JSTデータベース検索における索引の有効性と索引作業の重要性

JSTPlusファイル, JMEDPlusファイルにおける索引語の分析

Efficacy of indexes of JST databases search and importance of indexing

Analysis of index terms of JSTPlus file and JMEDPlus file

堀内 美穂¹ | 中村 徹¹ | 永井 賢吉¹

HORIUCHI Miho¹; NAKAMURA Tohru¹; NAGAI Kenkichi¹

1 独立行政法人科学技術振興機構 文献情報事業本部 文献情報部 (〒102-8666 東京都千代田区四番町5-3)
Tel : 03-5214-8470

1 Department of Literature Information, Office of Science and Technology Information, Japan Science and Technology Agency (5-3
Yonban-cho Chiyoda-ku, Tokyo 102-8666)

原稿受理 (2009-02-13)

(情報管理 52(1), 032-042)

著者抄録

科学技術振興機構 (JST) が作成, 提供しているJSTPlusファイルおよびJMEDPlusファイルを用いて索引に関する調査を実施した。検索時に利用される検索フィールドといった検索動向の分析や検索テーマ (52テーマ) ごとの検索フィールドによる再現率・適合率への寄与の調査分析を行い, 利用者は索引を意識して検索することは多くないが, 気付かぬうちに索引およびシソーラスのメリットを受けていること, 索引を検索対象とすることで検索精度が向上し, 書誌のみではたどり着けない文献も検索できていることを確認した。また, 論文の標題・抄録・全文中における索引語の出現割合の調査により, 索引語の4割は, その同義語も含めて論文中には存在しないものであり, 索引者が本文から判断して付与していることを明らかにした。これは人手による索引の有効性を意味する。

キーワード

文献データベース, 情報検索, シソーラス, 索引語, JSTPlusファイル, JMEDPlusファイル, JDream II, 索引作業

1. はじめに

科学技術振興機構 (JST) は, その前身である日本科学技術情報センター (JICST) の時代から30年

以上にわたって科学技術文献データベースを作成している。検索インタフェースは, 1976年に提供を開始したJOISシステム, 2003年に提供を開始したエンドユーザー向けのJDream, そして現在提供してい

るJOISとJDreamを統合したJDream II と時代に応じて変わっているが、提供データの中心が抄録を含む書誌事項とシソーラス¹⁾に基づく索引であることは変わっていない。かつては原稿用紙に抄録や索引を手書きしていたものが、現在は抄録・索引支援システムを使用しているものの、最終的には索引者が索引語を選択して付与していることも変わりはない。しかしながら、検索システムの変化により、検索方法には変化が見られる。例えば、以前のJOISでは標題や抄録中の用語は、独自の切り出しシステムによる切り出し語を検索タグとしていたが、JDream II システムでは標題や抄録のフルテキスト検索が可能であることから、利用者は以前に比べて検索フィールドを意識しない検索を行っていることが予想される。また、エンドユーザーの利用も増えていることから、シソーラスそのものを意識せず、あるいは索引の存在を知らずに検索を行っている場合も多いであろうことも予想される。シソーラスに基づいた統制語による索引については、これまでさまざまな場面でその有効性が述べられている。Google等の検索エンジンを用いた検索が一般に広く普及した現時点で、旧来のシソーラスを用いて索引が付与されたデータベースの有効性について検証することで、索引を意識しないで利用している利用者に索引の価値を知ってもらいたいと考えた。そこで、JSTのデータベースを対象にいくつかの調査を試みたので、その結果を報告する。調査の内容は、利用者の検索動向、索引を検索対象とした場合としなかった場合の再現率・適合率の変化である。また、論文の標題・抄録・全文中における索引語の出現割合についても調べた。

2. 調査対象データベース

JSTが作成しているJSTPlusファイルおよびJMEDPlusファイルを調査対象とした。これらのファイルはJDream II を利用して検索可能である。JSTPlus

ファイルは、科学技術全分野に関する文献情報であり、世界50数か国の情報を含んでいる。JMEDPlusファイルは、日本国内発行の資料から医学、薬学、歯科学、看護学、生物科学、獣医学等に関する文献情報を収録している。

抄録を含む書誌事項をデータとして収録しているが、本文言語が日本語であるか否かにかかわらず、日本語標題、和文抄録データを付与している（ただし、会議要旨集については抄録は付与していない）。抄録は、抄録作成者による独自の抄録のほか、許諾を得られたものについては著者抄録も使用している。

また、これらの文献には主題分析の結果、シソーラスに基づき索引語が付与されており、上位語を含めて検索タグとなっている。さらに、分類コード、準シソーラス用語（シソーラスには存在しないが索引者が重要であると判断した用語であり、索引語の一つとして付与している）、化学物質名も付与される。以降、シソーラス用語、準シソーラス用語、化学物質名、分類コードを合わせて索引と呼び、付与された用語を索引語と呼ぶ。これら索引語の付与方法は『情報管理』vol.50, no.4「索引作業におけるより高度な支援辞書の利用～JST抄録・索引支援システム「NAISS」について～」に詳しく述べられている²⁾。

3. 調査1：利用者の利用動向

3.1 調査内容

利用者の利用動向を調べるため、JDream II におけるJSTPlusファイルおよびJMEDPlusファイル1か月間（2007年8月）の利用統計を用いて、利用された検索フィールドを調べた。JDream II は検索上級者向けのコマンドモードと初心者向けのシンプルモードを提供しているため、それぞれ分けて値を取得した。利用者は「あらかじめわかっている書誌事項から特定の文献をヒットさせる検索」を行う場合と、「あ

図1 JSTPlus, JMEDPlusファイルの利用状況
コマンド利用回数割合 (著者名, 巻, 号等の書誌事項を除く)

るテーマに関する文献を探す」場合の大きく分けてふた通りの検索を行うことが考えられる。ここでは後者のような検索を行う場合に、索引を意識して検索しているのかどうかを調べるため、全体の統計から、前者のような検索で使用される著者名, 巻, 号, 資料名等の書誌事項の利用を除いたものを分析対象とした。また, 利用者が入力している言葉とその頻度についても調べた。

3.2 結果と考察

各ファイルの利用状況を図1に示す。ファイルの種類にかかわらず, コマンドモードとシンプルモードのいずれも, 検索フィールドを指定しないキーワード検索 (和文標題, 和文抄録, 下位語を含むシソーラス用語, 準シソーラス用語, 化学物質名が検

索対象) が主流であり, コマンドモードでは60%以上, シンプルモードは85%以上を占めていることがわかった。キーワード検索以外の検索の内訳は, 約6割が標題に限定した検索であり, 残り4割が索引を指定した検索であった。

この結果のみを見ると利用者が索引に注目をして検索を行っているとは考えにくい, 利用者へは, データベースの使い方として, キーワード検索でまず検索を行い, 必要に応じて標題での絞り込み検索を行うことを研修会などで勧めていることから, この結果は予想通りと言える。

次に, キーワード検索で利用者が入力している検索語を調べたところ, 1か月間に約15万種類の言葉が入力され (表1), そのうち検索頻度が月50回以上の言葉は約3,500種類あった。うち42%はシソー

表1 キーワード検索に入力される検索語の種類数とシソーラス用語数

キーワード検索入力頻度 (1か月間)	検索語の種類数 (A)	検索語の種類数 累計 (a)	a/総数	(A)に含まれていたシソーラス用語数 (B)	シソーラス用語数 累計 (b)	(B)のうち下位語を持つシソーラス用語数 (c)	下位語を持つシソーラス用語の割合 (c/B)	検索語の種類数に占めるシソーラス用語の割合 (B/A)	検索語の種類数に占めるシソーラス用語の割合 累計 (b/a)
50回以上	3,457	3,457	2%	1,459	1,459	915	63%	42%	42%
20~49回	6,033	9,490	6%	1,715	3,174	802	47%	28%	33%
10~19回	9,831	19,321	13%	1,890	5,064	689	36%	19%	26%
1~9回	130,162	149,483	100%	8,227	13,291	2,386	29%	6%	9%
合計	149,483			13,291		4,792	36%		

参考) シソーラス用語全体の下位語を持つ割合は27%

ラスにある言葉であり、その6割以上は下位語を持つシソーラス用語であった。シソーラス収録語全体で下位語を持つシソーラス用語の割合が27%であることを考えると、下位語を持つシソーラス用語ほど利用される頻度が高い。すなわち、利用者は比較的上位概念の語を用いて検索を行う傾向にあることがわかる。下位語を含むシソーラス用語で検索した場合、下位語が付与された文献もヒットすることから、漏れの少ない検索ができる。

このことから、利用者は検索フィールドやシソーラス用語を意識せずに検索しているものの、入力される用語にはシソーラス用語、特に下位語を持つシソーラス用語も多く、索引は有効に活用されていることが確認できた。

4. 調査2：サンプル検索による検索効率の分析

4.1 調査方法

シソーラス用語が実際の検索で使用されていることは調査1の結果から明らかとなったが、さらに、検索における索引の寄与を明らかにするため、実際の検索に近い検索テーマに対して索引を用いない検索式と用いた検索式を作り、検索フィールドごとのヒット件数と再現率、適合率を求め、比較した。

使用した検索テーマは、「JSTPlusファイル検索事例集2008年版」および「JSTPlusファイル活用テキスト2008年版」（以下、事例集）に記載されているテーマ（46テーマ）と予備調査のために独自に作成した検索式6テーマの合計52テーマ（表2）である。検索はJDream II のコマンドモードで実施し、異

表2 検索テーマ一覧

1. エネルギーの未来	27. 電気二重層キャパシタ
2. 高架・立体交差	28. 体細胞に由来する多能性幹細胞, iPS細胞
3. 高速道路トンネル内の換気	29. 文化財保護, 歴史的景観・古建築の保存
4. 事故防止用のセンサー, レーダー	30. グリホセートのダイズにおける吸収・代謝
5. 自動車用耐油性特殊ゴム	31. グルタミン酸ナトリウム
6. 酒類についての総説記事	32. ニトロメタンの危険性および事故事例
7. 小麦の品種の違いが製パン性に及ぼす影響	33. 食品中のGABA
8. 日本における石油・ガスの地下貯蔵	34. 窒化ガリウム系半導体を用いた発光ダイオード
9. 地球外知的生命の探査	35. マグネシウムを含んだアルミニウム合金の航空機産業への利用
10. 半導体洗浄技術	36. ウリミバエのガンマ線不妊化処理
11. 量子コンピューター	37. 機能性高分子(2007年以降とする)
12. 米国人の生活習慣病	38. La _{0.9} Sr _{0.1} MnO ₃ の検索
13. ヒアルロン酸の抗老化, アンチエイジング用化粧品としての応用, 効果	39. クロム規制とクロムフリー技術
14. ホンダ“ASIMO”のような二足歩行ロボット	40. 小児における偏頭痛と起立性低血圧について
15. サラブレッドの調教	41. POEMS症候群
16. つり橋の地震による震動応答解析	42. ノルバスクについて
17. ドーム球場の非常用設備	43. パクリタキセルの副作用について(2006年以降)
18. ネットワーク犯罪	44. 北海道におけるスパイクタイヤの被害および防止策
19. バイオエタノール産生における, ベータグルコシダーゼの役割	45. 橋梁の建設に使用されるアーク溶接ロボット
20. プラスチックごみ, 廃棄物のリサイクル	46. ワイン中のポリフェノールのHPLCによる分析
21. 韓国語の機械翻訳	47. トキの人工繁殖
22. 道路交通における最短時間経路の決定	48. 鉄道の事故対策・安全管理
23. 流動層触媒反応装置の数学モデルとシミュレーション	49. 抗体医薬の副作用
24. サッカーボールや野球の変化球に働くMagnus力	50. イノシシの防除(被害対策)
25. ICタグ	51. 風力発電の立地条件
26. 常圧での脂肪族アミンの合成	52. 砂質地盤の液化化に及ぼす異方性の影響

40～43はJMEDPlusファイルで検索, それ以外はJSTPlusファイルで検索

47～52は事例集にはない独自に作成した検索テーマ

表記は自動展開しなかった。使用するデータベースは、JSTPlusファイルあるいはJMEDPlusファイルであり「事例集」の各テーマが使用しているデータベースを使用した。また、比較する上で基準を合わせるために、検索対象データは、抄録が付与されているもので、索引が完了しているものとした。

具体的な検索方法を、検索テーマの一つ「ワイン中のポリフェノールのHPLCによる分析」の検索例を用いて説明する。

まず、検索テーマから切り出した単語を用いて、キーワードを作成した。これは、エンドユーザーが検索を行う際に最初に行う行為と考えられる。このとき、検索テーマにある「分析」「影響」「効果」などの言葉はキーワードとしなかった。

これらのキーワードを用いて、標題 (/ti) の検索を行った。

ワイン/ti and ポリフェノール/ti and HPLC/ti

続いて、検索対象を標題+抄録 (/ab) として検索を行った。

(ワイン/ti or ワイン/ab) and (ポリフェノール/ti or ポリフェノール/ab) and (HPLC/ti or HPLC/ab)

次に、各キーワードに対応するシソーラス用語を調べ、標題+抄録+索引 (/ct) を検索対象として検索を行った。このとき、準シソーラス用語 (/st)、化学物質名、日化辞番号も索引として検索対象とした。

(ワイン/ti or ワイン/ab or ぶどう酒/ct or ワイン/st)

and

(ポリフェノール/ti or ポリフェノール/ab or 多価フェノール/ct or ポリフェノール/st)

and

(HPLC/ti or HPLC/ab or 高速液体クロマトグラフィ /ct or HPLC/st)

「事例集」に掲載されている検索式を検索上級者が作成した適切検索式とし、適切検索式によりヒッ

トした文献を「適切文献」とした。この適切検索式では、前方一致、分類コード等による検索も駆使されている。また、このときは異表記を自動展開させた。以下がこの例における適切検索式である。

(ワイン or ぶどう酒 or ブドウ酒 or WINE/tien)

and

(ポリフェノール or 多価フェノール or POLYPHENOL/tien)

and

(HPLC or 高速液体クロマト or HPLC/tien or HIGH PERFORMANCE LIQUID CHROMATO?/tien)

作成した検索式でヒットした文献と、「事例集」より求められる「適切文献」を付き合わせ、作成した検索式でヒットした文献中の「適合文献数」をカウントし、再現率、適合率を求めた。再現率および適合率は以下の計算式を用いて計算した。

$$\text{適合率 (\%)} = (\text{適合文献数}) / (\text{ヒットした文献数}) \times 100$$

$$\text{再現率 (\%)} = (\text{適合文献数}) / (\text{適切検索式による適合文献数}) \times 100$$

検索結果は、検索テーマごとに、表3のようなシートを作成し管理した。

4.2 結果

それぞれのテーマの再現率、適合率を二次元にプロットした結果は図2のようになった。

標題のみの検索では、適合率は高いものと、低いものの両極端に分かれ、再現率は低くなった。抄録も追加して検索を行うと、各テーマの再現率、適合率を示す点が散らばってくることがわかる。索引の検索も加えた場合は、適合率、再現率が全体的に高くなっている。図2において矢印が示す点が全テーマの平均値である。

標題のみでの（再現率、適合率）の平均は（6%, 63%）であるが、抄録を含めると（24%, 70%）であった。さらに索引を含めた場合は（74%, 80%）となり、索引を検索対象とした場合、再現率が顕著に増加す

表3 検索シート例

検索テーマ	ワイン中のポリフェノールのHPLCによる分析									
適切検索式	(ワイン or ぶどう酒 or ブドウ酒 or WINE/tien) and (ポリフェノール or 多価フェノール or POLYPHENOL/tien) and (HPLC or 高速液体クロマト or HPLC/tien or HIGH PERFORMANCE LIQUID CHROMATO/tien)									
適切文献数	393									
	辞書なし	ヒット件数	適合件数	再現率	適合率	辞書あり(同義語, 異表記)	ヒット件数	適合件数	再現率	適合率
標題	ワイン/ti and ポリフェノール/ti and HPLC/ti	3	3	1%	100%	(ワイン/ti or ぶどう酒/ti) and (ポリフェノール/ti or 多価フェノール/ti) and (HPLC/ti or 高速液体クロマトグラフィー/ti)	7	7	2%	100%
抄録	ワイン/ab and ポリフェノール/ab and HPLC/ab	28	28	7%	100%	(ワイン/ab or ぶどう酒/ab) and (ポリフェノール/ab or 多価フェノール/ab) and (HPLC/ab or 高速液体クロマトグラフィー/ab)	66	66	17%	100%
標題+抄録	(ワイン/ti or ワイン/ab) and (ポリフェノール/ti or ポリフェノール/ab) and (HPLC/ti or HPLC/ab)	30	30	8%	100%	(ワイン/ti or ぶどう酒/ti or ワイン/ab or ぶどう酒/ab) and (ポリフェノール/ti or 多価フェノール/ti or ポリフェノール/ab or 多価フェノール/ab) and (HPLC/ti or 高速液体クロマトグラフィー/ti or HPLC/ab or 高速液体クロマトグラフィー/ab)	72	72	18%	100%
索引(アップポ スティング, ST, 化合物を含む)	(ぶどう酒/ct or ワイン/st) and (多価フェノール/ct or ポリフェノール/st) and (高速液体クロマトグラフィー/ct or HPLC/st)	294	294	75%	100%					
標題+抄録+ 索引	(ワイン/ti or ワイン/ab or ぶどう酒/ct or ワイン/st) and (ポリフェノール/ti or ポリフェノール/ab or ポリフェノール/st or 多価フェノール/ct) and (HPLC/ti or HPLC/ab or 高速液体クロマトグラフィー/ct or HPLC/st)	364	364	93%	100%	(ワイン/ti or ぶどう酒/ti or ワイン/ab or ぶどう酒/ab or ぶどう酒/ct or ワイン/st) and (ポリフェノール/ti or 多価フェノール/ti or ポリフェノール/ab or 多価フェノール/ab or 多価フェノール/ct or ポリフェノール/st) and (HPLC/ti or 高速液体クロマトグラフィー/ti or HPLC/ab or 高速液体クロマトグラフィー/ab or 高速液体クロマトグラフィー/ct or HPLC/st)	383	383	97%	100%
索引#(標題+ 抄録)		274	274	70%	100%					

備考：上記ヒット件数は、抄録あり索引付与済の文献数である

図2 サンプル検索における再現率、適合率の変化

図3 サンプル検索において適切文献数を100とした場合の適合率とノイズ

図4 サンプル検索において適切文献数を100とした場合の適合率とノイズ (大規模辞書を使用した場合)

ることがわかった。

また、適切検索式による適切文献数を100とした場合に、上記の適合率とノイズ（ヒットした文献の中で適切文献以外のもの）の平均をグラフにすると図3となった。

標題と抄録でヒットしたものを除き、索引のみでヒットしたものの再現率の平均は40%であり、索引がなければヒットしない文献もあることがわかった。

4.3 大規模辞書の効果

さらに、キーワードの同義語や異表記語を追加して検索した場合の再現率、適合率の変化を調べた。JDream II の機能であるJSTシソーラスブラウザおよび日本語異表記辞書ブラウザを用いて、JSTが作成する大規模辞書¹⁾に登録されている同義語、異表記語を調べ、検索式に同義語、異表記語を追加し検索を行った。大規模辞書は、科学技術用語の同義語や表記のゆれ（異表記）を整理したもので、平成20年

12月末時点で、約50万語が登録されている。大規模辞書を使用することで、図4の通り標題や抄録の検索では、ノイズの増加に比べて再現率はより増加した。大規模辞書を使用した場合と使用しなかった場合（4.1の調査結果）の再現率および適合率を表にまとめると表4の通りとなった。

4.4 考察

今回の検索で適切検索式とした「事例集」に掲載されている検索式は、検索フィールドのすべてを検索対象としたものであるため、作成した検索式においても、検索フィールドを増やすほど再現率が高くなるのは当然の結果ではある。しかし、漏れのない検索を行うような場合には、標題だけではなく、抄録や索引も検索できるということが重要であると言える。特に、索引のみでヒットした文献の再現率の平均が40%であったことは、標題や抄録のフルテキスト検索だけでは見つけることができない文献があるということを示しており、索引が有用であること

表4 サンプル検索における再現率、適合率の平均値

	タイトル		タイトル+抄録		タイトル+抄録+索引		索引でのみヒット	
	再現率	適合率	再現率	適合率	再現率	適合率	再現率	適合率
大規模辞書なし	6%	63%	24%	70%	74%	80%	40%	87%
大規模辞書あり	9%	60%	35%	70%	81%	72%	—	—

がわかる。

一方、索引を使用し、さらに大規模辞書による同義語を加えた検索式の場合でも再現率が100%にならないものがほとんどであった。これは、適切検索式では前方一致検索などの機能や分類コードなどの検索フィールドも使用されていたり、テーマから推察される関連語なども含まれているためである。このあたりも含めて検索するところが検索上級者の腕の見せどころである。

5. 調査3：論文の標題・抄録における索引語の出現頻度

5.1 調査内容、手順

前章までは検索という行為から見た索引語に関する分析であったが、現在JSTデータベースで実施している索引作業の必要性についての観点からも調べてみることにした。

索引者は全文に目を通して索引付けを行っているが、その結果付与される索引語は標題、抄録にどの程度出現しているものなのかを調べた。もし、索引された言葉のほとんどが標題や抄録にある言葉なのであれば、標題や抄録からのシソーラス用語の抽出が索引者による索引の代替となりうる可能性があることになる。

JSTPlusファイル、JMEDPlusファイルに2008年に

収録された文献のうち、それぞれ12,000件を基に分析した。索引されたシソーラス用語および準シソーラス用語と同じ文字列が標題あるいは抄録にあるかどうかを機械的に比較し、索引された言葉の数だけこの処理を繰り返した。

この調査においては、索引語の中でも自動的に付与される構造上位語は対象とせず、人手による索引語のみを対象とした。構造上位語とは、化学物質名に対する組成・構造に着目した上位概念の用語のことで、これらはデータベース作成の工程で自動的に付与されている。

また、著者が標題や抄録を書く際にはシソーラス用語を使用しているとは限らないため、大規模辞書を使用して、索引語の同義語、異表記語が標題、あるいは抄録にあるかどうかについても同様に調べた。

5.2 結果

索引したシソーラス用語と同じ文字列が標題あるいは抄録中に存在した割合は日本語文献の場合、JSTPlusファイルで39%、JMEDPlusファイルで37%であり、ファイルによる差は見られなかった。本文が英語の場合、索引語が標題あるいは抄録に存在した割合は、JSTPlusファイルで43%、JMEDPlusファイルで45%とやや高い傾向にあるが、総じて本文の言語による差はないといえ、約4割の索引語が標題、あるいは抄録に存在していることがわかった(表5)。

表5 標題・抄録における索引語の出現頻度

		記事数	CW総数	1記事平均 (CW語数)	索引された CWが標題 あるいは抄 録にあった 数	索引された CWが標題 にも抄録に もなかった 数	ST総数	1記事平均 (ST語数)	索引された STが標題あ るいは抄録 にあった数	索引された STが標題に も抄録にも なかった数
JSTPlus	日本語文献	4,427	46,951	11	18,238 39%	28,713 61%	13,390	3	8,133 61%	5,257 39%
	日本語文献以外	7,640	90,237	12	38,733 43%	51,504 57%	16,492	2	10,072 61%	6,420 39%
JMEDPlus	日本語文献	11,171	108,224	10	39,995 37%	68,229 63%	46,486	4	21,794 47%	24,692 53%
	日本語文献以外	891	8,599	10	3,884 45%	4,715 55%	3,347	4	1,715 51%	1,632 49%

(注) CW：下位語を含まないシソーラス用語

ST：準シソーラス用語

表6 標題・抄録における索引語（同義語を含む）の出現頻度

	記事数	索引語総数	標題か抄録に索引語か 索引語の同義語があった 索引語数	標題にも抄録にも索引語も 索引語の同義語もなかった 索引語数
JSTPlus	12,067	137,188	68,112 50%	69,076 50%
JMEDPlus	12,062	116,823	55,854 48%	60,969 52%

表7 全文における索引語（同義語を含む）の出現頻度

	記事数	索引語 総数	JDream II		J-STAGE	
			標題か抄録に索引語か 索引語の同義語があった索引語数	標題か抄録に索引語か 索引語の同義語があった索引語数	「全文のどこか」に索引語があった索引語数	「全文のどこか」に索引語か索引語の同義語があった索引語数
J-STAGEに全文がある 記事(索引語のうち、構造 上位語を除いた場合)	90	1,164	510 44%	599 51%	591 51%	712 61%

大規模辞書収録のシソーラス用語の同義語が標題あるいは抄録に出現していたものも含めるとJSTPlusファイル、JMEDPlusファイルでそれぞれ50%、48%の索引語が標題あるいは抄録に出現していた(表6)。

準シソーラス用語が、標題あるいは抄録に出現する割合はシソーラス用語の割合に比べて高く、JSTPlusファイルで約6割、JMEDPlusファイルで約5割であった。

5.3 考察

この結果から、索引語あるいはその同義語は、おおむね50%程度が標題あるいは抄録に出現していると言える。

大規模辞書は継続的に用語の追加を行っているため、今後のさらなる辞書の整備でその割合が高くなる可能性はあるものの、索引語が標題、あるいは抄録に存在するのは約6割にとどまると考えられ、残り約4割の索引語についてはその同義語も含めて、標題あるいは抄録には存在しないと言える。この約4割の索引語は、索引者が文献の本文あるいは文脈から選定していると考えられる。そこで次章では調査対象を全文に広げた場合の索引語の出現頻度について報告する。

6. 全文中における索引語の出現頻度

6.1 調査内容、手順

JSTPlusファイルおよびJMEDPlusファイルに収録されている日本語文献で、J-STAGE(科学技術情報発信・流通総合システム)から全文が提供されているもののうち90記事を対象に、JSTPlusファイルおよびJMEDPlusファイルの索引語(自動的に付与される構造上位語を除く)が全文中に出現する割合を調べた。また、5章の調査との整合性を確認するために、JSTPlusファイルあるいはJMEDPlusファイルに収録されているこれら90記事の標題および抄録を対象に5章と同様の調査も行った。さらに、全文中に出現しなかったシソーラス用語については、3章で使用した統計データを基に利用頻度を調べた。

6.2 結果

JSTPlusファイル、JMEDPlusファイルに収録されている90記事の中の索引語のうち、索引語と同じ文字列が標題、あるいは抄録に出現する割合は44%であり、大規模辞書を用いて同義語、異表記語までを含めた場合は51%であった。

全文に対象を広げて調べたところ、索引語のうち51%は標題、抄録を含む全文中に存在し、大規模辞書を用いて同義語までを含めた場合は、61%が存在

した（表7）。全文に出現しない索引語で検索頻度が高いものを調べたところ、ヒト、老人、症例報告、成人、腫瘍、二酸化炭素、画像処理、薬物療法などであった。

6.3 考察

5章の調査とこの章の調査は調査対象が異なるが、JSTPlusファイルおよびJMEDPlusファイルに収録されている90記事の索引語が標題あるいは抄録に存在する割合は、5章の結果とほぼ同じであったため、5章の結果とこの章の結果は比較可能であると判断した。

全文中に存在しない4割の索引語は、チェックタグ的なものや、シソーラスで定義している用語の下位語と考えられるがシソーラス用語となっていないためその上位語を付与しているもの、あるいは文章中には用語が出てこないがその文章が示す概念を付与しているものなどであり、機械的なシソーラス用語の抽出だけでは現時点の索引の代替とはならないと考えられる。

7. 全文中に存在しない索引語の利用状況

全文に出現した索引語と出現しなかった索引語について、利用者の検索頻度を比較したところ、後者の検索回数は前者に比べ1.7倍であった。全文に出現しない索引語には、6.2で示したようにチェックタグ的に使用される用語が含まれていることから、これが平均検索回数を押し上げているという事実もあるが、全文に出現しない4割の索引語も検索における有用性では全文中に出現する6割と差がないと考えられる。

8. 結論

以上の結果から、シソーラスに基づく索引は検索において有効であり、利用者は索引を意識して検索していない場合でも、下位語を含めて検索できるというメリットを受けていること、書誌データのみではたどり着けない文献も検索できていることが確認できた。また、付与されている索引語の4割については、その同義語も含めて論文中には存在しておらず、索引者が本文から判断して付与しているものであり、現時点では、標題や抄録等からの機械的なシソーラス用語の抽出だけでは索引者による索引の代替とはならないことがわかった。

9. 最後に

シソーラスに基づく索引の有効性についてはこれまでも述べられているところであるが、今一度、客観的にその有効性を検証したいと考え、今回の調査を行った。今回の調査の結果、索引がJSTデータベースの付加価値を上げていることが明らかになった。現在、JSTデータベースでは、索引者が全文を読んで主題分析を行い人手により索引作業を行うことで、索引の一定のレベルを確保している。一方、日本語処理の分野ではさまざまな研究が行われていることから、全文からの重要語抽出の研究が進み実用化が可能となった場合は、索引の自動化についても検討していきたいと考えている。

今回の調査では、調査方法を試行しつつ実施したため、各調査によって調査対象の母集団に違いがあるが、整合性が確認できるよう調査したつもりである。調査の結果出てきた数値の統計処理については、いささか素人的な部分もあるので、調査方法を含めてご意見などいただければ幸いである。

参考文献

- 1) 岩沢一男, 倉田重子, 山崎文枝, 中村徹. JST科学技術用語シソーラスの改訂—2008年版シソーラスと大規模辞書との連携—. 情報管理. 2008, vol. 51, no. 2, p. 129-137.
- 2) 富永祥平. 索引作業におけるより高度な支援辞書の利用～JST抄録・索引支援システム「NAISS」について～. 情報管理. 2007, vol. 50, no. 4, p. 210-217.

Author Abstract

In this study we analyzed the efficacy of index terms of JSTPlus and JMEDPlus databases. Those databases are produced and provided by Japan Science and Technology Agency (JST). Statistical analysis of user behavior showed that not many users were aware of indexes but got the benefit of indexes and thesaurus. From 52 theme searches, analysis of recall ratio and precision ratio showed that indexes raised retrieval precision. And by using indexes, users could retrieve the articles those were not able to be retrieved only from bibliographic data search. Appearance ratio of index terms in titles, abstracts and full texts showed that 40% of index terms and their synonyms didn't appear in full text. It showed that 40% of index terms were indexed by indexer's subject analysis of full text. These analyses meant efficacy of indexing by hands.

Key words

literature database, information retrieval, thesaurus, index term, JSTPlus file, JMEDPlus file, JDreamII, indexing