Library and Information Science No. 9 1971

二次情報ファイルにおけるキーワードの自動変換

Automatic Conversion of Keywords in the Secondary Information File

菊　池　敏　典

Toshinori Kikuchi

Résumé

In information retrieval it is sometimes better to use a vocabulary not coincident with that used for indexing. Here comes the problem of automatically converting a vocabulary of magnetic tape storage of documents indexed with keywords into that retrieval language.

First we need a conversion dictionary indicating correspondence between index language and retrieval one. By the use of this dictionary, following two methods of conversion are possible:

1. Keyword-for-keyword method, i.e., individual keywords in storage file are converted into those of retrieval file. As to this method conditions of the dictionary and feasibility of automatically preparing the dictionary come into question.
2. Document-for-document method, i.e., a set of keywords in a given document is converted into a set of retrieving keywords. In other words, this is a method based on statistics of frequency of occurrence of keywords in sample documents.

The storage file should preferably have a vocabulary controlled by thesaurus, indexed from various points of view, and represented with specific descriptors as possible. However, it is difficult to convert storage file satisfactorily into retrieval file by any single method of automatic conversion. Consequently combination of several conversion methods, that of automatic conversion and automatic indexing, human revision after conversion, etc. are required.

Standardization of thesauri is important in order to facilitate auto-conversion as well.

菊池敏典：日本科学技術情報センター電子計算機室調査役.
Toshinori Kikuchi, Staff, Computer Division, Japan Information Center of Science and Technology.

V 総　括
参考文献

要　旨

文献検索では，索引に使用した語彙とは一致しない語彙で検索する方がよい場合がある。そこで，文献を索引してキーワードを蓄積した磁気テープファイルの語彙を，検索用の語彙に自動的に変換する方法が問題となる。

自動変換には，索引語彙から検索語彙への対応を与えた変換辞書が必要である。この変換辞書を利用して，次の2種類の変換法が考えられる。

1. 蓄積ファイルの個々のキーワードを，検索用ファイルの個々のキーワードに変換する，キーワード単位の変換法。この方法では，変換辞書の条件，変換辞書の自動的作成の可能性などが問題となる。
2. 文献ごとに，キーワードの集合をキーワードの集合に変換する，文献単位の変換法。これは，標本文献におけるキーワードの出現に基く統計的な方法である。

蓄積ファイルは，語彙がシソーラスにより統制されていること，いろいろな観点から索引されていること，詳しいデスクリプタで記述されていること，が望ましい。しかしながら，特定の自動変換法により，満足できる検索用ファイルに変換することは困難である。したがって，いくつかの自動変換法の組合わせ，自動変換と自動索引の組合わせ，変換後に人間による修正などが必要となる。

シソーラスの標準化は，自動変換を容易にするためにも必要である。

I 序　論

現在のほとんどの文献検索システムは，キーワードの組合わせによる方式である。この方式では，システムで使用するキーワードをシソーラスで制限することにより，システムの効果を高めることができるとされている。すなわち，一般の情報検索論では，シソーラスを索引者と検索者を直結する手段として捕えているので，索引者と検索者が同一のシソーラスを使用することがたてまえになっている。

しかしながら，電子計算機による文献検索では，文献のキーワードファイルをそのファイルとは異なるキーワードの語彙で検索しなければならない場合が，今後発生すると思われる。

そのいくつかの例を次に記す。

(1) シソーラスの改訂

改訂前のシソーラスで索引されたキーワードファイルを，改訂後のシソーラスのキーワードで検索する。

(2) シソーラスの翻訳[1)]

ある言語のシソーラスで索引されたキーワードファイルを，別の言語のシソーラスのキーワードで検索する。

(3) キーワードファイルの統合[2)]

互に異なるシソーラスで索引された複数のキーワードファイルを統合して，どのファイルのキーワードも特定の検索用シソーラスのキーワードで検索する。

特に，二次情報が磁気テープファイルとして盛んに流通するようになると，索引者と検索者との間の語彙の食い違いの問題が表面化すると思われる。すなわち，ファイル提供機関は不特定多数の利用者集団を想定した平均的シソーラスで索引を行なうので，このシソーラスが特定の利用者集団を対象とする個々のファイル利用機関にとって最適とは限らない。たとえば，Hammond, W.[3)]は，二つの情報システム間の両立性 (compatibility) を，〝両システムに共通の主題範囲について，ある情報システムが，他の情報システムの元の索引データと抄録データを受用しうる能力〟と定義した上で，DDC ファイルと NASA ファイルの間のキーワードの両立性を論じている。

このような場合，対象文献について改めて人間が検索用シソーラスにより再索引を行なうのは，労力や費用の点から得策ではない。そこで，電子計算機により元のファイルから検索用シソーラス向きのファイルへ，キーワードを自動的に変換する方法が考えられる。

現状では，自動変換が成功する可能性は少ないと思われる。しかしながら，これは自動変換を無視する理由にはならない。むしろ，自動変換を可能にする条件をさがし出して，そのような条件を満足するようなシソーラスおよび索引への努力を行なうことが必要であろう。

そのためには，まず自動変換の性質を明らかにすることが必要である，という観点に立って，自動変換の一般的な分析を試みる。

II 準　　備

A. 定　義

本論文では，ファイル，シソーラスなどの用語は次の意味で使用する。

ファイル

文献検索用に作成したキーワードファイル。自動変換を受けるファイルを原ファイル，原ファイルから自動変換により作成されたファイルを目標ファイルという。

シソーラス

ファイルの作成または利用において，必要な概念を特定のキーワードで表示するために，キーワードの概念を規定した辞書。シソーラスでは，個々のキーワードに対してその関係語を表示することにより，キーワードの概念を相対的に規定している。シソーラスの語彙はデスクリプタと USE 参照語とから成る。

原ファイルの作成に使用するシソーラスを原シソーラス，目標ファイルの利用に使用するシソーラスを目標シソーラスという。

デスクリプタ

シソーラスの語彙のうち，ファイルにおいて概念を表示するために使用されるキーワード。

原シソーラスのデスクリプタを原デスクリプタ，目標シソーラスのデスクリプタを目標デスクリプタという。

USE 参照語

シソーラスの語彙のうち，ファイルではデスクリプタで代置されるキーワード。個々の USE 参照語の代りに使用するデスクリプタは，シソーラスで与えられる。

変換辞書（辞書）

原シソーラスの語彙から目標シソーラスの語彙への対応を与えた辞書。

自動変換（変換）

変換辞書を利用して，原ファイルのデスクリプタを目標ファイルのデスクリプタへ機械的に置換する操作。

B. 前　　提

本論文では，次の基本的な前提を設ける。

(1) 原ファイルのキーワードは，かならず原シソーラスのデスクリプタであり，かつ原シソーラスのデスクリプタは，かならず原ファイルのキーワードとして使用されている。

原ファイルのキーワードが無統制の用語ならば，不確定な語彙の統制という自動索引の問題となる。

もし変換の対象となる原ファイルに存在しないデスクリプタが索引用シソーラスに存在するならば，そのようなデスクリプタを除外したシソーラスを原シソーラスとする。

(2) 目標ファイルのキーワードは，かならず目標シソーラスのデスクリプタであり，かつ目標シソーラスのデスクリプタは，かならず目標ファイルの検索に使用される。

(3) 原シソーラスと目標シソーラスは異なる。

もし原シソーラスを検索に使用するならば，変換が不要なことは明らかである。ただし，説明の便宜上，原シソーラスと目標シソーラスが同一である変換を恒等変換として扱う。

(4) ファイル中の，ある文献に与えられたデスクリプタの集合には，同一のデスクリプタは含まれない。

C. 記　　号

本論文では，デスクリプタ，シソーラスの語彙などに次の記号を使用する。

$x_1, x_2, \cdots, x_m$: 原デスクリプタ

$y_1, y_2, \cdots, y_n$: 目標デスクリプタ

$X=\{x_1, x_2, \cdots, x_m\}$: 原シソーラスの語彙

$Y=\{y_1, y_2, \cdots, y_n\}$: 目標シソーラスの語彙

$X^* \subset X$: 原ファイル中の，ある文献に与えられたデスクリプタの集合

$Y^* \subset Y$: 目標ファイル中の，ある文献に与えられたデスクリプタの集合

Γ : 原シソーラスの語彙から目標シソーラスの語彙への対応

Γ^{-1}: 目標シソーラスの語彙から原シソーラスの語彙への対応

III デスクリプタ単位の変換法

A. 概　　要

この方法の一般的な特徴を次に記す。

(1) 変換辞書として，個々の原デスクリプタから個々の目標デスクリプタへの対応の有無を表示した辞書を使用する。

(2) 変換は次のように行なう。まず原ファイル中のデスクリプタを変換辞書を引いて対応のある目標デスクリプタに置き換える。次いで，文献単位に同一デスクリプ

タの重複があれば，重複を削除する

すなわち，この方法では，変換辞書において $\Gamma X=Y$ ならば，個々の文献のデスクリプタの変換では $\Gamma X^*=Y^*$ となる。

B. 変換の形式

デスクリプタ単位の変換では，変換辞書の対応形式が変換の形式を支配する。そこで，まずこの関係を検討する。

1. 変換辞書形式の分類と定義

デスクリプタの対応範囲と対応関係とにより，辞書を次のように分類・定義する。

(1) 対応範囲

全対応

すべての原デスクリプタは，少なくとも一つの目標デスクリプタに対応する。同時に，すべての目標デスクリプタは，少なくとも一つの原デスクリプタに対応する。

すなわち，$\Gamma X=Y$，かつ $\Gamma^{-1}Y=X$。

全：部分対応

すべての原デスクリプタは，少なくとも一つの目標デスクリプタに対応する。ただし，原デスクリプタと対応しない目標デスクリプタが存在する。

すなわち，$\Gamma X\subset Y$，かつ $\Gamma^{-1}Y=X$。

部分：全対応

すべての目標デスクリプタは，少なくとも一つの原デスクリプタに対応する。ただし，目標デスクリプタと対応しない原デスクリプタが存在する。

すなわち，$\Gamma X=Y$，かつ $\Gamma^{-1}Y\subset X$。

部分対応

目標デスクリプタに対応しない原デスクリプタ，および原デスクリプタに対応しない目標デスクリプタが存在する。

すなわち，$\Gamma X\subset Y$，かつ $\Gamma^{-1}Y\subset X$。

(2) 対応関係

1：1 対応

二つ以上の目標デスクリプタに対応する原デスクリプタは存在しない。同時に，二つ以上の原デスクリプタに対応する目標デスクリプタは存在しない。

すなわち，すべての $x_i\in X$ について，$\Gamma x_i\in Y\cup\phi$。かつ，すべての $y_j\in Y$ について，$\Gamma^{-1}y_j\in X\cup\phi$。

ここに，空集合を ϕ で示す。

多：1 対応

二つ以上の目標デスクリプタに対応する原デスクリプタは存在しない。ただし，二つ以上の原デスクリプタに対応する目標デスクリプタは存在する。

すなわち，すべての $x_i\in X$ について，$\Gamma x_i\in Y\cup\phi$。かつ，すべての $y_j\in Y$ について，$\Gamma^{-1}y_j\subset X\cup\phi$。

(3) 対応関係

1：多対応

二つ以上の原デスクリプタに対応する目標デスクリプタは存在しない。ただし，二つ以上の目標デスクリプタに対応する原デスクリプタは存在する。

すなわち，すべての $x_i\in X$ について，$\Gamma x_i\subset Y\cup\phi$。かつ，すべての $y_j\in Y$ について，$\Gamma^{-1}y_j\in X\cup\phi$。

多：多対応

二つ以上の原デスクリプタに対応する目標デスクリプタ，および二つ以上の目標デスクリプタに対応する原デスクリプタが存在する。

すなわち，すべての $x_i\in X$ について，$\Gamma x_i\subset Y\cup\phi$。かつ，すべての $y_j\in Y$ について，$\Gamma^{-1}y_j\subset X\cup\phi$。

2. ファイルデスクリプタ数

変換に伴ない，ファイル中のデスクリプタ数は一般に増減する。デスクリプタ数としては語彙量（異なりキーワード数）と総数（延べキーワード数）の2種類がある。デスクリプタ数の増減と辞書形式との関係は次表のようになる。

(1) 全対応または全：部分対応

デスクリプタ数 ＼ 対応関係	語彙量	総数
1：1 対応	不変	不変
多：1 対応	減少	増加しない
1：多 対応	増加	増加
多：多 対応	不定	不定

(2) 部分：全対応または部分対応

デスクリプタ数 ＼ 対応関係	語彙量	総数
1：1 対応	減少	減少
多：1 対応	減少	減少
1：多 対応	不定	不定
多：多 対応	不定	不定

3. 変換性

原ファイルと目標ファイルの間のデスクリプタの相互変換の可能性を変換性とし，復元性と可逆性を定義し変

換性を保証する辞書形式を検討する。

(1) 復元性

目標ファイルの作成に使用した変換辞書を使用した逆変換により，目標ファイルから原ファイルが復元できるならば，その変換辞書による変換は復元性がある，ということにする。

すなわち，復元性のある変換では，すべての $X^* \subset X$ について，$\Gamma^{-1}\Gamma X^* = X^*$ である。

復元性のある変換を可能にする変換辞書にとって，必要かつ充分な条件は次のようである。

全対応または全：部分対応であり，かつ 1：1 対応または多：1 対応であること。

(2) 可逆性

ある変換辞書による変換と逆変換とにより，二つのファイルのどちらを原ファイルとする変換においても原ファイルが復元できるならば，その変換辞書による変換は可逆性がある，ということにする。

可逆性のある変換を可能にする変換辞書にとって，必要かつ充分な条件に次のようである。

全対応であり，かつ 1:1 対応であること。

C. 検索の正確さ

一般に文献検索では，必要であるにもかかわらず検索されない文献（検索もれ）がある一方，不要であるにもかかわらず検索される文検（ノイズ）がある。検索もれとノイズの少ない検索を正確な検索ということにすれば，自動変換により検索の正確さがそこなわれては困る。したがって，変換辞書では，概念が一致するデスクリプタ間には対応が与えられ，概念が一致しないデスクリプタ間には対応が与えられてはならない。しかしながら，一般には異なるシソーラス中のデスクリプタが表現する概念の一致関係は，一致するかしないかという二者択一ではなく，どの程度類似しているか，ということである。したがって，変換辞書ではたかだか，概念が類似しているデスクリプタ間には対応が与えられ，概念が類似していないデスクリプタ間には対応が与えられてはならない，ということでしかない。

そこで，類似概念のデスクリプタ間の対応が問題となる。

1. 関係語の分類と定義

便宜上，デスクリプタを，それが指示する概念の内包を表現する事物を要素とする集合名とみなす。あるデスクリプタ x_i に対して類似概念のデスクリプタを y_j とすれば，y_j は x_i の関係語といい，関係語を次のように分類・定義する。

関係語

x_i と y_j との間に共通部分が存在する $(x_i \cap y_j \neq \phi)$ とき，かつそのときにのみ，y_j は x_i の関係語，y_j は x_i と関係がある，という。

同義語

x_i が y_j と等しい $(x_i = y_j)$ とき，かつそのときにのみ，y_j は x_i の同義語，y_j は x_i と同義である，という。

上位語

x_i が y_j の部分である $(x_i \subset y_j)$ とき，かつそのときにのみ，y_j は x_i の上位語，y_j は x_i の上位である，という。

下位語

y_j が x_i の部分である $(x_i \supset y_j)$ とき，かつそのときにのみ，y_j は x_i の下位語，y_j は x_i の下位である，という。

関連語

y_j が x_i の同義語，上位語，下位語でない関係語 $(x_i \cap y_j \neq \phi,\ x_i \neq y_j,\ x_i \not\subset y_j,\ x_i \not\supset y_j)$ のとき，かつそのときにのみ，y_j は x_i の関連語，y_j は x_i と関連がある，という。

y_j が x_i の上位語，下位語，関連語であることを $x_i BT y_j$，$x_i NT y_j$，$x_i RT y_j$ と記すこともある。

2. 関係語への変換

原デスクリプタ x_i または $\{x_i, x_j, \cdots, x_l\}$ から，それと関係のある目標デスクリプタ y_j への対応を変換辞書で与えたとする。この辞書による変換後に，y_j で目標ファイルを検索したとき，概念の不一致による検索の正確さの低下は次のようになる。

(1) 同義語への対応 $(x_i = y_j)$

x_i と y_j の概念は一致するので，変換は検索の正確さに影響を与えない。

(2) 上位語への対応

y_j を x_i とのみ対応させたとき。

x_i が与えられなかった文献のうち，$y_j - x_i$ の概念を含むものは検索もれとなる。一方，ノイズはない。

y_j を $\{x_i, x_j, \cdots, x_l\}$ と対応させたとき

$x_i, x_j \cdots, x_l$ のどれもが与えられなかった文献のうち，$y_j - (x_i \cup x_j \cup \cdots \cup x_l)$ の概念を含むものは検索もれとなる。一方，ノイズはない。したがって，このときは x_i とのみ対応させたときと比較して，ノイズは不変で

あるが，検索もれは減少する。

(3) 下位語への対応

y_j を x_i とのみ対応させたとき。

x_i が与えられた文献のうち，x_i-y_j の概念を含むものはノイズとなる。一方，検索もれはない。

y_j を $\{x_i, x_j, \cdots, x_l\}$ と対応させたとき。

$x_i, x_i, \cdots, x_l$ のどれかが与えられた文献のうち，$(x_i \cup x_j \cup \cdots \cup x_l)-y_j$ の概念を含むものはノイズとなる。一方，検索もれはない。したがって，このときは x_i とのみ対応させたときと比較して，検索もれは不変であるが，ノイズは増加する。

(4) 関連語への対応

y_j を x_i とのみ対応させたとき。

x_i が与えられなかった文献のうち，y_j-x_i の概念を含むものは検索もれとなる。一方，x_i が与えられた文献のうち，x_i-y_j の概念を含むものはノイズとなる。

y_j を $\{x_i, x_j, \cdots, x_l\}$ と対応させたとき。

$x_i, x_j, \cdots, x_l$ のどれもが与えられなかった文献のうち，$y_j-(x_i \cup x_j \cup \cdots \cup x_l)$ の概念を含むものは検索もれとなる。一方$x_i, x_j, \cdots, x_l$のどれかが与えられた文献のうち，$(x_i \cup x_j \cup \cdots \cup x_l)-y_j$ の概念を含むものはノイズとなる。したがって，このときは x_i とのみ対応させたときと比較して，ノイズは増加するが検索もれは減少する。

3. 変換辞書における概念対応

したがって，変換辞書では，原デスクリプタから目標デスクリプタへの対応は次のようであることが望ましい。

なお，NOT 検索，すなわちあるデスクリプタを含まない文献を求める検索での検索もれとノイズ関係は，そのデスクリプタで検索したときの逆となる。しかしながら，実際に NOT 検索を行なうことはまれなので，無視する。

(1) 同義の目標デスクリプタが存在する原デスクリプタは，かならずその目標デスクリプタへの対応が与えられ，それ以外の目標デスクリプタへの対応が与えられていない。

(2) 目標シソーラス中に同義のデスクリプタは存在しないが上位のデスクリプタが存在する原デスクリプタは，かならず上位の目標デスクリプタへの対応が与えられている。

(3) 下位または関連の目標デスクリプタへの対応が与えられている原デスクリプタは，目標シソーラス中に同義のデスクリプタは存在しない。ただし，下位または関連への対応の存否，またはその程度は，検索の正確さとして検索もれとノイズのどちらを重視するかによる。すなわち，ノイズよりも検索もれを重視するならば，変換辞書での下位または関連への対応は増加する。

(4) 2 個以上の上位の原デスクリプタからの対応が与えられている目標デスクリプタは存在しない。

D. 変換辞書の作成法

個々のデスクリプタの概念は，シソーラスで規定されているので，シソーラスに基いて変換辞書を作成することができる。すなわち，原シソーラス中のすべてのデスクリプタについて，一つずつ目標シソーラスから関係のあるデスクリプタをさがし出して対応を与える方法である。

この方法を人間が手作業で行なうことも不可能ではない。特に語彙の小さいシソーラスについて，どちらかのシソーラスを熟知した人が作成するのは比較的容易であろう。しかしながら，語彙の大きいシソーラスを対象とした変換辞書を人間が手作業で作成するのは相当の苦労が予想される。そこで，電子計算機を利用した変換辞書の作成法の開発が必要となる。

ただし，III. C. 3. で述べた条件を満足する変換辞書を全く自動的に作成することは一般には不可能である。そこで，機械処理と人間の判断との組合わせ方が問題となる。

次に，変換辞書の作成に電子計算機を利用する二つの方法について検討する。

1. 人間機械相互交信による方法

Auerbach Corp.[4] は変換辞書を作成するアルゴリズムを開発した。このアルゴリズムには，次の特徴がある。

(1) 原デスクリプタに対して，ある目標デスクリプタが同形異義語か否かなどのデスクリプタ間の概念の一致性は人間が判断する。電子計算機処理の過程で人間の判断が必要になれば，どういう判断を人間が行なうべきかを，電子計算機は人間に指示する。人間の判断を電子計算機に与えれば，電子計算機は処理を続行する。すなわち，このアルゴリズムによる変換辞書作成システムは，人間機械相互交信 (man-machine communication) システムである。

(2) シソーラスのみでなく，件名標目表などシソーラス類似の辞書類にも適用できる汎用アルゴリズムであ

る。ただし，語彙は英語に限られる。

(3) 原デスクリプタから同義または上位の目標デスクリプタへの対応のみが，このアルゴリズムで与えられる。したがって，変換辞書は一般に部分：部分対応，かつ多：多対応である。

このアルゴリズムの手順を原著の実例によって要約する。

原シソーラス：CAS Search Guide

目標シソーラス：Agricultual/Biological Vocabulary

目的：原デスクリプタ ARTERIES (BIOLOGICAL) に対応する目標デスクリプタの発見

手順（第1図，第2図参照）：

(1) A/B Vocabulary から ARTERIES (BIOLOGICAL) をさがす（機械）。

結果はなし。

(2) CAS Search Guide から ARTERIES (BIOLOGICAL) の上位語をさがす（機械）。

結果は BLOOD VESSELS があり。

(3) A/B Vocabulary から BLOOD VESSELS をさがす（機械）。

結果はあり。

(4) CAS Search Guide と A/B Vocabulary の BLOOD VESSELS が同形異語か否かをしらべる（人間）。

結果は否。

(5) ARTERIES (BIOLOGICAL) から BLOOD VESSELS への対応を変換辞書に記入する（機械）。

以上

第 1 図 CAS Search Guide

```
        ⋮
ARTERIES (BIOLOGICAL)
  BT BLOOD VESSELS
  NT AORTA
  RT ATERIOSCLEROSIS
  RT BLOOD STREAM
  RT CAPILLARY
  RT CARDIOVASCULAR SYSTEM
  RT CIRCULATORY SYSTEM
  RT HEAT
  RT OCCLUSIONS
  RT VEINS
ARTERIOSCLEROSIS
        ⋮
```

第 2 図 A/B Vocabulary

```
        ⋮
ARTERIES
  BT BLOOD VESSELS
  NT AORTA
  NT DUCTUS ARTERIOSUS
ARTERIOSCLEROSIS
        ⋮
BLOOD VESSELS
  NT ARTERIES
  NT CAPILLARIES
  NT VEINS
  RT ANGIOGRAPHY
BLOOD VISCOSITY
        ⋮
```

ARTERIES (BIOLOGICAL) の同義語として，A/B Vocabulary には ARTERIES が存在するにもかかわらず，Auerbach Corp. のアルゴリズムではこの対応は発見できない。

もし，両デスクリプタの下位語として AORTA が共通なことも利用できるアルゴリズムならば，ARTERIES (BIOLOGICAL) から ARTERIES への対応を発見できよう。

しかしながら，これは機械処理の過程に人間が介入する人間機械相互交信システムのアルゴリズムなので，人間の負担を増加させないために，アルゴリズムをこれ以上複雑にすることには問題があろう。

実用的な立場からは，次の理由により変換辞書を人間機械相互交信システムにより作成する効果は薄いと思われる。

(1) 人間機械相互交信システムで対応が与えられなかった原デスクリプタが残るので，あとで改めてこれらからの対応を人間が調査して追加しなければならない。Auerbach Corp. では前述のアルゴリズムにより，MeSH など11種の原語彙から A/B Vocabulary への変換辞書を作成する実験を行なった。その結果，もっとも成績のよかった MeSH でさえも，アルゴリズムが対応を与えた原デスクリプタは原語彙の 76.0% であり，残りの 24.0% のうち 13.0% は人間が対応を与えることができた。このことは，人間機械相互交信システムが変換辞書作成システムとして中途半端であることを意味している。

(2) ある原語彙からある目標語彙への変換辞書は一度作成すればよいので，この作業に特に即時性が強く要求されることはない。

2. 候補デスクリプタの自動的選定

変換辞書は次の手順で作成する。

(1) 電子計算機は，それぞれの原デスクリプタと対応する可能性のある複数の目標デスクリプタを目標シソーラスから選定し，リストとしてアウトプットする。

(2) リストで与えられた目標デスクリプタから，人間が対応するものを選定する。

(3) リストで目標デスクリプタが与えられなかった原デスクリプタ，およびリストで与えられた目標デスクリプタが不適切な原デスクリプタは，人間が調査して，対応する目標デスクリプタが存在すれば追加する。

この目的に使用するための，候補デスクリプタリストの作成法を提案する。

次の三つの仮定から出発する。

(1) 語形の一致する原デスクリプタと目標デスクリプタとは，同義語である可能性が強い。

(2) USE 参照語は，それに代るデスクリプタの同義語である可能性が強い。

(3) 上位語，下位語，関連語の一致数の多い原デスクリプタと目標デスクリプタとは，同義語である可能性が強い。

以上の仮定に基く次の二つの方法により，目標デスクリプタを機械的に選定する。

(1) 直接法

次の4種類の目標デスクリプタを選定する。

(a) 原デスクリプタと同形の目標デスクリプタ。

(b) 原デスクリプタと同形の目標 USE 参照語から，USE で参照されている目標デスクリプタ。

(c) 原デスクリプタに UF で参照されている原 USE 参照語と同形の目標デスクリプタ。

(d) 原デスクリプタに UF で参照されている原 USE 参照語と同形の目標 USE 参照語から USE で参照されている目標デスクリプタ。

(2) 間接法

直接法で対応が与えられたデスクリプタは互に同義であるとして，原デスクリプタの関係語と，ある程度以上関係語が一致する目標デスクリプタを選定する。

たとえば，前記の ARTERIES (BIOLOGICAL) に対して，A/B Vocabulary の ARTERIES は，BT BLOOD VESSELS, NT AORTA が一致する。一方，BLOOD VESSELS は VEINS が一致するが，RT と NT で関係の種類が異なる。したがって，間接法では，BLOOD VESSELS ではなく，ARTERIES が選定される。

間接法では，関係語の一致度の測定法，および目標デスクリプタの選定基準となる基準値を，あらかじめ設定する必要がある。もっとも簡単な方法は，共通する上位語，下位語，関連語の数をもって一致度とし，この数が n 個（基準値）以上の目標デスクリプタを選定する方法である。n の値を大きくするほど，選定語数は少なくなるが，語は厳選される。一方，n の値を小さくするほど，選定語数は増加するが，不適当語が混入する割合も増加する。もっと精密な方法としては，あるデスクリプタの概念を規定する能力は，下位語，上位語，関連語の順であることに着目して，この順に重みづけを与えることも考えられる。たとえば，$w_1, w_2, w_3\ (w_1>w_2>w_3)$ をそれぞれ下位語，上位語，関連語に与えた重み，n_1, n_2, n_3 をそれぞれ下位語，上位語，関連語の一致語数，w_c を基準値とし，$w_1n_1+w_2n_2+w_3n_3 \geq w_c$ の目標デスクリプタを選定する方法も考えられる。

〔実施例〕

実際に，直接法と間接法を机上実験により試みた。

原シソーラス : Thesaurus of Documentation Terms[5]

目標シソーラス: Information Science Thesaurus[6]

原シソーラスでは，キーワードの概念は平面図上の一定の領域で規定されている。そこで，非デスクリプタはその領域のデスクリプタと同義語とみなした。さらに，間接法では2個以上の関連語が一致する目標デスクリプタを選定した。

次に，索引法に関するデスクリプタを処理した実例を記す。

実例 1

原デスクリプタ: Coordinate Indexing

目標デスクリプタ

直接法: Coding systems, coordinate indexing

間接法: analyzing, authors, coordinate indexing, keyword-in-context indexing, storage and retrieval

この例では，直接法でも間接法でも coordinate indexing が選定されている。このように，直接法でも間接法でも原デスクリプタの同形語が選定されれば，同型異義語ではないとみなして，機械的に選定できると思われる。

実例 2
原デスクリプタ： Keyword Assignment
目標デスクリプタ
直接法： indexing
間接法： analyzing, authors, indexing, keyword-in-context indexing, semantics, storage and retrieval

この例では，原デスクリプタと同義の目標デスクリプタは存在しない。そこで，上位語として indexing を選定するのが適当と思われる。

実例 3
原デスクリプタ： Roles
目標デスクリプタ
直接法： role indicators
間接法： なし

この例では，直接法により同義語が選定されているが，間接法で選定されたデスクリプタはない。

実例 4
原デスクリプタ： KWIC indexes
目標デスクリプタ
直接法： なし
間接法： なし

この例では，同義語として keyword-in-context indexes が存在するにもかかわらず，語形が異なるため選定されない。

E. 問題点

デスクリプタ単位の変換法には限界がある。たとえば，目標シソーラスでは，cards, files, card files がデスクリプタ，原シソーラスでは，cards, files はデスクリプタで，card files はデスクリプタでなかったと仮定しよう。この場合，cards と files が与えられている文献は，card files に変換するのが最適であるが，デスクリプタ単位の変換では cards と files に変換される。

この問題を解決するために，次の変換方法が考えられる。

原デスクリプタ cards は，その文献に files が与えられていれば card files に変換し，files が与えられていなければ cards に変換する。原デスクリプタ files は，その文献に cards が与えられていれば card files に変換し，cards が与えられていなければ files に変換する。

この方法は，その文献に与えられている他の原デスクリプタの存否によって異なる目標デスクリプタに変換されるので，条件付変換法とでもいえよう。

条件付変換のための辞書を作成し，実際に条件付変換を行なうことも不可能ではないだろう。しかしながら，一般的にはこのように精密な辞書を人手で作成することは容易でなく，自動化にも限界がある。

そこで文献単位に複数の原デスクリプタを複数の目標デスクリプタに変換する方法が問題となる。

Ⅳ 文献単位の変換法

A. 概要

この方法の一般的な特徴を次に記す。

(1) 変換辞書として，個々の原デスクリプタから個々の目標デスクリプタの対応の度合を数量化した相関値を表示した辞書を使用する。

(2) 変換は文献ごとに次のように行なう。まず，すべての目標デスクリプタについて文献に対する適合値を計算する。ある目標デスクリプタの適合値は，その文献に与えられたすべての原デスクリプタからその目標デスクリプタへの相関値から，あらかじめ定めた計算式により算出する。

ついで，適合値があらかじめ定めた一定の基準値以上の目標デスクリプタを，その文献のデスクリプタとして選定する。

したがって，この方法では，変換辞書の作成法，および相関値から適合値を算出する計算式と基準値が，問題となる。

変換辞書の作成には，標本文献を使用する。すなわち，原ファイルから抽出した標本文献について，実際に目標シソーラスを使用して索引を行なう。次いで，それぞれの原デスクリプタとそれぞれの目標デスクリプタとの文献における出現の一致度に関する統計値を自動的に求め，この値を相関値とする。このような目的に使用する標本文献を，学習標本と呼ぶことにする。

具体的には，どのような統計モデルを採用するかによっていろいろな方法が考えられ，統計的自動分類法と統計的自動索引法が参考になる。たとえば，Maron, M. E.[7] が自動分類に応用した条件付確率を基礎にすれば，次の方法になる。

B. 条件付確率による方法

原デスクリプタ $\{x_i, x_j, \cdots, x_l\}$ が与えられている文献に対する目標デスクリプタ $y_j(1<j<n)$ の適合値

$p(x_i \cdot x_j \cdot \cdots \cdot x_l, y_j)$ を次式により計算する。

(1) 適合値の計算法

$$p(x_i \cdot x_j \cdot \cdots \cdot x_l, y_j)$$
$$= k \cdot p(y_j) \cdot p(y_j, x_i) \cdot p(y_j, x_j) \cdot \cdots \cdot p(y_j, x_l)$$

ここに

k: $\sum_{j=1}^{n} p(x_i \cdot x_j \cdot \cdots \cdot x_l, y_j) = 1$ とするための定数。

$p(y_j)$: 学習標本において，全文献に対する y_j が与えられた文献の割合で，辞書で与えられる値。

$p(y_j, x_k)$: 変換辞書で与えられる相関値。学習標本において，y_j を与えられた文献に対する y_j と x_k を与えられた文献の割合。

(2) 目標デスクリプタの選定法

適合値が，あらかじめ定めた基準値以上の目標デスクリプタを，その文献のデスクリプタとして選定する。基準値は，学習標本において，索引と変換によって得られた目標デスクリプタ総数が同数となるように定める。

C. 問題点

III. E. で述べたように，文献単位の変換法では，デスクリプタ単位の変換法では不可能な変換が可能になる。さらに，原シソーラスと目標シソーラスの言語が異なるときでも，変換辞書を自動的に作成できる利点がある。

しかしながら，次のような問題がある。

(1) 統計モデルの妥当性が問題である。

たとえば，前記の条件付確率を応用する方法では，次の二つを仮定している。

(a) 目標デスクリプタに関して，原デスクリプタの出現確率は独立である。

(b) 文献にすべての目標デスクリプタを与えることが妥当である確率は，いかなる文献でも常に 1 である。

この二つの仮定は，実際には成り立たない。そこで，実用上支障のない範囲で成り立つか，あるいはこれらの仮定が成り立たないことを考慮に入れてモデルを修正するか，という問題がある。

(2) 変換前後のデスクリプタ間の概念の対応が不明なので，検索結果に基いて変換辞書を修正するなどの人手による改良の余地がない。

(3) 実用上の立場からは，学習標本の大きさが問題となる。変換辞書を作成するために，学習標本について，実際に目標シソーラスを使用して人間が索引を行なわなければならない。このために必要な文献数は，原シソーラスの語彙量と文献当りの平均デスクリプタ数によって異なるが，かなりの量となろう。これに伴なって，標本抽出法もまた問題となる。

文献単位の変換法をデスクリプタ単位の変換法と併用することも考えられる。すなわち，両方法で変換したファイルを適当に合併して，目標ファイルを作成することもできる。

見方を換えれば，デスクリプタ単位の変換法を一般化すると文献単位の変換法になる。すなわち，文献単位の変換法において，変換辞書の相関値を対応の有無に応じて1と0とし，相関値の合計が1以上の目標デスクリプタを選定する方法が，デスクリプタ単位の変換法である。このことから，学習標本について計算により求めた相関値の一覧表をアウトプットすれば，このリストを III. D. 2. で述べたデスクリプタ単位の変換法のための候補デスクリプタリストとして利用できることがわかる。

V 総　括

国際レベル，国家レベル，機関レベルで各種の情報システムの設立が進むにつれて，レベル間およびレベル内でのシステムの両立性と変換性が重要な問題となっている。これには，システムを構成するいろいろな要素の検討が含まれるが，特にキーワードの問題が重要である。このためには，有効な変換方法の開発と並行して，シソーラスの標準化の推進が必要である。

1. シソーラスの標準化

正確な変換辞書を作成することが，変換を成功させる前提である。このためには，豊富な関係語が正確に与えられているシソーラスにより原ファイルが索引されていることが望ましいことは，Auerbach Corp.[4] の実験からも立証できる。

このため，シソーラスの標準化が必要となる。

まず第1に，シソーラス形式の標準化である。関係語の定義，キーワードの記載法などのシソーラスの形式に関する一般的な規則が作成され，どのシソーラスもこの規則に基いて作成されれば，変換は今までよりも容易になる。これと関連して，UNESCO では科学技術に関するシソーラス作成のための国際的な指針[8] を提案している。

次に，標準シソーラスの作成である。目的や利用者集団の特性に応じて，各種のシソーラスが並存することはやむを得ないとしても，ある範囲の基本語彙が共通なら

ば，少なくとも基本語彙に関しては，完全な変換が保証される。Thesaurus of Engineering and Scientific Terms はこのような意図の下に作成され，科学技術庁では科学技術に関する日本語の標準シソーラス[10]を作成している。

しかしながら，システムの両立性の観点からは，変換辞書による自動変換には一つの限界がある。

2. 自動変換の限界

Hammond, W.[3] は DDC のファイルと NASA のファイルとに共通の 996 件の文献のデスクリプタを調査した。その結果，この標本文献において，NASAファイルと DDC ファイルの両方に出現したデスクリプタは 740 種，NASA の語彙には存在するが DDC ファイルにのみ出現したデスクリプタは 187 種，DDC の語彙には存在するが NASA ファイルにのみ出現したデスクリプタが 323 種となっている。この結果は，両機関の索引基準が異なるので，変換辞書で対応が与えられているデスクリプタといえども，希望通りの変換が行なわれないことを示している。

すなわち，利用者集団が異なればキーワードの選定基準が異なるのは当然であるが，原ファイルで選定されなかったキーワードを自動変換で追加することはできない。

しかしながら，この事実は自動変換の価値を否定するものではない。なぜならば，これは利用者にとって最適とはいえない選定基準によるファイルを検索に利用することの是非の問題であり，たとえ原シソーラスを検索に使用しても解決できない問題だからである。

3. デスクリプタの修正

自動変換で希望通りのファイルが得られないときには，自動変換後に人間がデスクリプタの追加，削除を行なうことが考えられる。修正量が比較的少なく，かつ抄録が磁気テープファイルとして利用できるならば，ディスプレー装置を使用する人間機械相互交信システムで，抄録により人間が修正を行なうことも可能であろう。

この考え方をさらに発展させれば，抄録からの自動索引とデスクリプタの自動変換を並用する自動化法も考えられる。

4. 検索デスクリプタの自動変換

文献ファイルのデスクリプタを変換する代りに，検索時に質問式のデスクリプタを文献ファイルのデスクリプタに自動変換することも可能である。

この方法は，III. で述べたデスクリプタ単位の変換と同じ原理である。すなわち，変換辞書により検索デスクリプタを文献デスクリプタに変換すれば，その変換辞書を逆に適用してファイル変換を行なったときと同一の検索結果が得られる。ただし，質問式でデスクリプタ間の AND と OR を使用する論理検索では，これらの追加，変更，削除による質問式の修正が必要となる。

最後に，I. で述べた 3 例について考察する。

(1) シソーラスの改訂

キーワード選定基準，シソーラスの形式は改訂により変らない。また，改訂の対象となったデスクリプタのみを変換の対象とすればよい。したがって，自動変換は比較的容易であると思われる。

ただし，再改定に伴ない再変換が必要となる。この場合，変換で作成されたファイルから逆変換により復元された原ファイルから再変換を行なう方が正確である。したがって，この変換には復元性のあることが望ましい。

(2) シソーラスの翻訳

ある国際システムにおいて，日本の文献は日本語のシソーラスで索引し英語のデスクリプタに変換して提供し，提供された外国の文献の英語デスクリプタを日本語デスクリプタに変換して検索することも考えられる。

このための変換辞書は人間が作成しなければならないが，キーワードの選定基準および専門分野が一致しているので，原理的にはそれほど困難ではない。

ただし，この変換には可逆性が要求されるにもかかわらず，言語が異なるので 1：1 対応の変換辞書は不可能である。したがって，自動変換後の人間による修正が，ある程度必要となる。

(3) キーワードファイルの統合

次のファイルからの変換は，一般に容易である。

(a) 語彙がシソーラスにより統制されている。

(b) いろいろな観点から索引されている。

(c) 個々の概念は詳しい（specific）デスクリプタで記述されている。

しかしながら，この 3 条件を満足するファイルの作成は困難である。実際には，キーワードの選定基準，シソーラスの語彙などの異なるファイルの変換となり，変換の成否はこれらの要素がどの程度近似しているかによる。

NAL[2] では，MEDLARS, CAS などのファイルから NAL にとって必要な文献を自動的に抽出・変換して，NAL で作成したファイルに追加するシステムを検討し

ている。このように，ある機関にとって補助的なファイルは，検索の正確さを犠牲にしても，変換して主ファイルに吸収することが便利な場合もある。

本論文では，いろいろな変換法の一般的な性質を考察した。これらの具体的な適応条件の解析は，今後の課題として残されている。

1) Holst, W. "Mechanical Translation by Coordinate Indexing," *Amer. Doc.*, vol. 17, 1966, no. 3. p. 140–141.
2) Landau, H. B. "Design Criteria for a Multi-Input Data Base for the National Agricultural Library," *Proceedings of the ASIS*, vol. 6, 1969, p. 101–104.
3) Hammond, W. Dimensions in Compatibility. <Newman, S. M., ed. *Information System Compatibility*. Spartan, 1965> p. 7–17.
4) Auerbach Corp. "*Intersystem Compatibility and Convertibility of Subject Vocabularies*," no. 1582–100–TR–5 (PB184144).
5) Rolling, L. *The Role of Graphic Display of Concept Relationships in Indexing and Retrieval Vocabularies, Including a Thesaurus of Documentation Terms*. Euratom, 1965. No. EUR–2291e.
6) Schultz, C. K. "*Information Science Thesaurus*," Drexel Institute of Technology, Drexel Library Series, No. 17, 1967.
7) Maron, M. E. "Automatic Indexing; An Experimental Inquiry," *J. Assoc. Comp. Mach.*, vol. 8, 1961, no. 3, p. 404–417.
8) UNESO. *Guidelines for the Establishment and Development of Monolingual Scientific and Technical Thesauri for Information Retrieval*. 1970. SC/MD/20.
9) Engineers Joint Council. *Thesaurus of Engineering and Scientific Terms*. New York, 1967.
10) 渡辺　茂：〝科学技術標準シソーラスの作成〟**情報管理**, vol. 12, 1969, p. 487–492.