論　文

柔軟な問い合わせのための弱構造化表現*

花川 賢治** ・武田 英明*** ・西田 豊明***

Weakly Structured Representations for Flexible Queries*

Kenji Hanakawa** ・Hideaki Takeda*** ・Toyoaki Nishida ***

Most information systems always require us to represent our information in a formal and strict way. This lack of flexibility reduces the usefulness of the information systems. In this paper, we introduce two weakly structured representation techniques for easily constructing and retrieving ill-organized information. The first technique is a word-list-query which is a free ordered sequence of words used to represent user's information needs. The second is an associative graph which is a kind of semantic network consisting of concept nodes, IS-A links and unnamed horizontal links. To parse a word-list-query, we use a concept reduction method which is recursive rewriting of words according to information represented in associative graph. This integration of parsing and information retrieval disambiguates word-list-queries. We implemented an experimental system using these techniques, and show that this system is easy for unskilled user and its accuracy is reasonable.

1. はじめに

近年のパーソナルコンピュータ，ネットワークなどの情報機器の急速な普及に伴い，情報科学について特別な知識を持たない非専門家ユーザが身近な情報をコンピュータで処理する機会が非常に増えてきた．このようなコンピュータの日常化が進むにつれて，情報システムは従来とは異なった性質を持つようになった．それは以下のように三種類に大別できる．

(1) 利用目的・利用形態の多様化

情報システムは，かつては組織の定型的な業務にのみ利用されていたが，現在では情報機器のコストが低下したために特定の用途を想定しないで導入して個人のレベルで利用することが可能になった．また，情報機器の処理能力が増大し，応用範囲は飛躍的に広がった．その結果，情報システムの利用目的と利用形態は非常に多様化した．たとえば，現在急速に普及している WWW の利用目的はビジネス，教育，娯楽など非常に多彩である．

(2) 開発形態の多様化

非専門家向けシステム構築ツール (たとえば，HyperCard，スプレッドシートなど) の進歩に伴い，システムエンジニアが設計を行ない，それに基づいて組織的に開発するという方法に代わって，エンドユーザが自ら情報システムを構築することが可能になった．その結果，エンドユーザが採りがちな，全体像が明らかでない状態でも可能な部分から着手するというようなボトムアップの手法による開発形態が定着しつつある．

(3) 情報そのものの多様化

情報の種類が少なく，情報システム間の相互関係が単純なときには，コンピュータで処理しやすいように情報の表現形式を統一することは比較的容易であった．しかし，現在のように情報源が多様化すると，それぞれに合わせて形式を変換するのには莫大な労力がかかる．また，すべての情報の種類について表現形式を標準化するのも

* 原稿受付　年月日
** 大阪府立工業高等専門学校 電子情報工学科 Department of Electrical Engineering and Computer Science, Osaka Prefectural College of Technology; 26-12 Saiwai-cho, Neyagawa, Osaka 572, JAPAN
*** 奈良先端科学技術大学院大学 情報科学研究科 Graduate School of Information Science, Nara Institute of Science and Technology; 8916-5 Takayama, Ikoma, Nara 630-01,JAPAN
Key Words: knowledge representation, semantic network, natural language interface, knowledge based system, user-frendly interface

困難である．その結果，情報が情報源の形式 (たとえば，自然言語で書かれた文書) のまま格納されるようになり，情報システム内で多様な形式が共存するようになった．

本稿ではこのような性質に対応するために，情報を弱構造化して表現するアプローチを提案する．ここでの構造化の強さとは，情報の表現形式における規程の厳密性，複雑性の度合いのことを指す．弱構造化とはそれを意図的に低下させることである．

構造化の強さという観点でみると，現在電子化されコンピュータに蓄えられている情報は非常に幅が広い．たとえば，自然言語で書かれた文書ファイルはコンピュータからみると 1 次元の文字列であり，極めて弱い構造化による情報である．それとは対照的に，リレーショナルモデルや一階述語論理などによるデータベースは厳密な意味論に基づく表現形式を採用した強い構造化による情報であるといえる．これらの間に位置するものとして，表形式，タグを含む文書など人間とコンピュータが共に利用可能な比較的単純な形式化を行った情報がある．

一般に，情報を強構造化する程，複雑な情報処理を効率良く実行することが可能になる．たとえば大量の情報からユーザが欲しい情報を探し出す処理を考えた場合，文書ファイルに対するテキスト検索[6]よりもリレーショナルデータベースのほうがより複雑な検索条件を指定することができ，大量のデータを高速に処理することもできる．しかしながら，強構造化する程，情報の構築と利用に情報科学の知識，大きな労力，対象となる情報の高い整合性が要求されるという傾向もある．したがって，このトレードオフの関係において，構造化の強さの最良の点をみつけることが重要になる．

従来から情報の組織化は強い構造化の方法が主流であり，体系化された専門領域での応用では大きな成果を収めて来た．ところが，上に示した性質を持つ日常的な情報処理においては，強い構造化がもたらす問題がより顕著に現れるため，ある程度弱い構造化の方が適していると考えることができる．

そこで，本稿では，弱構造化の表現として単語列と連想グラフを提案し，それを事実的知識を蓄えユーザの簡単な質問に答えるシステムに応用することについて述べる．単語列とは一般的な人工言語とは対照的な，文法的制約のない自由な順序で単語を並べた質問の表現形式である．連想グラフは，データベースを記述するための，リンクに名前をつけない点に特徴がある，ある種の意味ネットワークである．単語列と連想グラフは，多様なユーザが容易に使用することができる非常に簡単な表現形式である．

単語列による質問は構造的な曖昧性を持つが，データベースに存在する具体的な情報を利用することにより，部分的な解消が可能である．本稿では，概念縮約法という質問の解析とデータベースの参照を統合的に行う手法により，そのような曖昧性を解消することについて述べる．さらに，旅行情報を題材にしたデータベースシステムを作成し，これらの手法の有効性を確認する．

2. 日常的な情報を強構造化することの問題点

従来の情報システムのほとんどは専門分野を対象とし，高度な機能と高い品質を目指したものであった．そのために，一階述語論理，リレーショナルモデルのような構文と意味論が厳密な表現が採用されてきた．一般に，このような方法による情報システムの構築は専門的な知識を持った人間による緻密な手作業によるため，莫大な量の労力とコストが要求されてきた．

一方，我々が日常的な場面で必要とする情報システムは，それほど高度な機能と品質は要求されない．それよりも，対象となる情報が乱雑で，論理的一貫性を維持するのが困難な点や，ユーザが情報科学の知識を十分に持ち合わせていないことによる問題の解決が望まれる．

日常的な情報処理に，従来の手法を適用したときの問題点を以下に示す．

(1) 表現の構造に一貫性が要求される．

複雑な情報を表現する構造は幾通りも考えられ，どの構造を採用するかはデータベースの設計者に任されている．一旦構造が決定されると，情報の構築者とユーザは常にその構造を頭に入れておいて，それに従って情報の登録や問い合わせを行う必要がある．また，一度決定された構造を変更することは容易ではない．

(2) 対象となる情報が整理済みで完全であることが要求される．

従来のトップダウンの構築手法はあらかじめ情報が整理済みで完全であることを前提としているが，日常的な情報は未整理で不完全な場合が多く，それを適用するのは難しい．

(3) 不自然なオブジェクトの命名が必要になる．

一階述語論理，リレーショナルモデル等に基づきデータベースを構築すると，我々が直観的に持っている概念に対応しないオブジェクトが多数発生する．データベースの構成要素の名前には自然言語の単語を使用するのが一般的であるが，全てのデータベースベースのオブジェクトに自然な一対一対応の名前を付けるのが難しい．

実際に，一階述語論理に基づく KL-ONE 系の知識表現言語である LOOM[1)2)4)] を用い，奈良の観光情報を知識ベース化する作業を行ったところ，多くの困難さに直面

(a) Definition of geographical things and "Todaiji is 500m north of Nara-park"

```
(defrelation is-in :range area  :domain thing)
(defconcept geographical-relation
  :is (:exactly 1 distance))
(defconcept directed-geographical-relation :is
  (:and geographical-relation  (:exactly 1 direction_)
      (:exactly 1 from)      (:exactly 1 to)))
(defrelation distance
  :range integer
  :domain geographical-relation)
(defrelation direction_
  :range direction :domain directed-geographical-relation)
(defrelation from
  :range location  :domain directed-geographical-relation)
(defrelation to
  :range location  :domain directed-geographical-relation)
(tellm (:about between-Todaiji-and-Nara-park
               directed-geographical-relation
               (distance 500m) (direction_ north)
               (from Nara-park) (to Todaiji)))
```

(b) "Deers in Nara-park."

```
(defconcept deer :is-primitive animal)
(tellm (deer deer-1))
(tellm (is-in deer-1 Nara-park))
```

Fig. 1 Representation in LOOM

した．その一例を Fig. 1 に紹介する．

(a) は地理に関する概念の定義と"東大寺は奈良公園の 500m 北にある"という事実を表現したものである．人工言語の構文と意味論に厳密に従う必要があるために，このような単純な事実を示すことでさえ，情報科学の専門知識を持たない人間にとっては非常に難しい．

(b) は奈良公園に鹿がいることを表現したものである．実世界では鹿の個体には名前は付けられていないが，LOOM では概念と個体は厳密に区別され後者にもユニークな名前を付けることが要求されるので，*deer−1* というような不自然な名前が必要になった．

以上のような問題点の原因の一つに，情報の表現形式が複雑で厳密すぎることが考えられる．そこで，厳密性と複雑性を低下させ，自然でわかりやすい表現を行うこと，すなわち弱構造化表現が必要になる．

3. 質問の弱構造化表現

3.1 単語列による質問

情報の弱構造化表現を応用した具体的なシステムとして，事実的知識を検索するシステムについて考える．一般ユーザが容易にこのようなシステムを利用できるようにするには，以下のことが重要になる．

- 問い合わせのための難解な人工言語を覚える必要がない．
- データベーススキーマをあらかじめ知っている必要がない．

これを目的とした研究は古くからデータベースインターフェースの分野で盛んに行われてきた．それらで提案されている質問の表現方式には，自然言語，メニュー，アイコン，記号列などがある．このうち記号列は[8)7)]は，データベースを構成する記号を自由な順序で並べたもので，質問を弱構造化表現したものであると考えることができる．本稿でも，これと類似した単語列による質問の表現を採用する．

単語列とは，一般的な人工言語で書かれた文とは対照的な以下のような特徴を持つ文である．

- 単語の順序が自由である．
- 人工言語特有の定義がされている単語を含まない．
- 単語が属するカテゴリが設定されない．

一般的な問い合わせ言語と同様の複雑な検索条件が指定できるように，単語列でも，入れ子構造，すなわち質問文の部分が質問文とみなせる構造を許す．ただし，入れ子構造は意味的なものであり，それを明示的に表現するための構文的な手段は設けない．

単語列は，付属語などを省略したインフォーマルな自然言語の文に似ている．これを用いれば，ユーザは非常に小さな労力で質問を行うことができる．

3.2 情報レベルでの単語列の曖昧性解消

質問の単語列を構成する単語は一対一でデータベースのオブジェクトと対応づけ，語義の曖昧性はないものと仮定する．しかしながら，このように仮定しても，単語列の質問は必ずしも意味を一意に特定することはできない．たとえば,「太郎 愛する 女性」という質問には,「太郎が愛する女性」と「太郎を愛する女性」のどちらを答えればよいのか，判別できない．さらに，単語列の質問では入れ子構造を明示的に表現しないので，それに関する曖昧性も発生する．

このような曖昧性を解決するために，現実には質問を行う人は質問の対象について既にある程度の情報を持ち合わせていることが多いということに着目する．すなわち，コンピュータのデータベースに対し質問が行なわれるときの条件について次のような仮定を導入する．質問者は質問の対象について部分的な情報を持っているが，完全な情報は持っていない．データベースには質問の対象についてのより完全な情報が格納されていて，質問者はそのことを知っている．当然データベースに格納されている情報は質問者があらかじめ持っている部分的な情報を包含する．つまり，質問が行なわれる際には，質問者とデータベースは部分的に情報を共有する．

上記のような仮定は多くの場合現実にあてはまる．たとえば,「太郎が愛する女性の勤めている会社」という意味で「太郎 愛する 女性 会社」という質問が行なわれた