人文科学とコンピュータ　18－6
（１９９３．５．２１）

テキストデータベースを利用した研究情報検索

春山　暁美

（前）愛知淑徳大学

文字データを検索する上でのテキストデータベースの柔軟性・簡便性に着目して、化学物質情報及び研究文献情報のデータベースとして利用した二つの事例を報告する。環境中の化学物質に関する種々な情報の検索、及び研究文献データベース改良のためのキーワード付与作業用ファイル作成に利用した結果、処理が簡単なこと、記述文や入力形式が一定でない文字データなどの検索が容易に行えることなどの利点が認められた。実用のためには、さらに新しい機能の追加が望まれる。

Application of a Text Database System to Research Information Retrieval

Akemi HARUYAMA

(formerly) Aichi Shukutoku University

Rm. 301, 50-1 Choda-cho, Meito-ku, Nagoya 465 Japan

Two cases of utilizing a text database system for retrieval of various information on chemical substances and bibliographic data are reported. Easiness in handling data files and efficiency in retrieving character data was recognized. For future utilization new functions should be added to present text database systems.

1．はじめに

テキストデータベースシステムは，著作物などのテキストの内容解析のために開発されたが，これを文献情報のデータベースとしても利用することができる。とくに，一部に英文の文章を含むデータベースや，和文と英文の文字データが混在し，さらにそれらの入力形式（大文字，小文字など）が一定していないデータを検索しようとするときには，従来の汎用大型コンピュータを利用したデータベースなどの場合にくらべて検索が簡単で容易である。また，データの追加や訂正も容易である。これらの特性をいかし，環境中の化学物質に関する種々な情報の検索，及び研究文献データベースの改良に利用した例について報告する。

2．環境中の化学物質に関する情報検索への利用

2.1　利用の目的と背景

環境中の化学物質が人間の健康に及ぼす影響を評価するためには，多くの化学物質に関する情報が必要であるばかりでなく，個々の化学物質について，多種多様な情報が必要となる。例えば，ある特定の化学物質について，物理化学的特性・毒性・代謝・分析法・環境中の測定データ・法規制など，種々の観点からの情報が必要とされる。

化学物質に関するデータベースは多数存在するが，各物質について上に述べた多様な観点からの情報を網羅したデータベースは，実際には存在せず，検索に当たっては，多くのファイルを探索しなくてはならない。また，あまりにも大量の情報が収録されているために，実際に必要とするデータに到達するまでに，多大の労力と時間を要することが多い。

一方で，研究者は，それぞれの研究分野において，独自の方法により化学物質に関する種々な情報を収集し，蓄積している。また，これらの情報の集積結果として，化学物質のリスクに関する優れたリファレンス・ブックもいくつか刊行されている。しかし，これらは，データベース化されていないために，その利用の範囲が，限られたものとなっている面がある。

そこで，これらの従来データベース化されていなかった情報を主な対象として，テキストデータベースシステムの活用により，既存の商用大型データベースを補完するものとしての，研究者のための実験的な環境化学物質データベースを試作した。

2.2　対象としたデータ

国立環境研究所（1974-1990年6月まで国立公害研究所，1990年7月より現名称）が収集した化学物質に関する情報ファイルの中から，163物質の物質名・物理化学的特性などとともに，これらの物質の毒性・中毒症状・救急処理などに関する記述文を対象として，テキストデータベースを作成した。

2.3　機器構成及び使用ソフトウェア

NEC PC9801シリーズ本体及び周辺機器を使用した。ソフトウェアとしては，テキストデータベース研究会（事務局：千葉大学人文学部哲学研究室）から提供を受けたＴＥＸＡＳ（Text Analysing System，原作者　弘前大学人文学部　清水　明氏）を使用して，テキストデータベースシステムの構築・運用を行った。

2.4 方法

1) テキストの作成

国立環境研究所が収集した化学物質情報ファイルの内，163物質に関する英文の情報をキーボードから入力してテキストファイルを作成した。テキストの内容は，大まかに次の二つの部分に分けることができる。

① 化学物質の物質名，別名，商品名，CAS Registry Numberなどの登録番号，分子式，分子量，沸点，融点，比重，屈折率，その他の物理化学的特性，参照文献など
② 急性毒性，慢性毒性，代謝，中毒症状，救急処理，分解生成物，など（一部に記述文を含む）

2) テキストデータベースの作成

TEXASでは，関係型データベースのように「データ項目」の集まりである「レコード」を単位としたデータベース構築は行わない。しかし，このデータベースでは，一つの物質に関する種々な情報が，全体として一つにまとまっている必要がある。このために，各物質毎の情報を1ページのテキストとみなし，本来のテキストである記述文のほかに，「データ項目」に当たる各項目をそれぞれ1行のテキストとして，データベースを作成した。また，検索結果を見やすくするために，見出しとして，各物質名及び物質を特定するためのＩＤ番号に当たるCAS Registry Number を割り当てた。

2.5 利用の結果と考察

試作データベース「CHEMRISK」は，パーソナルコンピュータとフロッピーディスクによるシステムであるため，可搬性があること，また，研究者がデータベースの配布を受け，さらに独自のデータを追加したり，修正を行ったりが容易にできることが，大きな特色である。

本報告の場合は，大量の化学物質データを収集することよりも，ある程度の数の物質について，より詳しい広範囲のデータを一覧できることが重要と考えられる。この点で，テキストデータベースシステムの利用は，目的にかなうものと考えられる。

テキストデータベースシステムでは，研究者がワープロと同様の感覚で自由にデータを追加したり，修正したりすることができる。そのとき，大文字，小文字などの入力形式が不統一でもかまわない。和文・英文の混在も許される。これらは，従来の汎用大型コンピュータを利用したデータベースなどの場合に比べてはるかに柔軟であり，有利な点であるといえる。

しかし，その反面，数値演算ができないこと，テキストデータを行単位で管理するため，現在の TEXAS では，ページ番号（この場合は各物質ごとの情報の集まり）を要素とする検索集合の和，差，積を作成できないことなどの不利な面もある。実用のためには，これらの機能を追加する必要があると考えられる。

なお，この研究は，著者が1989年3月まで在職し，その後1992年3月まで客員研究員としてかかわった国立環境研究所の特別研究「先端技術における化学環境の解明に関する研究」（昭和62年度～平成3年度）の一環として行われたものである。

3．研究文献データベースの改良への利用

3.1 利用の目的と背景

労働衛生分野は，労働の場における人間・環境系の中で，健康を阻害する種々の要因に

ついて，調査・研究を行い，有害因子の除去，作業環境の改善など，健康をまもるための対策を講じる分野である。わが国では，労働省の付属研究機関である産業医学総合研究所（1956年～1976年6月までは労働衛生研究所，1976年7月より現名称）が，この分野での唯一の国立専門研究機関として，研究を行ってきた。

1990年に，産業医学総合研究所の研究成果のデータベース化が行われたが，現段階では抄録を含まず，また，キーワードや分類を付与していないことなどから，検索効率の面では不十分な点が多い。

労働衛生分野が学際的な領域であること，及び研究成果が多方面から利用されることを考慮すると，できるだけ多くの面からの検索が行えることが望ましい。

そこで，著者らは，検索効率の向上を図るために，データベースの収録文献に対して，遡ってインデクシング（ここでは検索のためのキーワードなどの付与をいう）を行い[1]，検索実験によりその効果を検証する試みを行った。[2]

このとき，作業をしやすくするために，原データベースからオフラインデータベースをテキストデータベースシステムにより作成し，これを検索してインデクシング作業用のファイル作成に利用した。

3.2 対象データベース

汎用大型コンピュータ FACOM M-730/6 を使用して，対話型情報検索システム FAIRS-I により作成された書誌データベースである。表1に収録文献の種類と蓄積件数，図1に出力例を示した。

表1 データベース収録文献数

種類		蓄積件数
原著	（和文）	258
〃	（英文）	558
総説	（和文）	365
〃	（英文）	3
著書	（和文）	179
〃	（英文）	55
発表講演	（和文）	1743
〃	（英文）	153
報告書	（和文）	448
〃	（英文）	6
合計		3768件

3.3 テキストデータベースの機器構成及び使用ソフトウェア

2．と同様に，NEC PC9801シリーズ本体及び周辺機器，またソフトウェアとして TEXAS を使用した。

```
==発表講演（和文）==

#1      DOCUMENT NUMBER         76GA001
        TITLE                   低密度精神作業のストレスと生理的変動について
        AUTHOR                  柿崎　敏雄
        PROCEEDINGS             第４９回日本産業衛生学会
        JURNAL                  第４９回日本産業衛生学会講演集
        PAGE                    306-307
        YEAR                    1976
        NUMBER                   34

==原著（英文）==

#1      DOCUMENT NUMBER         83GE003
        TITLE                   TIME COURSE OF THE CHANGES OF CATECHOLAMINE LEVELS IN
        AUTHOR                  AYAKO SUDO
        JURNAL                  BRAIN RES.
        VOLUME                  276
        PAGE                    372-374
        YEAR                    1982
        NUMBER                   12
```

図1 データベース出力例

3.4 方法

1) オフラインデータベースの作成

原データベースの主ファイル内容（マスタファイル作成時の入力データファイル）を，フロッピーディスクに出力して，MS-DOS 標準テキストファイルに変換し，このファイルに基づく可搬性のオフラインデータベースを，TEXAS により作成した。

オフラインデータベースをテキストデータベースとした主な理由は，大文字，小文字などの入力形式が統一されていないデータをそのまま検索できること，及び処理が簡単なことである。

FAIRS-I によるデータベースのマスタファイル作成用入力データファイルは，本来の意味での「テキスト」ではないが，テキストとみなして処理しやすい形式である。このファイル全体を一つのテキストとみなしてテキストデータベースを作成した。

2) インデクシング作業用ファイルの作成

オフラインデータベースを TEXAS により検索してできるだけ同一の主題に関する文献レコードを抽出し，フロッピーディスクに出力してインデクシング作業用のファイルを作成した。これは，キーワード等の付与及び入力時に，同一主題に関する文献をまとめて処理することにより作業の効率を上げるためである。（図２）

```
#  1:S:     8 items          *ＶＤＴ *
ITEM    1 :  +-6     7 items left
                  ⋮
-TI ＶＤＴ作業時の頭部運動と視線移動
-AU 斉藤　進*斎藤　真*大久保　堯夫
-PR 第６回姿勢シンポジウム
-JN 第６回姿勢シンポジウム講演集
-PA 19-20
-YR 1985
                  ⋮
                  ⋮
ITEM    8 :  +-6     0 items left
                  ⋮
-TI 眼球運動の疲労特性とＶＤＴ作業
-AU 斉藤　進
-PR 第３回日本眼科医会ＶＤＴ研究会
-JN 日本の眼科
-VL 58
-PA 1203-1204
-YR 1987
```

図2　TEXASによる検索例

この作業用ファイルを用いて，ファセット分析に基づくシソーラス用語及び分類コードを各収録文献に追加後（図３），著者校正を経て，このファイルに基づくデータベース更新を実験的に行った（図４）。さらに，付与されたキーワードなどが，検索結果に与える影響を検証するための検索実験を行った。それらの詳細は，別に報告した[1][2]ので，ここでは省略する。

```
-TI 眼球運動解析におけるＶＤＴ作業の定量的評価
-AU 斉藤　進
-PR 第２回日本眼科医会ＶＤＴ研究会
-JN 日本の眼科
-VL 58
-PA 733
-YR 1987
-KW CRT DISPLAY TERMINALS(Jaxuc)*COMPUTER TERMINALS(Jaxu)*OFFICE
    EQUIPMENT(Jax)*VISUAL DISPLAYS(Jzed)*ELECTROMAGNETIC RADIATION(Bo)*
    VISUAL TASKS(Kika)*EYES(Pansa)*OPHTHALMOLOGY(Qoxg)*VISUAL FATIGUE
    (Phunv)*ERGONOMIC EVALUATION(Qwe)
```

図3　遡及インデクシングの例

#63	DOCUMENT NUMBER	73GA003
	TITLE	自由作業と規制作業について
	AUTHOR	須藤　綾子
	PROCEEDINGS	第４６回日本産業衛生学会
	JURNAL	第４６回日本産業衛生学会講演集
	PAGE	198-199
	YEAR	1973
	NUMBER	12
	KEYWORD	OCCUPATIONAL PHYSIOLOGY(PHO)
		PHYSIOLOGY OF ENDOCRINE SYSTEM(PHE)
		AMINES(DO)
		EPINEPHRINE EXCRETION(PHEACE)
		NOREPINEPHRINE EXCRETION(PHEACO)
		URINARY EXCRETION(MGUR)
		DETERMINATION IN URINE(QIBN)
		DETERMINATION IN EXCREMENTS(QIBX)
		DETERMINATION OF CONCENTRATION(QI)
		HORMONE SECRETION(PHEH)
		REPETITIVE WORK(KID)
		MENTAL WORK(KIM)
		CONTINUOUS WORK(KOF)
		SPEED OF WORK(KOH)
		INTERMITTENT WORK(KOG)
	FREE TERM	PACED TASK*UNPACED TASK*HUMAN SUBJECTS

図４　更新後のデータベース出力例

3.5　利用の結果と問題点

既存のデータベースの収録内容を改変するに当たり，データベースの収録レコードを直接修正することも可能であるが，大量に行う場合は別のファイルを作った方が処理しやすい。また，修正する内容が最終的に決まるまで，数回の再修正作業が必要となる今回のような場合は，この方法が適していると思われる。

作業用のオフラインデータベースをテキストデータベースとしたことにより，文献単位の検索が困難な反面，入力形式が統一されていない文献データを，原データベースの場合よりも容易に検索することができ，作業用ファイルの作成に役立った。

また，データをテキストとして扱うことにより，インデクシング担当者も，校正を依頼した研究者も，ともにデータベースの構造や入力の規則などを意識することなく，ワープロと同様の感覚で，キーワード付与及びその校正作業を行うことができた。

問題点として，テキストデータベースの検索結果をインデクシング作業用ファイルに加工する際に，多少手間がかかることがあげられる。また，２．の場合と同様に，検索の際，文献単位で（検索集合の要素を行番号だけでなくページ番号で）検索集合の和，差，積を作成できる機能があることが必要であろう。

４．参考文献

[1] 春山暁美，久保田均：労働衛生分野データベース改良のための遡及インデクシングの試み（１）．第28回情報科学技術研究集会発表論文集．日本科学技術情報センター，東京，PP.95-104（1992）

[2] 春山暁美，久保田均.労働衛生分野データベース改良のための遡及インデクシングの試み（２）．キーワード付与が検索結果に与える影響．第29回情報科学技術研究集会発表論文集．日本科学技術情報センター，東京，PP.361-368（1993）