# 3G−9　情報検索におけるサーチャの知識を用いた自然言語からの検索式生成

谷 幹也　市山 俊治　久保加奈子　会森 清

日本電気(株)

## 1　はじめに

データベース(DB)技術の発展、普及にともない、情報検索を日常的な業務としていないユーザによる検索が増加している。しかも、その検索は従来に比べ非定型かつ複雑なものになりつつある。従来、大規模DBや商用DBの検索はその情報検索専門家(以下サーチャ)がユーザとの対話によってユーザの情報要求の明確化・検索式の生成を行なっていた。ユーザはDB内のスキーマ情報、DBの内容の領域情報、DB検索の知識を欠いていることが多い。ユーザが直接DB検索を行なうような場合、サーチャが持っているような知識を代用するようなシステムが必要となる。本論文では、サーチャがDB検索に果たしている役割をまとめ、その一部の知識を使用して自然言語の検索質問文からキーワード検索のための検索論理式を作成するアルゴリズムを提案する。

## 2　サーチャの役割と検索式の生成

サーチャはユーザとの対話を介しながら、ユーザが欲しい情報(情報要求)を明確化して、DBの語彙で情報検索のための検索式を作成する。その時にサーチャが果たす役割とその知識は図1の通りである。

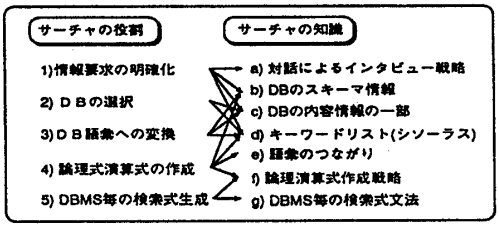


図1　サーチャの役割と知識

このサーチャの役割の中でも4)論理演算式の作成を計算機で行なうために、自然言語の名詞句からの論理検索式の生成をサーチャの知識の一つであるf) 論理演算式作成戦略を用いて行なう。

## 3　自然言語からの検索式自動生成の問題点

キーワード検索において自然言語の名詞句を検索式に変換する場合問題となるのは1) 並列のスコープ、2) 不要語の削除、3) 並列演算子の種類があげられる。ここでは話を簡単にするために名詞句を「と」と「の」のみによって構成するものとする。

### 3.1　並列のスコープ

名詞句の並列の範囲は、その入力文の構造だけでは判断できない。図2の1),2)では名詞、助詞の並びは等しいものの並列の範囲が異なっている。1)では「バス」に対する並列要素は「電車」であり、2)では「単車」の並列要素は「自動車」である。この違いは、それぞれの語彙の関連度情報から決定できる。

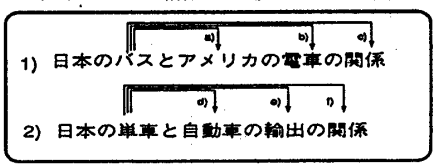


図2　並列のスコープ

### 3.2　不要語の削除

入力文中の名詞の中には直接検索式に反映しないほうが良いものが存在する。図3の例のように、この入力文に対する検索式には、「日本電気」と「関係」は陽に現れない。「日本電気」は、「PC-9801」のために不要となるがキーワードとして必要になる場合もある。これはシソーラスの情報から削除することが出来る。また、「関係」の方は、入力の自然言語文に現れるがキーワード検索式上ではキーワード間の論理演算子を決定することにのみ関係し、検索式に陽に現れない語彙である関係語である。

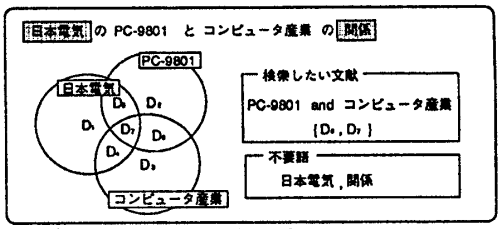


図3　不要語の削除

### 3.3　並列演算子の種類

並列になっている語彙動詞の論理接続はその語間の関連度による。した図4のように、1)に対する演算子は「or」であり、2)に対する演算子は「and」と決めることが出来る。これも各キーワード間の関連度から決定できる。

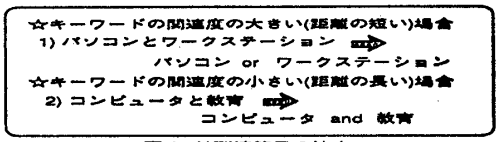


図4　並列演算子の決定

## 4　検索式の自動生成方式

前章で述べたような問題点を解決するために必要な知識としては、a) シソーラス、b) 語間の関連度、c) 関係語のリストの3つがあげられる。このうち、シソーラスは商用データベースがほとんど備えているものであり、使用することを前提とする。この b),c) はDBに最初から存在するものではない。

---

Generation fo Query Equations from Natural Language by Searcher's Knowledge

Mikiya Tani, Shunji Ichiyama, Kanako Kubo and Kiyoshi Emori

*NEC* Corp.

### 4.1 シソーラス

商用データベースのシソーラスに通常記述される語彙の関係は以下の通りである。

(1) 階層関係：一般的階層、全体 - 部分、例示 etc.

(2) 等価関係：同義語で登録されるもの関係 (例)「計算機」と「コンピュータ」

(3) 関連関係：上の2つの関係に入らないものの語彙の間に関連の強いもの。研究領域 - 研究対象、行為 - 動作主、物 - 作成者などの関係を表す。

### 4.2 語彙の関連度とシソーラス上の距離

語彙の関連度を計るために、本論文ではシソーラス上の距離を導入する。シソーラス上の3つの概念 $A, B, C$ に対して、$A, B$ 間の距離 $d(A, B)$ と $B, C$ 間の距離 $d(B, C)$ の比較は、$A, B$ の共通の上位概念 $C_{AB}$ と $B, C$ の共通の上位概念 $C_{BC}$ の関係により以下のように定義する。

1) $C_{AB}$ と $C_{BC}$ が同じ。 $d(A, B) = d(B, C)$ (図5.a)

2) $C_{AB}$ が $C_{BC}$ の上位概念。 $d(A, B) > d(B, C)$ (図5.b)

3) 上位下位関係でない。 決定不可能 (図5.c)

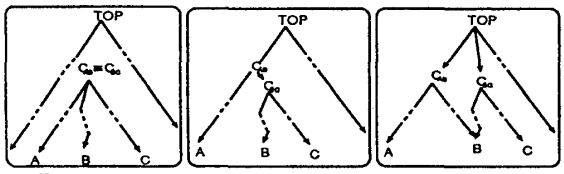

図5.a 図5.b 図5.c

1) の場合や 3) の場合はそのままでは距離を計ることは出来ない。この場合、共通の関連語彙の数、共通の上位概念・下位概念が多いほど距離が近いとする。

## 5 実現手法

前章で定義したシソーラスとシソーラス上の距離を用いて入力名詞句から検索式を生成する手順を示す。「の」と「と」で連結された名詞句は以下のように表すことが出来る [田村 87]。

「$A_1$ の‥の $A_n$ と $B_1$ の‥の $B_m$ と $C_1$ の‥の $C_o$」

手順の概要は次のようになっている [山形 91]。主要な部分についてのみ以下に述べる。

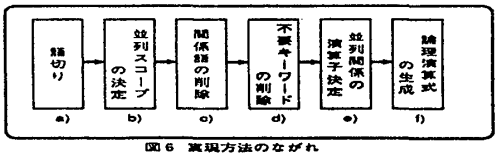

図6 実現方法のながれ

b) 並列スコープの決定「と」による並列要素を次のように決定する。

1. 並列要素の一方を「と」の直前のものとする。
2. 「と」の直後から次の「と」迄の名詞を並列要素の候補 $(B_j)$ に対して次式を満たすような $B_k$ を求める。
$$\min_i d(A, B_i) = d(A, B_k)$$

(a) k=m の場合、同じ処理を $B_m$ の直後の「と」について繰り返す。

(b) k≠m の場合、$(A_1$ の.... の $B_k)$ に関して $(B_1$ の‥の $B_k)$ と並列概念をなす $(A_l$ の‥の $A_n)$ となる l を探す。この決定は、$B_1$ から $B_k$ の概念の数と $A_1$,‥ $A_n$ の中で $B_1$ からの距離が最小となる $A_i$ の二つから求める。

c) 関係語の削除 名詞句の中に3章で述べた関係語の削除を行なう。関係語はその前にある並列句の論理関係に影響を与えるかどうかで表1のような2つのカテゴリにわかれる。

表1 関係語の機能と動作

| 種別 | 機能 | 動作 |
|---|---|---|
| 関係語 | 関係語直前の並列関係を and にする | 関係語と「の」を削除し、「と」に関係Aマーキングをする。 |
| 関係語B | 関係語直前の並列関係に影響を及ぼさない。 | 関係語と「の」を削除 |

d) 不要キーワードの削除 「AのB」という構造に関して次のルールでキーワードAの削除を行なう。

1) シソーラスに複数のBが存在する ⟶ A and B
2) Bの複数の上位概念が存在しその一つがA ⟶ A and B
3) Bの存在するTREEの上位にAが存在 ⟶ B
4) Bの関係語にAがある。 ⟶ B
5) その他 ⟶ A and B

図7 不要キーワードの削除

e) 並列関係の決定 並列関係の演算子の決定は「並列要素の距離」と「関係Aマーク」で行なわれる。関係Aマークがあれば「and」接続、そうでない場合、並列要素を構成する要素 $A, B$ の共通の概念が TOP から1レベル以内であれば「and」接続 $(d(A, B)$ が大)、違うなら「or」接続 $(d(A, B)$ が小) と決定する。

## 6 例

次のような例文を考えてみる。

「日本の東京の学校図書館と公共図書館の利用」

DBのシソーラスの一部と関係語を図8、表2に示す。前節のアルゴリズムを適用すれば、次の論理演算式が正しく得られる。

「東京 ·(学校図書館 + 公共図書館)」

表2 関係語の例

| タイプ | 語彙 | タイプ | 語彙 |
|---|---|---|---|
| 関係語A | 関係、比較、違い、..... | 関係語B | 利用、応用、影響、..... |

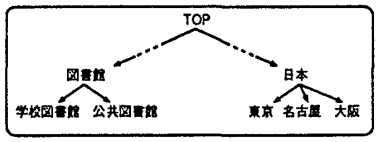

図8 シソーラスの例

## 7 おわりに

サーチャの論理演算式作成のための知識を用いて、シソーラスの意味距離計算による並列スコープの解析、並列演算子の決定と不要なキーワードの削除情報検索を特長とする検索式生成アルゴリズムを提案した。本方式はシソーラスのみを用いているため、語間の関係の種類などからの関連度を出すことが出来ず距離自体が完全に正確なものではない。今後は、多数の例に対して評価を行なうと共に、その中で距離の定義を洗練する予定である。

## [参考文献]

[山形 91] 山形、会森、川西、谷、：検索式自動生成の一考察，ドクメンテーションシンポジウム,Vol.21,1991

[田村 87] 田村直良、田中穂積：意味解析に基づく並列名詞句の構造解析，情処自然言語処理研究会,Vol59,No.2