17 空間・場所」「31 文書・表現*」「37 資本・金

銭*」(*は作成者が修正した分類項目名)の分類項目(仮に分類Xとする)と、上記以外の分類項目(仮に分類Yとする)とをわけて、以下の提示順序をとる。

- (a) メニューの最上位で分類Xの項目だけを提示する。
- (b) 上記の項目で選択できるものがない場合のみ、その他の項目(分類Y)を提示する。

これによって、3aで言及の、漠然としたわかりにくい項目名の第一階層を避けられ、3bの第二階層の項目を二つに分けて別々に提示することでメニュー提示に適した項目数になる。

分類Xを分類Yより先に提示することについて以下に理由を述べる。

分類Xは、そのほとんどが目に見える「もの」か、分類基準が物理的に決まっているものを指す言葉の分類である。よって、項目名が具体的でわかりやすいか、細分類がユーザにわかりやすい物理的な分類になっていると判断した。それに対して分類Yは、そのほとんどが抽象的な概念を指す言葉の分類であるため、項目名が漠然として抽象的で、下位にどのような語が分類されているか想定しにくいと考えられる。また、企業の規定文書は人や物品など「もの」の扱いについて記した文書であるため、具体的な「もの」を指す言葉は比較的重要な意味を持ちやすいと考えられる。以上のことより、分類名や分類の枠組がユーザにわかりやすく、文書中でも比較的重要な意味を持ちやすい語を分類したXの項目を先に提示することにした。

この分類Xの項目が、規定文書の語のどれぐらいを占め、規定の条文にどれぐらい含まれているかを次の3.1.1章で分析する。

3.1.1 規定文書の分析

表3は、規定文書を形態素解析し、抽出した名詞の語句(一つ以上の名詞の形態素から成る語)が、分類Xの項目に登録されている形態素を含むかどうかで分類したものである。(但し、分類語彙表に未登録の形態素だけで構成される語句は除く。)その結果、規定文書の名詞の語句の約60%は分類Xの語を含むことがわかった。

表3: 規定文書の名詞語句の分析

語句の種類	異なり語数	延べ語数
分類Xを含む語句	2475(61%)	10656(58%)
分類Xを含まない語句	1587(39%)	7749(42%)
名詞語句の総数	3822	18405

また、規定文書を条文ごとに分割し、分類 X の項目の語を含有する条文の数を調べた。(条文という単位を用いたのは規定文書の項目の中で、ユーザが参照する最小の単位であると考えたためである。)その結果、分類 X の語は 677 の全ての条文に含まれていた。

以上より、規定文書を対象に、分類 X の項目を先に提示する索引メニューを構築した場合、最上位のメニュー項目から語句の 60% と条文全文を参照できることが確認できる。分類 Y の項目からは、語句の 40% しか参照できず、それ以外の語句を参照するには別の階層の項目を選択しなければならない。よって、分類 X を先に提示すれば、分類 Y を先に提示するより索引メニューを参照する効率が良いと考えられ、その有効性が確認できた。

3.2 索引メニューシステムの構築

3.2.1 システムの構成

本論の索引メニューシステムは、提示プログラム、分類語彙表をもとにしたシソーラス、企業の規定文書のデータベースからなる。分類語彙表から作成したシソーラスの索引メニューに、対象文書から抜き出した名詞の語句が割り付けられており、ユーザにはシソーラスの分類項目を階層ごとに提示して、求める情報に関連しそうな項目を選択させていくものである。このシステムをどのように構築していくかについて、以下で手順を説明する。

3.2.2 構築の手順

3.1 章で述べた修正手法を取り入れて、分類語彙表から索引メニューを構築する手順が以下である。

1. 分類語彙表の第二階層の項目に名前をつける。
2. 分類 X の項目の第二階層以下を抜き出す。
3. 抜き出した項目を並べて、最上位の索引メニューとする。
4. 最上位の索引メニューの最後尾に、「その他」という名前の分類項目をつくる
5. 分類 Y の項目の第二階層以下を抜きだし、「その他」の項目以下に並べて提示する。

次に索引メニューの末端の階層に、文書中の用語を割り付ける手順について述べる。

1. 文書を形態素解析し、名詞語句単位に用語を抽出する。(形態素解析には、JUMAN 2.0 [7] を利用した)

2. 名詞語句を、そのうちの各形態素ごとに対応する索引項目に分類する。つまり、複数の形態素からなる語は、複数の索引項目に分類される。また、もとの分類語彙表で複数の項目に登録されている形態素を含む語は、それぞれの項目に分類される。
3. 割り付ける語がなかった項目は、不要なので削除する。

用語の割り付けについて以下で例を使って説明する。例えば、「住宅資金援助」という言葉は「住宅／資金／援助」の三つの形態素から成る。よって、それぞれの形態素の意味により、「1.440 住居」「1.3721 資本・金銭」「1.365 救護・世話」の三つの索引項目に重複して分類される。また、「衣料」のように、シソーラス上で「1.410 資材」「1.4200 衣料・綿・皮・糸」の二つの分類項目に登録されている形態素は、二つの索引項目に分類される。

形態素を、単独ではなく名詞語句単位で割り付ける理由は、最下位の階層までたどって、語の一覧を参照する場合、形態素単独のものより、語句単位で表示したほうが、どのような語として使われているかわかりやすいからである。例えば、「業務」とだけ出ているよりは「業務課」のほうがわかりやすいと考えられる。

上記の手順で作成したメニュー体系を図 1 に示す(図中で * がついているものは作成者が修正した項目名)

最後尾の「.00 その他」の下位には以下の図 2 の索引メニューがある。

仮に、分類語彙表の体系をそのまま採用して第二階層から提示する索引メニューを作成すると、図 1 と図 2 の項目が入り混じっており、見にくいので、どの項目を選択すべきかわかりにくいと考えられる。

4 索引メニューシステムの動作例

この章では、作成した索引メニューの動作例を示す。

以下では、仕事上どうした場合に上司の判断や、承認印が必要なのか、自分の仕事と上司の仕事の関わりについて、規定文書の記述を調べる場合を例にとって説明する。

図 1 のメニューの第一階層から求める情報に関連がありそうな項目を探すことを想定する。上司の仕事内容に関する調査なので、仮に関連の深そうな「.24 成員・職」の項目を選択すると、図 3 の下位メニューが提示される。この階層では、選択に多少迷うが、「.243 長」を選択すると「上司」を指す「所属長」の言葉にあたり、「所属長」の職務内容について書かれた箇所を規定文書からひろって参照できる。

また、同じことを調べるために、図 2 の「.38 仕事」を選択すると、下位の索引メニューの項目は図 4 になる。

分類 No.	索引項目名
.16	時間 (位置・地点・場合) *
.17	空間・場所 *
.20	人一般 *
.21	親族
.22	相手・仲間
.23	人種・民族
.24	成員・職
.25	地域 *
.26	社会的場所 *
.27	機関
.28	同盟・団体
.31	文書・表現 *
.37	資本・金銭
.40	物品
.41	資材
.42	衣服
.43	食料
.44	住居
.45	道具
.46	機械 *
.50	性質・材質 *
.51	自然・物体・物質
.52	宇宙・地形 *
.57	身体
.58	生命・健康
.00	その他

図 1: 最上位の索引メニュー

このメニューでは、.3802～.3832 の項目は業種別という視点で分類されているのに、同階層上に.3840 以降は家事などの作業を分類したものが共存しているためわかりにくい。

仮に関連がありそうな「.3800 仕事」の項目を選び、登録されている語句をみても「仕事」「実務」「事業」といった仕事に関する一般的な語句しか参照できず、求める情報に関連がありそうな個所はなかなか参照できなくなっている。

5 考察：二つの分類の違いについて

本論では、「もの」を指す言葉の分類 X を、分類の視点が明確で、下位にどのような語や細分類あるか想定しやすい項目と考え、他の項目より優先的に提示した。この分類 X とそれ以外の項目との違いについて以下で考察する。

普通、検索の初期場面では、ユーザは文書中の言葉や文脈に関して知識をもたず、対象領域に関してどれぐらい知識を持っているのかもわからない。つまり、ユーザの知識として期待できるのは、領域や文脈に依存しないものだけである。

また一般に、汎用シソーラスは領域や文脈によらずに言葉を分類する方針で作成されている。しかし、言

分類 No.	索引項目名
.10	事柄
.11	関係 *
.12	有無
.13	様相
.14	力
.15	作用・変化
.18	形・型・姿・構え
.19	量
.30	心
.32	創作・著述
.33	文化・歴史・風俗
.34	義務・権利
.35	交わり
.36	支配・政治・革命
.38	仕事

図 2: 「.00 その他」の下位メニュー

分類 No	索引項目名
.240	成員・職
.2410	専門的・技術的職業
.2411	支配的・管理的・書記的職業
.2412	販売など
.2416	生産工程
.243	長
.244	相対的地位
.2450	臨時的地位 (役・役員)
.2452	(被告・仲人・審判・使者・持主など)

図 3: 「.24 成員・職」の下位メニュー

言葉の意味や、言葉と言葉の関係を規定するものは多様なので、領域や文脈を特定せずに語彙全体を分類するのは実は難しいことである。つまり、意味や、他の言葉との関係が領域や文脈に左右されやすい言葉があると考えられる。

文献 [8] では、言葉の意味を規定するものとして、以下の三つの関係をあげ、

1. 単語と指示物の関係
2. 単語と他の単語との関係
3. 文脈の関係

外界に指示物をもつ言葉は 1、2 が、そうでない言葉では 2、3 が意味を構成すると述べている。

これより、外界に指示物をもたない言葉 (分類語彙表では分類 Y の項目) は、領域や文脈に左右されやすいため、それを特定せずに分類しても、どのような分類なのかわかりにくくなると考えられる。

4章で例示した表図 4 は、微妙に意味が重複した項目名であり、その下位にどのような視点で語が分類分けられているかわかりにくかった。

分類 No	索引項目名
.3800	仕事
.3801	業
.3802	産業・生産
.3810	農事・営林
.3811	飼養・採取
.3820	製造工業
.3821	印刷・製本
.3822	土木
.3823	建築
.3830	運輸・交通
.3831	医療
.3832	出版・興行
.3840	家事
.3841	裁縫
.3842	洗濯など
.3843	炊事
.3844	掃除など
.3850	設備・作業・手当て・処理
.3851	練り・塗り・射ち その他
.3852	使用
.386	製造・荷造り

図 4: 「.38 仕事」の下位メニュー

一方、外界の指示物を指す言葉は、指すものの「もの」としての分類や物理的な分類に沿った視点で分類すると、領域や文脈に左右されにくく、ユーザもその視点に沿って言葉の分類をたどれると考えられる。

そのため、4章で例示した図 3 のメニューは図 4 に比べてわかりやすいメニューになっていた。

よって、検索の初期場面でユーザの知識を効率良く活用するためには、意味や他の言葉との関係が領域や文脈に左右されにくい言葉 (分類語彙表では分類 X の言葉) の利用が有効ではないかと考察した。

6 まとめと今後の課題

本論では、分類語彙表という汎用的なシソーラスを利用し、文書中の語を意味別に分類した検索用の索引メニューシステムの試作について報告した。

索引メニューに利用するための分類語彙表の主要な問題点は、同一階層の分類の視点が明確でないことや、項目数、項目名の関係でメニューの最上位として提示するのに適した階層がないことだった。

そこで本論では、項目数の多い第二階層の項目から「もの」を指す言葉の分類項目を抜き出し、他の項目より先にメニュー提示することを試みた。

そして、「もの」を指す言葉は、意味や他の言葉との関係が領域や文脈に左右されにくい言葉であり、検索の初期場面でのメニュー提示には有効ではないかと考察した。

また、今後の課題として以下の点があげられる。

分類語彙表という汎用のシソーラスを利用したにもかかわらず、第二階層につけた項目名を多少修正することで、作成者の主観的な判断が多少入ってしまった。今後は自動的に名前を付与する方法 (下位項目のうち出現頻度の高い項目名を幾つか併記するなど) を検討する必要がある。

また、現行のシステムでは分類語彙表に未登録のものは、索引メニューには載っていない。シソーラスの未登録語の扱いについて (分類語彙表の対応する分類にわりふるなど)、検討する。

現行では、索引メニューの末端の項目に割り付けた語句は、項目に対応する形態素ごとにまとめて一覧提示しているが、よりみやすい提示方法を検討する。

第二階層の項目を分ける基準となった、目に見えるまたは物理的な分類基準のある「もの」を指す言葉の分類であるか否かを、本論では作成者の主観で行ったが、客観的に定義することは今後の課題である。

また、他の汎用シソーラスを調査して、本論で行った手法に関しての考察を深めることも今後の課題である。

謝辞

本論文の執筆にあたり、御指導下さった電中研情報研人工知能グループのメンバーに感謝致します。

参考文献

- [1] 広がる文書データベース. 日経コンピュータ, No. 374, pp. 134-147, 1995.
- [2] 日本経済新聞社. 日経全文記事データベース - 検索の手引 [第二版]-. 1984.
- [3] 浦本直彦. Information outlining- 検索情報の可視化 - 行政情報の活用のために. 情報処理学会情報学基礎研究会研究報告, 第 95 巻, pp. 49-56, 1995.
- [4] 河合真宏, 中馬高彦, 櫛木好明. 知的処理機構を用いた知的文書管理機能. 情報処理学会第 37 回全国大会, pp. 1934-1935, 1988.
- [5] 橋本肇, 赤瀬幸夫, 辻洋. ソフトウェア常識集 ir システム socks(1). 情報処理学会第 35 回全国大会, pp. 1505-1506, 1987.
- [6] 国立国語研究所 (編). 「分類語彙表」. 国立国語研究所資料集 6. 秀英出版, 1964.
- [7] 松本裕治, 黒橋禎夫, 宇津呂武仁, 妙木裕, 長尾真. 日本語形態素解析システム J UMAN 使用説明書 Version 2.0. 奈良先端科学技術大学院大学 松本研究室, 1994.
- [8] 国広哲也. 意味論の方法. 大修館書店, 1982.