中国特許の中国語キーワード検索検証:
中国語を用いた特許調査の網羅性向上

○石田政司[1)], 山本光三[2)], 田畑文也[3)]
株式会社神戸製鋼所[1)], 株式会社カネカテクノリサーチ[2)], 富士フイルム株式会社[3)]
〒141-8688　東京都品川区北品川5丁目9－20
Tel: 03-5739-6420　FAX: 03-5739-6942
E-mail: ishida.seiji@kobelco.com

Effective Chinese patent search using Chinese Keywords.:
Improvement of Complete Chinese patent search using Chinese Keywords.

ISHIDA Seiji[1)], YAMAMOTO Kozo [2)] ,TABATA Fumiya[3)]
Kobe Steel Ltd. [1)], KANEKA TECHNO RESEARCH CORPORATION [2)], FUJIFILM Corporation [3)]
5-9-12, Kitashinagawa, Shinagawa-ku, Tokyo , Japan
Phone: +81-35739-6420　Fax: +81-35739-6942
E-mail: ishida.seiji@kobelco.com

【発表概要】

中国特許の中国語キーワードを用いた調査において、網羅性を上げるためには、①適切な中国語キーワードの網羅的かつ効率的な抽出、②データベースのキーワード検索ロジック（部分一致、インデキシング等）の把握が必要となる。本報告では、中国語キーワード抽出の手法事例を紹介し、更に中国特許の原語データベースの代表であるCNIPRについて、キーワード検索用インデキシングを検証した結果を紹介する。

【キーワード】

中国特許, 特許検索, 特許情報, 原語データベース, 中国語キーワード, 部分一致, ストリングサーチ, インデキシング, 網羅性, Nグラム法, 形態素解析, 中国語検索

1. はじめに

データベース（以下DB）性能の機能向上、分類付与精度の向上[1)]により、中国特許の調査環境は5～6年前に比べて飛躍的に向上してきたといえる。

しかし、中国特許調査の網羅性を上げるには、英語DB検索の漏れを補充する中国語のキーワードによる検索も依然として必要である。[2), 3)]

中国語キーワード検索においては、①適切な中国語キーワードの抽出法の検討、②DB上で認識される文字列の正確な把握、が漏れのない調査を行うために必須となる。

そこで、本稿では中国語キーワード抽出方法の検討、各DBの検索ロジックの検証を行い、更に、中国特許調査に汎用されているDBであるCNIPRを対象にインデキシング(indexing)ロジックの解析を行ったのでその結果を以下に報告する。

2. 検討内容

2-1. 中国語オンライン辞書の検討

中国特許DBで中国語を用いたキーワード検索をする場合、まずは翻訳サイトまたはオンライン辞書で中国語キーワードを調べ、テキストデータをDBにコピー・ペーストして検索してみるのが一般的と思われる。

しかし、翻訳サイトでは基本的に1つの訳語しか表示されず、異表記のバリエーションが少ないオンライン辞書もあることから、検索実務に有用な辞書の選択が望ましい。

そこで、インターネット上にある日中辞書、英中辞書、百科辞典、専門辞書の機能を調べ、中国語での網羅的なキーワード抽出に役に立つと思われる辞書をピックアップした。

2-2. マクロを用いたキーワード抽出

最先端の技術用語や非常に特殊な単語の場合、辞書に載っていないことが多い。また、外来語には様々な異表記が用いられることが多く、辞書だけでは網羅性が不十分な場合がある。

従って、辞書を用いないキーワード抽出方法の検討も必要となる。

一般的には、英語-中国語両方のデータを表示又はダウンロードできるDBを利用して、目視で英語文章と中国語文章を対比しながらキーワードを抽出するという作業が行われていると思われる。

しかし、この作業は非常に手間がかかることから、労力を軽減する手段が望まれる。

そこで、DBのダウンロードデータから特定の文字または単語の周辺文字を抽出できるツール「単切」を作成した。

このツールは Excel のマクロで動き、検索文字列と抽出開始位置、抽出文字数を指定すると検索語周辺の文字を含む KWIC 形式のリストを作成するもので、指定する文字数による、前方一致、中間一致、後方一致検索が可能である。

このツールを用いて中国語キーワードを抽出した事例を報告する。

2-3. DB毎のインデキシング確認

検証には、特許庁系DBとして、CNIPR[4)]，SIPO DB[5)]，PSS-SYSTEM[6)]を用い、商用DBとして HYPAT-i（(株)発明通信社)、専利SEARCH-i2（アイ・ピー・ファイン(株)）、Orbit.com(Questel 社)、PatBase（RWS グループ社）を用いている。 なお、商用DBの検証結果は、DB-A～D で結果を表示した。

又、検索対象は、2000 年～2012 年発行の中国公開特許を対象とし、公報型DB及びファミリー型DBもヒット数として同列に評価した。

(1) 検索フィールド毎のヒット解析
使用キーワード：糖醛酸 （ウロン酸）
①検索フィールド：
名称、要約、クレーム、明細書全文（以下説明書）

(2) DBごとのクレーム検索ヒット解析
①検索フィールド：クレーム
②使用キーワード：透明质酸（ヒアルロン酸）及び糖醛酸 （ウロン酸）
なお、CNIPRについてはワイルドカード(%)の有無による検索結果の違いも検証した。

2-4. CNIPR のインデキシング解析

CNIPR を使用して、クレームについて１～３文字ずつ先頭から分かち書きした文字列、及び中国語文章の形態素解析ツールを用いて分かち書きした文字列により検索を行い、どのような文字列がDBでヒットするかを検証した。

3. 検討結果

3-1. 中国語オンライン辞書

インターネット上の中国語オンライン辞書を調べ日中辞書9件、英中辞書30件、百科事典5件、医学辞書5件、化学辞書4件、その他4件をピックアップし検討した。

検討したオンライン辞書の中で、実務に有用と思われるものを表1.に示し、それぞれ特徴点を記載した。

上記の他に、医学・化学・コンピュータ系の専門辞書も挙げている。

表.1 中国語オンライン辞書

種 類	辞 書 名	ア ド レ ス	特 徴
日中辞書	weblio 日中中日辞典	http://cjjc.weblio.jp/	複数の辞書を縦断検索
	Naver 中国語辞書	http://cndic.naver.jp/	手書き文字入力可能
英中辞書	有道词典	http://dict.youdao.com/	事例が豊富
百科事典	维基百科	https://zh.wikipedia.org/	中国語版 wiki
専門辞書	医网打尽词典	http://dict.51daifu.net/	医学系辞書
	化工引擎 Chem YQ	http://www.chemyq.com/xz.htm	化学系辞書
	電脳漢和辞典	http://www.bjkoro.net/dict/it/	IT 系辞書

3-2. マクロを用いたキーワード抽出例

Excelのマクロ機能を活用したツール「単切」を用いて辞書に載っていない単語を調べる方法を検討した。

審査指南第二章2.2.7の規定“説明書の記載に関する他の要求”の趣旨により、説明書中の外来語については明瞭性を確保するために中国語音訳に隣接して外来語（英語等の原語）が注記される場合がある。

このように外来語が注記的に記載されている場合は、特定の文字の前後文字列を切り出して抽出するというのもキーワード検討の一つの有効な手法である。

そこで、最初の事例として抗老化蛋白として知られる“サーチュイン”（sirtuin）の中国語を抽出した。

まず、英語の“sirtuin”で特許DBを検索して中国語の文章を出力し、この出力結果から”(sirt”を検索し、その6文字前から11文字を抽出した（表.2）。

その結果、“(sirtuin)”に隣接して“沉默调节蛋白”が 34 件、“供新的瑟土因”が3件抽出された。 切り出し範囲を調節してさらに確認した結果、“サーチュイン”に該当するキーワードである“沉默调节蛋白”及び“瑟土因”が抽出された。

表.2 “サーチュイン”の中国語表記

検索語[(sirt]	**件数**
irtuin**(SIRT**	42
沉默调节蛋白**(SIRT**	34
默调节蛋白,**(SIRT**	6
供新的瑟土因(sirt	3

次に、中国語の異表記が多いキーワードの抽出例として“アルツハイマー病”の中国語キーワード抽出を試みた。

検索語を“Alzheimer”とし、オンライン辞書で予備調査したワードの中から特徴語“阿(a)”と“默(mo)”を含み、“阿”で始まり“默”で終わる単語を「単切」により抽出した。

表.3に示すとおり、“阿尔茨海默”、“阿耳茨海默”等の使用頻度の多いワードから“阿尔海默”、“阿尔采默”等の使用頻度の少ない異表記のワードまで幅広く抽出することができた。

表.3“アルツハイマー”の中国語表記

検索語［默］	件数
阿茨海默	2
阿滋海默	1
阿耳茨海默	163
阿兹海默	59
阿尔茨海默	1089
阿尔海默	3
阿尔采默	2
阿尔兹海默	51

3-3. DB毎のインデキシング検証

まず、CNIPR、PSS-SYTEM、SIPOの各検索フィールドで“糖醛酸”(ウロン酸)のキーワードを用いて検索を行い、件数を確認した。

(1)検索フィールド毎のヒット解析

表.4 検索フィールドと検索ロジック

項目	CNIPR	SIPO	PSS-SYSTEM
名称	127	127	127
要約	404	404	404
クレーム	428	不能	1456
説明書	2942	不能	2944

CNIPRとPSS-SYTEMを比べるとクレーム、説明書で件数が大きく異なることからCNIPRの名称・要約のフィールドとクレーム・説明書のフィールドは検索ロジックが異なるといえる。

(2)DB毎のクレーム検索ヒット解析

次に、“透明质酸“(ヒアルロン酸)、“糖醛酸”(ウロン酸)のキーワードで、中国語による検索が可能な複数のDBで検索しヒット件数を確認した。

表.5 クレーム検索ヒット解析

DB/ｷｰﾜｰﾄﾞ	透明质酸	糖醛酸
CNIPR	3320	428
CNIPR(ワイルドカード有)	3320	430
PSS-SYSTEM	3325	1456
DB-A	3325	1456
DB-B	3325	1456
DB-C	3224	500
DB-D	3033	1360

“糖醛酸”のキーワードでは、CNIPRとDB-Cは、ヒット数が他のデータベースよりも著しく件数が少ないことから部分一致の検索に対応していないと思われ、検索ロジックは以下のように推定される。

(3) 推定クレーム検索ロジック

表.6 推定クレーム検索ロジック

DB	推定検索ロジック
CNIPR	インデキシング型
PSS-SYSTEM	部分一致
DB-A	部分一致
DB-B	部分一致
DB-C	インデキシング型
DB-D	部分一致

全ての検索フィールドが部分一致に対応しているDBもあるが、一部のフィールドのみ部分一致に対応しているDBもあり、同じ母集団のデータに対して同じキーワードで検索しても、ヒット数が異なる場合がある。

3-4. CNIPR のインデキシング解析

(1)DB毎のクレーム検索ヒット解析

DBのインデキシングの方法として、主として辞書を使用しないで機械的に所定の文字数で分かち書きするNグラム法と、構文解析を行い解析用の辞書により単語単位で分かち書きする形態素解析とがあり、CNIPRのクレーム部分についてもいずれかの方法又は類似する方法を採用しているものと推定される。

そこで、Nグラム法に準じて1～3文字の文

字数で切出して分かち書きしたワードと、形態素解析ソフトICTCLAS[7]を用いて分かち書きしたワードがキーワード検索でどのようにヒットするかを検証した。

事例として、クレームに“糖醛酸(ウロン酸)”の文字列を含むCN1850106について解析した事例を示す。

表.7 CN1850106の解析結果

1文字切出		2文字切出		3文字切出		ICTCLAS	
1	1	1	1	1.聚	0	1	1
.	0	.聚	1	.聚甘	0	聚	1
聚	1	聚甘	0	聚甘露	1	甘露	1
甘	0	甘露	1	甘露糖	1	糖	1
露	0	露糖	0	露糖醛	0	醛	1
糖	1	糖醛	1	糖醛酸	1	酸	1
醛	1	醛酸	1	醛酸硫	0	硫酸盐	1
酸	1	酸硫	0	酸硫酸	0	在	1
硫	0	硫酸	0	硫酸盐	1	制备	1
酸	1	酸盐	0	酸盐在	0	防治	1
盐	0	盐在	0	盐在制	0	糖尿病	1
在	1	在制	0	在制备	1	药物	1
制	0	制备	1	制备防	0	中	0
备	0	备防	0	备防治	0	的	0
防	0	防治	1	防治糖	0	应用	1

表.7の数字は表中のキーワードと公報番号でAND検索を行ったヒット件数でキーワードがヒットする場合は“1”でヒットしない場合は“0”となる。

部分一致検索に対応していれば任意の文字数で分かち書きしたキーワード検索でヒットする筈であるが、ヒットが0件のものもあるため、CNIPRのクレームは部分一致検索に対応していない。

又、上記解析例では“盐(塩)”、“硫酸(硫酸)”という意味を持つ文字がヒットしていないことからNグラム法のような機械的な分かち書きでインデキシングされていない。

一方で、形態素解析を行っているICTCLASで分かち書きしたキーワードは高い率でヒットする傾向にある。

更に上記解析例と別の6件について、同様にICTCLASで分かち書きしたキーワードで索した結果、表.8に示すとおり略80%(ワード数は数字、記号、アルファベットを除く中国語のみで算出)以上の高い割合でヒットしている。

表.8 ICTCLAS抽出のキーワード検索

	抽出ワード数	ヒット数	ヒット率(%)
CN101928910A	91	88	97
CN202937406U	32	32	100
CN102516407A	46	44	96
CN102463987A	89	89	100
CN1850106	14	12	86
CN102988212A	38	35	92
CN1491660	26	20	77

(2)インデキシングのワイルドカード検索への影響

表.5で示したようにCNIPRのクレーム・説明書の検索ではワイルドカード“%”を用いても部分一致型のDBよりヒット件数が少ない場合があり、インデキシングによるデータの持ち方が検索ロジックに影響しているものとと推定される。

そこで、CNIPRの近接演算子“A pre/n B”(文字Aにn文字以内で文字Bが近接する案件を検索する演算子)を使用して、検索条件式“((%糖 pre/1 醛酸%) not (%糖醛酸%)”で検索した。 そしてヒットした案件で、“糖醛酸”の文字列を含む案件の中からCN101328228を選定し、上記表.7で示した事例と同じ方法により分かち書きを行い解析した。

クレーム中の“‥甘露糖基和葡萄糖醛酸基, ‥”の下線部は“葡萄糖/醛/酸/基”の“/”の箇所でインデキシングされており、“萄糖醛”、“糖醛酸”のように“/”を跨ぐワードは前後にワイルドカード“%”をつけてもキーワードとして必ずしも認識しないようなロジックになっている。

4. まとめと考察

(1)適切な中国語キーワードの抽出

①本稿ではいくつかのオンライン辞書を紹介した。

“Weblio日中中日辞典”は専門用語辞典

を含む複数の辞書を検索し、異表記のワードを多く得ることができる。

“Naver中国語辞書”は手書き入力できるので、テキストデータをコピーできない中国語を調べるときに便利である。

“维基百科”は中国版 Wikipedia で、中国語による説明がキーワード検討の参考になる。

このように、様々な機能・特徴を持つオンライン辞書があり、キーワード抽出に有効である。 本稿で紹介したもの以外にも有用な辞書があると思われ、使用目的に合わせて使いやすいオンライン辞書を探して使い分けるとよい。

②Excelのマクロ機能を活用したツール「単切」によりキーワードを抽出した事例を紹介した。

辞書に載っていない技術用語でも外来語（英語等の原語）が中国語に隣接して表記されている場合は「単切」により効率良く抽出することができる。

又、異表記の抽出も特徴語となる漢字を起点にして「単切」により異表記のキーワード抽出を幅広く抽出することができる。

（2）DBのインデキシング

①インデキシングの有無でキーワード検索の結果が大きく異なる。 検索漏れを回避するためには、使用するDBの各検索フィールドで部分一致検索に対応しているか否かを十分に把握した上で検索することが必要である。

②中国語キーワードを使用した調査の網羅性を上げるためには、部分一致検索に対応したDBの使用が好ましい。

やむを得ずインデキシング型DBの、部分一致に対応してない検索フィールドで検索する場合はインデキシングの分かち書きの位置によりワイルドカード検索でもヒットしない場合があるので、近接演算子を併用する等、検索式の構築の際には細心の注意が必要である。

③CNIPRのクレーム部分のインデキシング構造の解析を行った。 形態素解析により分かち書きしたキーワードはヒット率が高いことから、使用している辞書の詳細は不明であるがCNIPRのインデキシングは形態素解析を利用しているものと思われる。

5. おわりに

最後に、本報告は2013年度の「アジア特許情報研究会」のワーキングの一環として報告するものであり、研究会の皆様にデータ・情報の提供及び数々のアドバイスをいただきました。 ここに改めて感謝申し上げます。

6. 参考文献

[1]日本知的財産協会知的財産情報検索委員会.“中国特許調査に関する研究”, 知財管理, Vol.62, p67(2012)
[2]田畑文也 他.“英語・原語によるハイブリッド検索” 第 8 回情報プロフェッショナルシンポジウム INFOPRO 2011
[3]伊藤徹男.“中国特許調査におけるCNIPRデータベースの役割”
Japio YEAR BOOK2012,p144-149
http://www.japio.or.jp/00yearbook/yearbook2012.htm（参照 2013-07-24）
[4]CNIPR
http://search.cnipr.com/
[5]SIPO
http://www.sipo.gov.cn/zljs/
[6]PSS-system
http://www.pss-system.gov.cn/
[7]ICTCLAS
http://ictclas.nlpir.org/