情 報 学 基 礎 21－7
（1991. 5. 28）

文献情報データベースからの知識獲得

石塚英弘，宇陀則彦*)，山本毅雄
図書館情報大学図書館情報学部

知識獲得の一手法として，文献情報データベースからの知識獲得法について報告する．我々は，論文に書かれた知識を『報告された知識』と言い，専門家に認められた知識を『判断された知識』と言う．格文法に類似した手法を用いれば，CASearchの一般事項索引から報告された知識を獲得することができる．ここで，CASearchは詳細な索引付けで定評のある，化学分野の網羅的な文献情報データベースである．判断された知識の方も報告された知識から自動的に抽出し，専門家の判断を経て知識ベースに格納することができる．

我々はこの手法を用いて，報告された知識ベース，判断された知識ベース，それに知識ベース管理システムとから構成される知識ベースシステムを開発した．

Knowledge Acquisition from Bibliographic Database

Hidehiro Ishizuka, Norihiko Uda and Takeo Yamamoto
University of Library and Information Science
1-2 Kasuga Tsukuba-shi, Ibaraki 305

The present paper reports an approach to knowledge aquisition from a bibliographic database to overcome knowledge acquisition bottlenecks. We call knowledge writen in an article as "reported knowledge" and knowledge authorized by an expert as "judged knowledge". Using a computational linguistic technique which is similar to a case grammar, reported knowledge can be acquired from a general subject index(GSI) of CASearch, which is a comprehensive bibliographic database in chemistry, and famous for detailed indexing. Judged knowledge can be obtained through automatic extract from reported knowledge and through judgment by an expert.

We use the approach reported in this paper, and develop a knowledge-based system containing reported-knowledge-base, judged-knowledge-base and knowledge base management system.

---

*)現在は筑波大学大学院博士過程在学中

1. はじめに

知識ベースシステムの問題点として知識獲得の難しさが指摘されている．その解決を目指して種々の研究，たとえば，専門家へのインタビューによる知識獲得，専門書からの獲得，データベースからの獲得，帰納推論や演繹推論を用いる学習，類推，モデルを用いる定性推論，事例ベース推論等々の研究が活発に行われているが，知識獲得は依然として大きな課題である．

また，大部分のシステムでは『知識ベース』が対象とする『知識』とは『専門家が持つ知識』『既に確立し終わった知識』であって，知識ベースの内容が専門家が持つ知識と同じになれば最早変更されることはないと考えられているようである．しかし，実際には知識の内容はその分野の研究の進展によって変更されていくし，学界の定説も変更されることがある．

たとえば，ある高温超伝導物質の発見はそれ以前に定説となっていた理論モデルに反することであったし，その後に次々と高温超伝導物質が発見されたことにより，その現象を説明できる新しい理論の構築が始まった．また，環境問題において以前は『炭酸ガスは汚染物質ではない』とされていたが，炭酸ガスが地球温暖化の原因物質となることが判明してからは汚染物質とされている．

新しい知見を得ることが研究の主目的である以上，学問の進展に伴って知識が変化することはむしろ当然と言えるだろう．分野の専門家は情報検索などを行って，この種の新たに報告された知見を取り入れ，自分の持つ知識を変更していく．実際に各分野で毎年大量の論文が報告されている．たとえば，化学では年間約45万件に上るという．知識ベースシステムとしても，大量の新規報告の中から必要なものをシステマティクに取り込む仕組みを用意する必要がある．

そこで，我々は新しく発見された情報や知識に対応できるシステムとして，文献情報データベースから知識を獲得して知識ベースシステムを構築することを研究している.[1,2]ここでは，知識を『報告された知識』と『判断された知識』とに分けて考える．前者は論文に発表された知識であるが，これは文献情報データベースから獲得できることを示す．また，判断された知識は専門家に確認された知識である．これは報告された知識から抽出して，専門家の判断を経て知識ベースに格納することができる．このようにして開発したシステムについて報告する．

2. これまでの知識獲得手法の概観

ここでは，先に挙げた知識獲得手法の内，後に示す我々の手法との関連があるものについて述べる．

専門家へのインタビューによる知識獲得の研究は数多く見られる．しかし，それらによって開発された知識獲得ツールを用いても，専門家の持つ経験的知識を体系的に獲得することは容易ではない．

書籍からの獲得としては，人手で百科辞典の記述から知識ベースを作る巨大プロジェクトCYC[3]がある．また，専門書から自然言語理解の手法によってテキストを解析して知識を獲得するアプローチもある．これには，lexical functional grammarを用いてドイツ語の文を解析し，Prologの節に変換して知識ベースを構築するもの,[4] ハードウェア・マニュアルから自動的にモデルを構築するもの[5] などがある．しかし，現時点では文脈解析に困難な点があるため，限界がある．

データベースのデータは個々の事実を明確に記述しているため，専門家や専門書から知識を獲得する場合の問題は生じない．そのため，データベースのデータを知識ベース用に変換する研究がなされている．たとえば，リレーショナル・データベースのテーブルを命題・述語論理型知識の形式に変換して知識ベースを構築したもの[5]，テーブルをフレームに変換して知識ベースを構築したもの[7]などがある．そのほか，予め用意した知識抽出用ルールを用いて，データベースから診断用の知識を抽出する試み[8]もある．

しかし，データベースのデータは，表現は明示的であるが，セマンティクスは隠れている．そのため，データベースのデータから知識ベースを作るためにはセマンティクスを明示的にしなければならないが，そこに困難がある．

モデルを用いる定性推論[9]は深い知識を表現しており，得られる結論の信頼性も高い．しかし，深い知識の獲得は経験的知識をルールの形で獲得するよりも難しいという欠点がある．

事例ベース推論[10,11]は，過去の類似する事例と類推を用いて問題を解決する推論である．事例は事例ベースに蓄えられ，これが知識ベースとなる．事例の獲得は，経験的知識を体系的に獲得することに比べれば容易という長所がある．

3. 文献情報データベースに埋め込まれた知識

文献情報データベースには，著者名，文献タイトル，雑誌名，巻号頁のほかに文献の内容を示す主題索引や抄録が付いている．主題索引や抄録は，1)その内容が自然言語で書かれていること，2)そのため，セマンティクスが暗示的に示されていることの2点で通常のデータベースの項目と異なっている．

我々はCASearchのGeneral Subject Index(以下，GSI と略す）に注目した．ここでCASearchとはアメリカのChemical Abstracts Serviceが作っている，化学分野では最も信頼されている文献情報データベースである．GSI は単に名詞だけを並べたキーワード索引とは異なり，前置詞を含んだ英文のフレーズ(phrase)になっている．その理由は，まとまった意味が表現できるように，中心となる主題の名詞と関連する名詞とを前置詞によって接続してあるからである．また，GSI には抄録に書かれていないような詳しい知見も入っている．そこで，文脈解析の点で問題のある抄録の意味解析を行うよりも，GSI の解析を行う方が有効であると考えた．

ただし，GSI は元々印刷体の索引として作られたため，フレーズ中の重要語を見出し語に出す必要があり，残りの部分はカンマで区切られたサブフレーズが並んだ形式(articulated subject index）になっている．そのため，元のフレーズを復元してから解析する必要がある．GSI の例を図1に示す．また，これを復元して得たフレーズを図2に示す．なお，復元のアルゴリズムは前報[1]に示した．

ここで，Agropyron repensとはイネ科の雑草であり，herbicide は除草剤である．そこで，このフレーズを翻訳すれば『多年性の穀物の中に生えるAgropyron repensの抑制のための除草剤』となる．そして，この索引は，ある文献の単なる索引というばかりでなく，『多年性の穀物の中に生えるAgropyron repensを抑制する除草剤がある』という知識を得る元とすることもできる．

GCH: Agropyron repens
GTM: control of, in perennial grass crops, herbicides for

図1 articulated subject index形式のGSI
GCH: 概念見出し語. GTM: 説明語句. 見出し語を取り除いた残りの部分.

herbicides for control of Agropyron repens in perennial grass crops

図2 復元したフレーズ形式の索引

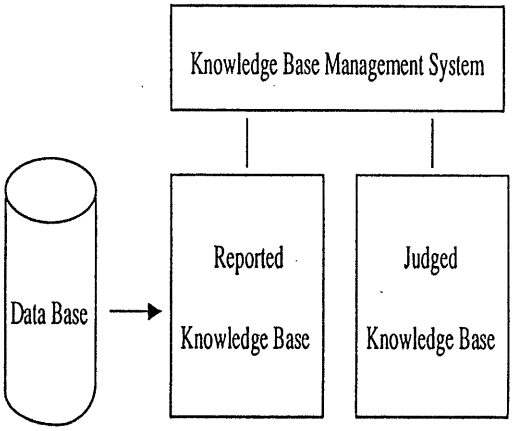

図3 システムの構成

4. 報告された知識と判断された知識

我々は，知識を『報告された知識』と『判断された知識』の2つに分けて捉える．報告された知識とは文献情報データベースから抽出して得た知識である．また，判断された知識とは専門家の判断を経た知識であり，報告された知識から必要な部分を抽出して得ることもできるし，別に外から与えることもできるようにした．図3のシステム構成図に示すように，我々の知識ベースは報告された知識と判断された知識の2つに分かれている．また，2つの知識ベースを管理する知識ベース管理システムがあり，後に述べるような知識の獲得，推論，2つの知識ベースの使い分けなどの機能を持っている．システムは Sun-3ワークステーション上で，ISAC社のK-Prologを用いて作成した．

4.1 報告された知識の知識ベース

報告された知識の知識ベース（以下，報告された知識ベースと略す）は，文献の書誌事項を

納めたbib-frame, GSI から作成したgsi-frame, CSI(Chemical Substance Index)から作成したcsi-frame の３つのフレームで構成されている. gsi-frame は一文献に複数存在するので, 書誌事項はbib-frame にまとめ, インヘリタンスによって参照できるようになっている. なお, いま現在で, 報告された知識ベースに収録されている文献数は 966, gsi-frame の数は3202である.

gsi-frame のスロットの名称とその意味を表１に示した. instrumentからrelationまでは, subject との関係（格）を示している. 格文法[12]の格は述語との関係を表現するのが普通であるが, 文で中心となるのは述語であるのに対して, GSI ではsubject であるため, subject との関係を格とした. また, 格文法ではinstrumentは述語で示される動作を実行するための道具であるが, 我々のシステムでは, 主題を得るための道具や手段とした. その理由はその道具や手段を使って或る主題を得るからである. なお, 論文によっては或る操作, 手法, 概念に内在する問題を主題に選ぶことがある. 主題の存在するものが操作や手法である場合はmicro range に, 概念の場合はmacro range とした. また, GSI では主題との関係を明示するように努めているが, 論文の中で関係が明確になっていない場合もある. この場合は『ＡはＢと何らかの関係がある』としか言いようがないので, GSI では A is relation to B" と表現している. この場合, Ｂをrelationとすることにした.

表１ gsi-frame のスロットの名称と格判定用のマーカ

スロット名	スロットの内容	前置詞	意味マーカ
ca#	文献番号, bib-frame とのリンク	−	−
subject	中心となる主題	−	−
instrument	（主題を得るための）道具, 手段	by, with	−
time	時間	after, before	−
purpose	目的	for	−
place	場所	at, in, on	point
source	出所, 材料, 原料	from, in, on	material
micro range	（主題が存在する）操作, 手法	in, on	job
macro range	（主題が存在する）概念	in, on	concept
relation	関連	in, on	relation

なお, 先に述べた例, "herbicides for control of Agropyron repens in perennial grass crops" の場合は, subject はherbicides, purpose はcontrol of Agropyron repens, source はperennial grass crops となる.

このように格を判定して格納することにより, たとえば, 『除草剤 Roundupは, イネ科の植物である Agropyron repens に有効か？』という問題が与えられた時, 知識ベース管理システムは次の手順で答えることができる.

1) Roundupをsubject, Agropyron repensをsourceとして, また格を変えて
Agropyron repens をsubject, Roundupをinstrumentとして, 報告された知識ベースに問い合わせる.

2)該当する知識を持つgsi-frame が探索される.

3)結果を出力する.

4.2 報告された知識の獲得

報告された知識ベースは，文献情報データベースのGSI から知識ベース管理システムの中のプログラムによって作成される．まず，既に述べたように GSIから元のフレーズを復元する．次いで，得られたフレーズについて格文法に類似した処理を行って格を求め，gsi-frame のスロットに格納する．

格文法では格の判定法として，

a)格特有の前置詞

b)意味マーカの利用

c)動詞によって接続する格が異なることの利用

があるが，本研究ではa)とb)の方法を用いる．GSI には動詞がないため，c)は適用できないからである．我々のシステムでは，個々の名詞ごとに意味マーカを書いた格判定用辞書を用意した．辞書にない名詞については，その殆どが化学物質であるため，意味マーカはmaterialとして処理する．

スロットへの割り当て方は以下のようにして行う．まず，マーカとなる前置詞から次のマーカとなる前置詞の直前までを一単位として取り出す．なお，"of"は前置詞ではあるが，マーカとはならないので区切りには使用しない．取り出した単位について，その前置詞が，by，with，for，after，before，atの場合は前置詞だけでスロットが決定できるので，該当するスロットに割り当てる．前置詞が in，on，from の場合はこれのみでは決定できないので，表1に示した意味マーカによって判定する．単位の中に複数の名詞がある場合は，優先順位の高い意味マーカによってスロットに割り当てる．優先順位は，relation，point，concept，job，material の順である．

4.3 判断された知識の知識ベース

判断された知識の知識ベース（以下，判断された知識ベースと略記する）は概念と概念の関係を表現している．用意している関係は今のところ，階層関係を表す isaと，原因と結果の関係を表すcause の2種類である．ここでは，個々の概念が一つのフレームを形成しており，スロットによってフレーム間の関係を示している．isa は階層構造を形成するが，cause もあるため，全体としてはネットワークになっている．

与えた isaの例としては，Elymus sibiricusとElymus repens とAgropyron repensがイネ科(Graminate) であること，また RoundupとUtalとatrazineは除草剤であることなどがある．なお，概念を指定すれば，それに関係する概念とその関係を出力する機能もシステムに用意した．

判断された知識ベースは，たとえば次のような場合に使用する．『除草剤 Roundupは，イネ科の植物である Elymus sibiricus に有効か？』という問題が与えられた時，知識ベース管理システムは次の手順で答えることができる．

1)報告された知識ベースを探索する．格の指定の仕方は4.1 で述べた方法と同じ．

2)報告された知識ベースには該当するものが存在しない.

3)関連知識を用いて探してみる旨のメッセージを出力する.

4)判断された知識ベースによって，Elymus sibiricusをElymus repens とAgropyron repensで置き換えて，報告された知識ベースを再探索する.

5)『Roundup はElymus repens とAgropyron repensに有効である』旨の答えを出力する.

4.4 判断された知識の獲得

判断された知識を知識ベースに格納する方法として，

A)問題解決を通した動的な知識獲得

B)専門家による知識の入力

の2つを用意した.

A)は，問題解決の過程で獲得する方法で，言わば動的な獲得である．実際には，次の手順で行う.

1)問題解決のための知識を報告された知識ベースから抽出する.

2)重複チェックを行い，判断された知識ベースに存在しないもののみを出力する.

3)専門家にこれを見せて，判断された知識ベースに格納するか否か指示してもらう.

4)格納を指示されたもののみ格納する.

たとえば，『greenhouse effects（温室効果）を引き起こす原因は何か』という問題を与えると，以前に述べたようにして，報告された知識ベースからfluorochlorohydrocarbons, chlorofluoro hydrocarbons, hydrocarbons, carbon dioxide の4つの化学物質が温室効果の原因である旨の答えを出力する．そして，重複チェックをした後，ユーザである専門家の指示を仰いで，格納することを指示した知識のみ，判断された知識ベースに書き込む．この場合は，cause 関係として書き込む.

なお，階層関係のisa はB)の方法で格納することが多い.

判断された知識は，一部に報告された知識と重複するデータを持つが，報告された知識とは別の世界を形成する．なぜなら，判断された知識には或る観点，たとえばgreenhouse effectsと原因物質という観点で抽出し判断された知識のみが納められているが，報告された知識ベースの方には，greenhouse effectsと原因物質が納められた格以外のスロットにも値が納められている．また，原因物質の方は別の概念とも関係を持つ可能性があり，互いに個別に存在する報告された知識とは別の世界となるからである.

5.まとめ

以上述べたように，本報告では文献情報データベースから報告された知識を獲得できること，報告された知識から更に濃縮した判断された知識を獲得できること，この2つの知識を使い分け，推論を行うことによって，能率的な問題解決が可能なことなどを示した.

文献

1)石塚英弘，王忠清，山本毅雄：化学文献情報データベースの知識ベース化－設計と試作，1989年情報学シンポジウム講演論文集，pp.53-61(1989).

2)宇陀則彦，石塚英弘，山本毅雄：文献情報データベースを情報源とする知識ベースシステム，情報処理学会・データベースシステム・人工知能合同研究会発表資料，p.89-97(1990.11).
3)Lenat,D.B, Prakash,M. and Shepherd,M.: CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks, *AI Magazine*, Vol.6, No.4, pp.65-85(1986).
4)Welner,F., Feyle,U. and Rohrer,C.: Automatic construction of a knowledge base by analysing text in natural language, *IJCAI'83*, pp.727-729(1983).
5)Nishida,T., Kosaka,A. and Doshita,S.: Towards knowledge acuisition from natural language documents, *IJCAI'83*, pp.482-486(1983).
6)小口琢夫，近藤秀文：知識ベース管理システム　知識ベースとデータベースの融合方式，情報処理学会第31回全国大会論文集，pp.1003-1004(1985).
7)森原一郎，牛島浩一，小野寺尚文：KBMSにおけるDB/KB 変換方式，情報処理学会第38回全国大会論文集，pp.571－572(1989).
8)柳吉洙，志村正道：故障診断用エキスパートシステムにおける知識獲得，人工知能学会誌，Vol.1，No.1，pp.93-100(1986).
9)総説としては，情報処理，Vol.32，No.2(1991)や，人工知能学会誌；Vol.4，No.5(1989) の特集があり，また書籍では，淵一博（監修），溝口文雄，古川康一，安西祐一郎（編）：定性推論，共立出版(1989)がある.
10)Hammond,K.J.: Case-based planning, Academic Press, 1989, 277p.
11)Bareiss,R.: Exemplar-based knowledge acquisition, 1989, 169p.
12)Bruce,B: Case systems for natural language, *Artificial Intelligence*, Vol.6, pp.327-360(1975).