D-042

技術文書管理支援システム検索用キーワード用辞書の検討について

Study of keywords for JAXA Digital Archives

田中 陽子† 祖父江 真一†
Yoko Tanaka Shinichi Sobue

1．まえがき

旧宇宙開発事業団において創立以来蓄積されてきた 60 万件にのぼる研究開発および開発成果に関する技術文書の蓄積、共有、活用を効率的に行うことを目的として、宇宙航空研究開発機構（JAXA）情報化推進部では技術文書管理支援システム(DARC)を整備・運用している。この DARC に保存されている技術文書には文書の作成者により任意にキーワードが付与され、検索キーワードとして利用できるようになっている。しかし、文書の特徴を表す適切なキーワードが付与されていない、正規化されていないなどの理由により、キーワードによる検索は、多くヒットしすぎたり、逆に少なすぎたり、文書の絞り込みに役立っていない。このため、JAXA においては、作成者が付与したキーワードと全文検索ソフトウェアが本文の特徴抽出をして作成したキーワードを比較・分析を実施したので、その結果について報告する。

2．現状

(1) 属性情報による検索

技術文書管理支援システム（JAXA Digital Archives: DARC)は、文書のタイトルや作成日やキーワードなどの属性情報を検索するシステムであり、MS-Word などのファイル形式の文書からテキストデータを抽出しての全文検索は行っていない。このため必要とする文書をキーワード等で検索するために、文書を特徴付ける適切なキーワードを付与することがポイントとなり、これによりほしい情報の絞込みが可能となる。

(2) キーワードの付与者

現在 JAXA では、文書へのキーワードの付与は文書の作成者自身が行っており、JAXA にはキーワード設定のガイドラインやキーワード用の辞書がないため、キーワードのつけ方は個人に依存し、キーワードの数や内容に個人差が生じている。

(3) ヒアリング結果

キーワードを利用して検索を行った結果について、システム運用者、利用者に対してヒアリングを行ったところ、検索結果が多すぎてほしい情報の絞込みができない、検索結果が少なすぎてほしい情報を得ることができないといった、相反する問題点を抱えていることがわかった。

(4) 仮説

以上より、キーワード付与方法に課題があると考え、以下の仮説を立てた。

a. 本文を表す特徴的なキーワードが選定されていない
b. 偏った語をキーワードとして選定している
c. キーワードが登録されていない文書が多い
d. キーワードの表記ゆれが大きく検索漏れが生じる

† 宇宙航空研究開発機構, JAXA

3. キーワード一覧の作成

(1) キーワード一覧の作成

本研究においては、キーワード調査・分析の母数を増やすため、DARC に加えて、2 つのプロジェクト用情報管理システム（超高速インターネット衛星（WINDS）プロジェクト情報管理システム（PIMS）と宇宙ステーション用統合文書ライブラリ（TBL））について、登録文書のキーワードフィールドから単語を抽出し、キーワード一覧の作成を行った。なお、PIMS と TBL についても、キーワード付与は文書作成者本人が任意に付与をしている。各システムの登録文書数、キーワード数などを表 1 に示す。

表 1　各情報システムとキーワード

	DARC	PIMS	TBL
登録文書数	227,657	3,524	76,235
キーワード付文書数	34,074	3,471	6,465
キーワード種類	26,976	2,283	6,692
最大キーワード数	33	-	25

表 1 から DARC,TBL は登録文書に対して非常にキーワード付与数が少なく、また PIMS は登録件数の 18%の文書に同一のキーワード（「超高速インターネット衛星」及び「WINDS」）が付与されている。これらの事実は、２項の現状で立てた仮説のｂとｃを証明するものである

(2) キーワードの分類／体系化

（1）で行ったキーワード一覧作成の結果を踏まえ、以下の手順で、キーワードの分類／体系化を行った。

①正規化の実施

詳細は、英数字は全て半角小文字へ、半角カナは全角カナへ、“ー”（全角マイナス）などの一般記号は半角へ変換したところ、表２の結果が得られ、キーワードの表記ゆれがみられた。

表２　正規化後のキーワード数

	DARC	PIMS	TBL
キーワード(正規化後)	26,591	2,169	6,425

②略語集に含まれる専門語と一般語の分類結果

正規化後のキーワードについて、JAXA が所有する約 8500 語からなる略語集に含まれる専門語と、それ以外の一般語に分類した。

表３　専門語の分類

	略語集	DARC	PIMS	TBL
1) 略語集 (略語)	5,663	1,055	146	359
2) 略語集 (正式名称)	8,743	727	122	205

表４　一般語の分類

	DARC	PIMS	TBL
3) 表３の 1),2)に出現せず	24,823	1,904	5,866
4) 3)の内スペルチェックを通った英単語	813	44	509
4) 3)の内スペルチェック通らない英文字列	1,401	157	606
4) 3)の内 1byte 文字のみの文字列	3,105	342	1,518
5) 3)の内 1,2byte の混合文字列	3,959	459	854
6) 3)の内 2byte 文字のみの文字列	17,759	1,103	3,494

(3) キーワードの分類／体系化の結果

表２、表３、表４より以下のことがわかった。

a. 表記を正規化して、重複した文字数を取り除くと、キーワード数が減少する。

b. 略語の方が正式名称より 1.5 倍程度、キーワード中に存在する。

c. 専門語より一般語の方が 10 倍以上のキーワードが存在する。

以上の事実は、２項の現状で立てた仮説の a と d を証明するものである。よって、２項の現状で立てた仮説の全てが証明され、現在のキーワード付与方法に問題があることがわかった。

4. キーワードの有効性の検証

(1) 本文中のキーワード出現件数の確認

市販の全文検索エンジンを用いて、キーワードの本文中での出現件数の確認を行い、現在登録されているキーワードの有効性について確認を行った。

ここでは、正規化したキーワードに文書のタイトルから単語を切り出して加えたものを用意し（表５の３項）、DARC に登録されている MS-Word 文書のうち 28,933 文書から本文の文字データを抽出し、全文検索 DB に登録を行った。今回得られたキーワードを、このデータベースに対する中間一致での検索を実施し、ヒットする文書数を求めた。表５の通り、半数近くのキーワードについては、本文中に出現していないことがわかり、本文に登場する言葉と、キーワードおよび文書のタイトルで使われる言葉との間にズレがあることを示している。

表５　キーワードと文書件名抽出単語

	DARC	PIMS	TBL
1.キーワード (正規化後)	26,591	2,169	6,425
2.タイトルから自動切り出しした単語数	111,271	3,677	49,043
3. 1 と 2 の重複を除いた数	127,273	5,201	52,972
4. 3.の内、本文中に出現箇所数が 0	72,334 (56.8%)	2,161 (41.5%)	28,970 (54.7%)

(2) 特徴語の抽出

DARC に登録されている MS-Word 文書 28,933 文書から、市販の全文検索エンジンの機能を利用して登録文書の特徴を現すキーワードを抽出する機能を用いて、標準辞書と、これまでに得られたキーワード一覧を追加した辞書を用いて特徴語抽出を行い、正規化したキーワードに文書のタイトルから単語を切り出して加えたもの（表５の３項）との比較を行った。

特徴語抽出により得られた単語と、各データベースとキーワードとの比較結果は表６のとおり。

表６　キーワードと特徴語との比較

	標準辞書	キーワード追加辞書
総抽出特徴語	15, 809	27, 565
DARC キーワードに含まれる単語	5, 690	17, 464
PIMS キーワードに含まれる単語	695	1, 581
TBL キーワードに含まれる単語	2, 851	8, 530

以上より、キーワードや文書のタイトル中に含まれる単語が特徴語との一致は、キーワード数の約 1～２割程度であり、本文を表す特徴的な用語がキーワードとして選定されていないことがわかった。したがって、現在のキーワードは、検索のための項目としては有効に機能していないといえる。

5. 今後の課題

この研究を通じて、現在のキーワード付与には、以下の問題があることがわかった。

・本文を表す特徴的なキーワードが選定されていない
・偏った語をキーワードとして選定している
・キーワードが登録されていない文書が多い
・キーワードの表記ゆれが大きく検索漏れが生じる

これらの問題を解決し、今後は、適切なキーワード設定を行うために、以下に示す対策を行い、JAXA 略語集を活用して検索用キーワード用辞書を構築していく。

・登録する文書の本文中に現われる特徴語をキーワードとして付与する。
・表記ゆれをなくすためにキーワードとして登録可能な文字の限定をする。
・JAXA 略語集を活用し、一般語でなく、できるだけ専門語をキーワードとして設定する。
・略語、正式名称の使い分けを明確化する。

なお、本文中に現れる特徴語については、作成者が予想もしていないものであることが多いため、市販の全文検索ソフトウェアなどを用いて自動抽出するなど、キーワードの設定のための工夫をしていく必要があると考える。

6. 参考文書

[1] 2004 年度技術情報データベース用辞書作成検討成果報告書、松下電器産業株式会社、2004