HOKKAIDO UNIVERSITY

Title	オンライン・データベース検索における適合率向上の考察 ： キーワードの差別化と関係子を付与したキーワードによるデータの選別
Author(s)	岩垂, 司; 三国, 景史; 小銭, 正尚 他
Citation	北海道大學工學部研究報告, 171, 19-26
Issue Date	1994-10-28
Doc URL	https://hdl.handle.net/2115/42434
Type	departmental bulletin paper
File Information	171_19-26.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北海道大学工学部研究報告
第171号（平成6年）

Bulletin of the Faculty of Engineering
Hokkaido University, No. 171 (1994)

オンライン・データベース検索における適合率向上の考案
－キーワードの差別化と関係子を付与した キーワードによるデータの選別－

岩垂　司　三国　景史　小銭　正尚
三枝　武男*　折登　一彦**

（平成6年6月17日受理）

Improvement of relevance in on-line database retrieval by discrimination of keywords and selection of relevant data using keywords accompanied by role indicators

Tsukasa IWADARE, Keishi MIKUNI, Masahisa KOZENI, Takeo SAEGUSA and Kazuhiko ORITO

(Received June 17, 1994)

Abstract

A procedure to descriminate keywords in on-line database retrieval was proposed. The discrimination was carried out by seeking major and minor (less essential) keywords in the article title and Basic Index respectively. This procedure afforded significant improvement in the relevance of output data. Roles of keywords appearing in article titles were analysed. There the keywords were classified into five categories, and the relationship between the roles of keywords and accompanying role indicators were investigated. Use of the keywords accompanied by proper role indicators gave better results in selection of relevant data than those by the conventional procedure.

1．はじめに

現在オンライン・データベース検索において情報の選別は，キーワードの論理積，和，差の組み合わせにより行われている。補助的な手段として近接演算，語幹一致，範囲指定などがあるにせよ，主たる手段がキーワードであることには変わりはない。キーワードによる情報選別に関しては従来から，(1)キーワード間の重み付けが出来ない，(2)キーワード間の関係付けが出来ない，(3)キーワードが名詞中心で形容詞，副詞的表現の抽出が出来にくい，などの難点が指摘されてきた。これに対し従来から入力されたキーワードと出力データ中のキーワードのあり方とを関係付け，適合率を向上させるための研究が行われてきた[1,2]。これらの中には妥当な結果を得ているものもあるが，手続きが煩雑で作業に長時間を要し実際的でなく，検索現場の用に供するにはより

学習情報通信システム研究所
＊北海道情報大学
＊＊分子化学専攻　精密合成化学講座

実際的な方法が望まれる。今回我々は，キーワード間の重み付け及びキーワード間の関係付けに関し，簡単で効果的な方法を考案実験したのでその結果を報告する。なお本研究は，現用オンライン・データベースを用いて行う実務の現場での検索技術に関するもので，システム／ファイルの機能の変更を意図するものではない。従って実験はすべてシステム／ファイルに付与されている機能を用いて行われた。

2．キーワードの差別

2.1 目的

検索に複数のキーワードを用いる場合，総てのキーワードの重要度（重み）は同一ではなく，そこには中心となる重要度の高いものと，補助的なものとの区別が存在する。現行のシステムでは総てのキーワードに同じ重みが与えられ，重要度に応じた差別が出来ない。重要度に応じたキーワードの差別は出力データの選択（絞り込み）上有効であろうと考えられる。我々はキーワード差別のため以下に述べる方法を提案する。

2.2 方法

ここで提案する方法の骨子は，検索主題の中心となる重要度の高いキーワードは論文表題(TI)を，副次的キーワードはベーシック・インデックス(BI)をフィールドとして検索することによりキーワードを差別することである。文献検索における適否判定は(1)論文表題，(2)抄録，(3)本文チェックの順序で行われる。このプロセスから分かるように，基本的選別は表題をチェックすることにより行われるのである。現代においては論文の速報性が益々重視される傾向にあり，このことはCurrent Contentsのような学術雑誌の目次を集めた情報誌の発行部数が大きく伸びていることによっても示されている。一方学術雑誌における論文の表題も出来るだけ内容を具体的に表現することが要求されるようになってきていて，論文表題中に研究内容に関する重要な言葉（語）が現れる事が多い。かかる現況から重要度の高いキーワードの論文表題中の存在はデータ選別の有力な手がかりになるものと考えられる。

この考えに基づいて，既に収集した検索事例[3]で用いられた検索主題の中からテーマを選びテストを実施した。テストにはJOIS/JICST(010)及びSTN/CAのシステム／ファイルを用いた。先ずBI(Basic Index)を検索しポスティング数が100以上の主題について更に表題をフィールドとして検索し，BIについては100件，表題についてはポスティング数が100以上の場合は100件，100未満の場合は全数のタイトルを出力した。出力された表題を調べ，適合，不明（更に調べなければ分からない），不適合に分類した。不明のものは抄録，キーワードを含む全項目を出力して適否を調べた。適合性の判定基準は「当該キーワードに関する事柄を主たる内容とするもの」とした。テストはキーワード数が，1，2，3の場合につき，キーワードがBI中にある場合と表題中にある場合とで行い適合率を比較した。またを物質と事象に分類し，各々の場合の適合率を調べた。更に2キーワード以上の場合には1（又は2）キーワード表題中，1（又は2）キーワードBI中の場合の適合率を比較し，有効性を考察した。

2.3 結果と考察

それぞれの場合につき20例の検索を行い，得られた結果を表1に示す。以下の結果から分かるように，何れの場合もキーワードをBI中に求める場合と比較してキーワードが表題中に存在する場合の適合性の向上が認められる。

本方法による適合性向上の理由としては上述の論文表題の付け方の他に，現用オンライン検索システムにおける検索語付与のあり方も考えられる。現用オンライン検索システムのBIは，ファイルにより多少の差異はあるが一般的には以下のようなフィールドから構成され，(1)−(5)の各フィールドはそれぞれ独立に検索できる。

表1　キーワードの所在と適合率

キーワード数	キーワード所在	表題中キーワード数	適合率
1	BI		0.31
	論文表題	1	0.66
2	BI		0.12
	論文表題，BI	1	0.36
	論文表題	2	0.70
3	BI		0.15
	論文表題，BI	1	0.55
	論文表題，BI	2	0.71
	論文表題	3	0.78

BI : Basic Index

(1)キーワード：著者が付与，自然語，6-12語
(2) INDEX TERM(IT)：データベース作成者が表題及び抄録中から切り出して付与，統制語，10-20語
(3)表題：著者が付与，自然語
(4)抄録：著者が付与，自然語
(5)化合物登録番号(RN)：CAが付与，科学技術システム全般に適用，論文により異なるが10-20程度

論文著者は情報に関し，データベース作成者は論文内容に関し必ずしも専門家ではないので，キーワードとITが常に適切であるとは限らない。それに対し著者は論文内容に関し専門家なので，表題は適切な専門語を用いて付けられている。故に検索者が用いているキーワードと一致する割合が高くなる。これはキーワード，ITをフィールドとした場合に比べてノイズやサイレンスの減少を意味する。表題は自然語なので表題中では単数，複数は区別される。表題が英文の場合はこの点に留意しなければならない。

本方法は文献検索だけではなく，広く一般的なデータについても適用可能である。多くの学術雑誌では論文中の図，表，写真などには，本文を参照する事なくそれだけで内容を表わすキャプションを付することが要求されている。従ってキャプション中のキーワードの存在を手がかりに図，表，写真などを検索することも原理的には可能である。これを敷衍すれば画像，表，図形の検索も可能となる。更に本方法はポスティング数が少なすぎる（絞りすぎ）場合のポスティング数の回復にも有効である。すなわちそのような場合には，より重用度の高いキーワードは表題を，

重要度の低いキーワードはBIをフィールドとして検索を行なえば適当なポスティング数を得ることが出来る。

本方法の問題点として，論文の主たる内容ではないが副次的に論文中に記されている興味ある事柄を拾い出せないことがある。この点については更に研究の要があるが，重要度の高いキーワードは表題を，副次的なキーワードはBIをフィールドとするなどの工夫により解決法を見いだせると考える。

3．キーワードの役割と関係子

3.1 目的

前述のように現在のデータベース検索手法の主流となっているキーワードを用いる検索法には，(1)名詞中心，(2)キーワードの重み付けが出来ない，(3)キーワード間の関係付けが出来ない，と言う難点が指摘されている。この内の(2)については，我々はキーワードが表題中に存在するか否かである程度重み付けが可能であることを前節で述べた。現在の検索システムで用いられているキーワードの和，差，積の論理演算では，キーワード間の関係付けは出来ない。ここでキーワードの役割を明らかにすることによりキーワード間の関係付けを行えば，より有効な検索をなし得ると考えられる。この目的に沿って論文表題中のキーワードの役割を明らかにし，その結果に基づいてキーワード間の関係付けの手法を考案し，実験により実用性を検証した。

3.2 方法

日本文においては，キーワードの役割は助詞或いは的，中等の語（以下関係子と称す）によって表わされるので，我々は関係子を利用してキーワード間を関係付けることを考えた。キーワードと関係子の組み合わせによりキーワード間の関係付けを行えば，より有効な検索を可能にし得ると考えられる。

キーワードの役割と関係子の使われ方を把握するため，論文表題中で用いられているキーワードと関係子を調べた。表2に示す物理，化学，生物，情報分野の専門誌と総合科学雑誌1993年掲載のそれぞれ100編(計500論文)の論文の和文表題からキーワードと関係子を抽出した。抽出はJICSTのキーワード切り出し規則[4,5]に従って行われ，更に各専門分野の研究者の校閲を受けた。論文表題中のキーワードの役割を主語（主体），状態（性質，限定），目的（対象，方向），述語（動詞，動作，行為），手段（理由，原因），並列に分類した。ここで並列とは，と，および，並びにのような関係子をもつもので，それだけでは役割を限定できず後続キーワードによって役割の決まるものを指す。またここでの分類は言語学的な自然言語解析ではなく，検索要求に応ずる為のものである。関係子もキーワードの分類に従って整理し，キーワードの役割に応じて用いられる関係子を纏めた。日本語では漢字の連結による様々な表現が可能なため，関係子をもつキーワードともたないキーワードが生ずる（表2)。

作業は以下の手順で行った：(1)表題出力，(2)キーワード抽出，(3)役割分類，(4)関係子整理，(5)実例検索。各論文につき，和文表題，英文表題，キーワード，フリータームを出力し，キーワードと関係子の抽出には和文表題を用いた。英文表題，キーワード，フリータームは抽出の際の参照に供した。作業にはJOIS-IIIシステム中のJICST(010)ファイルを使用した。本ファイルは保有データ数約680万件と情報量が多く場合の数が多い事と，予定している実験が出来る機能を備えていることが理由である。関係子をキーワードと組み合わせて使うために，JICSTファイルの2次検索に用いる表題の文字列検索機能（ストリングサーチ型検索）を利用した。

検索実験は，(1) Basic Index でキーワードを入力してポスティング数を掴み，(2)表題中にキーワードを持つデータに絞り込み，(3)キーワードに要求に応ずる格関係子を付してキーワード＋関係子の文字列を表題中に持つデータを抽出する，と言う手順で実施した。精査を要するものについては，抄録を出力し判定した。

3.3 結果

分析結果を表2，3に示す。和文表題においては漢字の連結により状態を表わすことがあるので切り出しの際関係子を持たないキーワードが生ずる。キーワードに関係子の付く場合と付かない場合との割合は全分野を通じて，関係子付き66%，関係子無し34%であった。1割程度存在する並

表2 調査対象

分野	雑誌名	表題数	キーワード	関係子有無
生物	Nature, Sci.	100	394	272/122
情報	情報処理学会論文誌	100	358	193/165
物理	Phy. Rev.	100	392	232/148
化学	Chem. Lett. Chem. Comm.	100	454	313/141
総合	Nature Science	100	420	280/140
合計	6学術雑誌	500	2006	1290/716

表3 キーワードの役割

関係子	キーワード数	役割					
		主語	状態	目的	手段	述語	並列
あり	1290	20	551	377	182	53	107
なし	716	228	271	32	26	159	

列を示す関係子「と」，「および」，「ならびに」をもつキーワードについても同様である。関係子付きのキーワードの役割毎の出現率(%)は全分野をまとめると，主語1.5，状態35.5，目的36.1，手段13.1，述語3.2，並列8.3で，状態，目的，手段のキーワードに関係子の付くことが多く，合計で出現する関係子の8割を越える。また関係子を持たないキーワードの割合(%)は主語31.8，状態37.7，目的4.5，手段3.6，述語22.2で，主語，述語に後続関係子を持たないものが多い。これは主語，述語（合成，分析のように名詞化された動詞）は表題文末に位置する場合が多いためである。これに対し目的，手段は文末に位置することは少なく，関係子を伴って文中にある場合の多いことを示している。状態を表すキーワードは関係子が付随する場合，しない場合ともほぼ同

じ割合で出現し，何れの場合も文中にあることが多い。状態を表すキーワードで関係子の付随しないものは，後続する語に直接接続して後続語の状態を表現（修飾）している場合である。以上から関係子を付けない主体，述語キーワードと，関係子を付した状態，目的，手段のキーワードの組み合わせで有効な検索が可能と考えられる。

以下に「関係子＋キーワード」を用いて試みた検索の例を示す。なお以下の検索例において，＊は論理積(AND)，＋は論理和(OR)，＆は前方一致型マスク文字検索（トランケーション）を意味する。また関係子を付する検索は1次検索によって作られた集合に対する2次検索なので第2のキーワード（(1)の場合「毒性＆」）は入力しなくても良い。

(1)海洋生物の生産する毒性物質：キーワードとして海洋生物と毒性物質を入力し，論理積をとると，目的文献の他に海洋生物に対する毒性物質，例えばテトラブチル錫を取り扱った論文がノイズとして混入してくる。関係子を付することにより海洋生物の生産する毒に関するデータと，海洋生物に対する毒に関するデータとを区別できる。

ポスティングNo.	キーワード	キーワード位置	関係子有無	適合数
605	海洋生物＊毒性＆	BI	なし	
49	海洋生物＊毒性＆	表題	なし	
19	海洋生物の＋海洋生物から	表題	あり	11
37	海洋生物に	表題	あり	17

(2) EDTA の分析：EDTA は各種金属イオンの分析に用いられるキレート剤であるが，食品などに安定剤として添加されることがあるので，製品中の EDTA を分析する必要のあることがある。キーワードとして EDTA ＊分析を入力すると，EDTA を用いる分析に関するデータが多数混入しノイズとなるので，関係子を付与して EDTA を用いる分析と EDTA の分析とを区別する。

ポスティングNo.	キーワード	キーワード位置	関係子有無	適合数
2994	EDTA ＊分析	BI	なし	
412	EDTA ＊分析	表題	なし	
60	EDTA の	表題	あり	35

(3)ホルムアルデヒドに関する還元：ホルムアルデヒドの還元に関しては，ホルムアルデヒド（手段）による還元，ホルムアルデヒド（対象）を還元する，他物質のホルムアルデヒド（方向）への還元，の3つの場合があるのでこれを区別する。

ポスティングNo.	キーワード	キーワード位置	関係子有無	適合数
881	ホルムアルデヒド＊還元	BI		
33	ホルムアルデヒド＊還元	表題	なし	
9	ホルムアルデヒドによる＋ホルムアルデヒドを用いる	表題	あり	7
9	ホルムアルデヒドの	表題	あり	8
6	ホルムアルデヒドへの	表題	あり	6

(4)有機金属錯体の合成と有機金属錯体を用いる合成：有機金属錯体は有機合成反応において触媒としての利用価値が高いので，有機金属錯体を用いる合成反応の研究と共に有機金属錯体の合成に関する研究も行われている。キーワードの論理積だけではこれらを区別できないが，関係子を付することによりそれぞれの関係論文を選択的に抽出できる。

ポスティングNo.	キーワード	キーワード位置	関係子有無	適合数
1087	有機金属錯体＊合成	BI		
70	有機金属錯体＊合成	表題	なし	
33	有機金属錯体の	表題	あり	23
15	有機金属錯体を＋有機金属錯体による	表題	あり	15

3.4 考察

(1)関係子的役割の語については，従来より言語学的見地から意味マーカー或いはロール・インジケーターの名称の下に研究が行われて来ているが，これらをデータベース検索手法として用いた報告はない。

(2)本研究で用いた文字列検索機能は，JICST ファイルの2次検索のために付与されているもので，BI から作り出された1次集合の原文表題(OT ; Original Title)，和文表題(TI)他2項目中の文字列を検索する。この機能は(1)切り出し位置が不明の言葉または切り出されていない言葉，(2)切り出しの際削除される数字，スペース，特殊記号などの不用語（ストップワード）を含む文字列などの検索のために付与されているもので，この機能を本研究のようにキーワードの役割付与に用いた研究例は報告されていない。

(3)関係子を持たないキーワードは関係子を用いた検索ではサイレンスになるので，これの取扱いは別に工夫を要する。関係子を持たないキーワードは主語，状態及び述語に多く，このうち主語および述語は文の最後尾に位置する事が多い。この事から検索においては主語および述語は関係子を付さず，状態，手段，対象のキーワードに関係子を付してキーワード間の関係づけを行うと良いと言える。主語，述語を除いた関係子を持たないキーワードの大部分を占める状態を表すキーワードは後続語に付着して後続語を修飾している。この事から「関係子を持たないキーワード」「後続語（任意とする)」「関係子」の語順で文字列検索すればサイレンスを解消できると考えら

れ，手法について検討中である。なお表題に用いられている関係子の種類は少なく，役割毎の出現頻度上位3位までの合計で殆ど7-8割に達する。検索の際，キーワードの役割を表現するのに出現頻度の高い関係子を利用するのが有効であろうと推測されるので，利用手法を検討中である。(4)本方法は現時点では論文表題中のキーワードについてのみ可能なので，要求主題を主たる内容とする論文についてのみ適用できる。また本方法は1次情報からの不要データの除去法としても有効である。発生するノイズは事前に予測し難い事が多いので，1次情報をサンプル出力してノイズの出方を見た上で2次検索として除去操作を行う事が望ましい。本方法は簡単な手法で合目的性の高い資料の入手法として有効であり，現場研究者への検索法の教育に利用し得るものと考える。現在検討中のサイレンスキーワードと並列キーワードの処理法が実用化されれば適用範囲が広がるであろう。

キーワードの切り出しについてご教示頂いた㈱ディック・アルファ，テクニカルスーパーバイザー北井隆氏に深謝申し上げる。

参考文献

1）海老沼幸夫：適合情報利用によるオンライン高性能自動文献検索法：情報管理，Vol. 27, No. 8, p. 692-703 (1984).
2）H. P. Frei and Schaeuble: Determining the effectiveness of retrieval algorithms: Inform. Process. Manag. Vol. 27. No. 2, p. 153-164 (1991).
3）岩垂　司，三国景史，小銭正尚，三枝武男：データベース検索における出力データ処理事例の収集と分析：第29回情報科学技術研究集会発表論文集，p. 131-136 (1992).
4）北井隆：日本文からのキーワードの切り出しについて，その1，2，3：JOISニュース，No. 64，p. 5-8，No. 65, p. 4-8，(1990), No. 7, p. 6 (1991).
5）JICST編：和文・和訳表題の切り出し法：JOIS活用の手引きII，JICST系ファイル，p. 29-31 (1991).