中国語キーワードによる中国特許情報解析：
調査精度向上への応用

○安藤俊幸[1)]，金澤祐孝[2)]，小山裕史[3)]，沖　祥嘉[4)]
花王株式会社[1)]，株式会社IHI[2)]，電気化学工業株式会社[3)]，東ソー株式会社[4)]
〒131-8501　東京都墨田区文花 2-1-3
Tel: 03-5630-9538　FAX: 03-5630-9712
E-mail: ando.t@kao.co.jp

Patent information analysis using　Chinese keywords.:
Application to improve investigation precision

ANDO Toshiyuki[1)], KANAZAWA Hirotaka[2)], OYAMA Hiroshi[3)] , OKI Yoshitaka[4)]
Kao Corporation[1)], IHI Corporation[2)] , DENKI KAGAKU KOGYO KABUSHIKI KAISHA[3)] , TOSOH Corporation[4)]
2-1-3, Bunka, Sumida-ku, Tokyo 131-8501 Japan
Phone: +81-3-5630-9538　Fax: +81-3-5630-9712
E-mail: ando.t@kao.co.jp

【発表概要】

最近、中国語キーワード抽出方法の選択肢が増え、キーワード解析とその応用の範囲が広がってきた。今回次の①～⑤の検討を行った。①複数の中国語キーワード抽出方法の比較検討、②テキストマイニング手法による重要キーワード、同義語抽出、③各種中国語特許データベース検索との相互補完的な活用、④中国語の概念（類似）検索の解析とその応用、⑤キーワード解析の応用として1次元（直線上）での公報の類似率ソート、2次元（平面上）でのクラスタリングによる可視化。

中国語キーワード抽出は化学分野：ヒアルロン酸、機械分野：風力発電をテーマにして確認した。中国特許公報より抽出した重要な中国語キーワードを用いることで検索集合の適合率向上に有用である。さらにダウンロードした検索集合に対して文書の類似率ソートを用いた調査の効率化手法を提案する。

【キーワード】

中国語キーワード抽出，中国特許情報解析，中国特許調査，中国語形態素解析，パテントマップ，キーワード辞書，テキストマイニング，クラスタリング，潜在的意味解析，ネットワーク分析

1. はじめに

中国特許データベースには様々なタイプがあり特徴および留意点も多様である[1)]。データベースで使える言語も中国語、英語、日本語はもとより機械翻訳の進歩とともに広がっている。Google 翻訳に代表される各種翻訳ツールや Web 上の中日、中英辞書等を使用すれば検索用の中国語キーワード(以下 KW と表記)を用意することも比較的容易である。但しこれらの方法で用意した中国語 KW が本当に適切かは保証の限りでない。また特許の性質上機械翻訳では対応できない専門用語や新語もしばしば現れる。

中国特許公報より中国語 KW を適切に抽出、利用できると上記問題は大幅に改善される。INFOPRO2012 でインパテック社のパテントマップ EXZ で中国語 KW 抽出して特許調査に応用した事例を紹介した[2)]。最近、低コストの中国語 KW 抽出方法の選択肢が増えており応用範囲も広がっている。

2. 目的

複数の中国語 KW 抽出方法の比較検討を行い重要 KW、同義語を抽出する。重要 KW として発明の構成、効果等の特徴を表す発明概念特徴語の抽出を目指す[3)]。

全体を大別すると下記(1)(2)の関係の解析に分けられる。

(1)KW 相互間の関係の解析

下記の検索精度、再現率向上への応用を検討する。

・KW 利用のブーリアン検索
・概念検索(類似検索)

(2)文書(公報)相互間の関係の解析

・1次元(直線上)での公報の類似率ソートによる検討(降順ソートで対象に類似の公報から確認する)
・2次元(平面上)でのクラスタリングによる可視化[4)]

3. 方法

化学分野:ヒアルロン酸、機械分野:風力発電を対象に検討した。

中国特許データベースの下記機能を使用した。CNIPR:KW 検索、概念検索、類似検索。Questel 社 Orbit.com:KW 検索、類似検索(英語)、中文ダウンロード。発明通信社 HYPAT-i 及びアイ・ピー・ファイン専利 SEARCH-i2:KW 検索、中文ダウンロード(タイトル、要約、請求項、全文)。

中国語 KW 抽出は下記のように2ステップで行った。

第1ステップ(入力された文書を KW に分解して出力)表1の6種類の抽出方法と下記 N-gram 法を検討した。

・N-gram 法(単語単位ではなく文字単位で機械的に分解して出力する)

第2ステップ(重要 KW、特徴語、同義語、類義語等の抽出)

・抽出語リスト作成/頻度解析
・抽出語の隣接頻度解析
・階層的クラスター分析
・ネットワーク分析
・潜在的意味解析
・人手抽出

文書中の特定文字列をサーチして位置を返す Excel VBA の InStr 関数を使用した KW 抽出と KW 数のカウント、正規表現により特定文字列を抽出する機能も場合に応じて使用した。

4. 結果

中国語 KW 抽出は化学分野ではヒアルロン酸関係の主題調査を念頭に特徴語の抽出を目指した。機械分野では風力発電関係の先行技術調査をテーマに発明のポイントとなる発明概念特徴語を抽出して調査対象に類似の文献から確認する手法を提案する。

4-1. 中国語 KW 抽出方法

表1に中国語 KW 抽出方法を示す。日本語の形態素解析器としてフリーで入手可能なものとして MeCab(和布蕪)、ChaSen(茶筌)等が有名である。中国語の形態素解析(分詞)ツールとして和布蕪や茶筌に相当するのは品詞情報も出力する ICTCLAS 2013(別名 NLPIR 中国語分詞システム)[5)]である。

IKAnalyzerNet[6)]は C#言語によるユーザーアプリケーションソフトから呼び出す中国語分詞ライブラリである。添付のサンプルプロ

グラムを位置情報を使用して隣接 KW を抽出してネットワーク分析に使用できるよう改良した。IKAnalyzerNet の KW の位置情報は自分で KW のインデキシングを行いたい人にとっても有用である。ICTCLAS と言選Web（中文版）[7]はWeb上でデモ機能を利用できる。

抽出方法	説明	特徴
ICTCLAS	中国語形態素解析	品詞情報出力
IKAnalyzerNet	C#から呼び出す中国語分詞ライブラリ	KWの位置情報出力
言選Web（中文版）	専門用語(KW)自動抽出サービス	専門用語抽出、多言語対応
パテントマップEXZ	パテントマップソフトの組み込み機能	日、英、中国語KW切り出し可
Orbit.com	中国語KW区切りコードを利用した抽出	中国語分詞ソフトにより分離?
Microsoft Word 中国語版	Wordの組み込み機能	VBAマクロより利用可能

表1. 中国語 KW 抽出方法

表2に CN1753913A の請求項1の特徴的な抽出結果を示す。太字（赤）で示した化学物質名アセチル化ヒアルロン酸の中国語に注目すると中国語 KW 抽出方法の特徴と課題が分かり易い。

CN1753913A の請求項1	1.一・ 眼用药物组合物，含有**乙酰化透明质酸**和可药用载体。
ICTCLAS	1./m 一/m・ /q **眼**/n **用**/p **药物**/n **组合**/vi **物**/ng，/wd **含有**/v **乙**/m **酰**/x **化**/k **透明**/a **质**/ng **酸**/a **和**/cc **可**/v **药用**/b **载体**/n 。/wj
IKAnalyzer Net 一部抜粋	12)14,17 = **乙酰化** 13)14,16 = **乙酰** 14)15,17 = **酰化** 15)17,20 = **透明质** 16)17,19 = **透明** 17)20,21 = **酸**
パテントマップ EXZ	**乙酰化透明质酸** **眼用**药物**组合**物
Orbit.com 注1）	1. 一・ □**眼用**□药物□**组合**□物，含有□**乙酰**□**化**□**透明质酸**□和□可□药用□载体。

注1）不可視の区切り記号&H200bを□に置換して可視化

表２． 中国語 KW 抽出結果

特許情報を解析する上でのポイントとして専門用語と新語に注目した場合、専門用語を抽出できるのは言選 Web、パテントマップEXZである。新語はどの抽出方法でもそのままでは十分な対応はできない。別の方法で新語を抽出して辞書登録を行うことでパテントマップEXZは対応可能である。係り受け解析等のさらに進んだ処理を行う場合にはICTCLAS の品詞情報は有用と思われるが本報では使用しない。Microsoft Word は対応する言語の単語の一覧情報を文書内部に保持している[8]。VBA マクロより利用できる。化学分野の専門用語は抽出できなかった。現時点で Orbit の中国語、日本語のテキストデータには KW の境目に特殊な区切り記号（16 進表記&H200b）が含まれている。この区切り記号を用いて KW を分離抽出した。Orbit の要約、請求項等のテキストデータを自分で処理する場合には区切り記号の存在を念頭において処理しないと予期せぬ結果を招く。Orbit の KW 検索も同様である。

表 3 は有機活性成分を含有する医薬品製剤の下位の IPC=A61K31/728(ヒアルロン酸)の検索結果 200 件（請求項）より抽出した KW の文字数分布である。正規表現を利用して英数記号を含む KW はノイズと見なして除去している。IKAnalyzer は部分的に N-gram による抽出をサポートしており文字数3以下の KW が多くなっている。

KW 文字数	EXZ KW	Orbit KW	IKAnalyzer KW	IKAnalyzer 頻度
1	74	1046	631	21473
2	923	2645	6031	73158
3	1074	735	1158	10276
4	1406	153	410	4934
5	856	26	90	608
6	687	11	22	132
7	410	3	6	27
8	244		0	0
9	154		1	5
10	96		1	1
11	51		2	5
途中略				
18	3			
計	6059	4619	8352	110619

表 3. 抽出 KW 文字数分布

各中国語キーワード抽出方法の特徴を考慮して IKAnalyzerNet、パテントマップEXZ、Orbit の区切り記号を用いた KW 抽出方法と N-gram 法による N が 1～6 の N-gram を次の第2ステップで比較検討した。

4-2. 重要 KW の抽出

特許DBを使用した同義語、類義語の抽出方法として WIPO PATENTSCOPE の多言語

検索 CLIR（Cross-Lingual Information Retrieval)[9]、CNIPR の概念検索結果の下部に表示される同義語、相関（関連）概念語がある。

テキストマイニングによる重要 KW、同義語、類義語の抽出方法として KH Coder[10]の分析手順を参考に基本検討として最初に抽出語リスト作成と頻度解析を行った[11),12)]。

ヒアルロン酸の主題調査を念頭に同義語、類義語の抽出を試みた。

No.	日本語	中国語	CNIPR	Orbit CN指定	HYPAT-i
1	ヒアルロン酸	质酸	4100	4327	4114
2	ヒアルロン酸	玻尿酸	59	151	55
3	ヒアルロン酸	透明質酸	0	43	0
4	ヒアルロン酸	透明质酸	3688	3667	3640
5	ヒアルロン酸	玻璃酸	228	257	230
6	ウロン酸	糖醛酸	823	1717	1723
7	ヒアルロン酸Na	质酸钠	650	646	632
8	ヒアルロン酸Na	玻璃酸钠	183	186	182
9	ヒアルロン酸Na	透明质酸钠	617	598	607

対象：CN公開、TI+AB+CLM　　検索：2013.03.06

表 4. ヒアルロン酸の同義語、類義語

ヒアルロン酸の同義語、類義語の抽出結果の一部を示す。No.1～5 が同義語、No.6～9 は類義語の例である。ウロン酸はヒアルロン酸に化学的に近い構造である。ヒアルロン酸の同義語の重要性を評価するため表 4 のDBでヒット件数を求めた。ヒット件数からは No.1,4 が重要である。Orbit(Fampat)はファミリー型 DB であり台湾特許が存在する場合ファミリーとして繁体字の中文データを保持している。No.4 の繁体字である No.3 は台湾特許の影響を受けている。

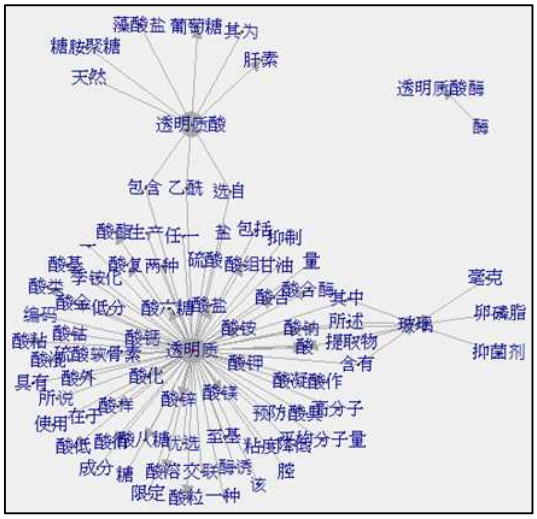

図1. ネットワーク分析結果（上位 100 組）

中国語 Windows 上の R 言語で行った中国語のヒアルロン酸関連隣接語上位 100 組のネットワーク分析結果を図1に示す。抽出母集団は表 3 と同じである。

4-3. 公報の類似率ソートの検討1

文書（公報）相互間の関係の解析における KW 解析の応用として1次元（直線上）での公報の類似率ソートの先行技術調査への適応を検討した。PCT 出願サーチレポートで多くの中国公報が引例として使われているCN1796780A（表5）をターゲット公報として引例の類似率と順位を確認した。検索集合は検索式を使用したブーリアン検索とCNIPR の類似検索機能[13)]により作成した。CN1796780A の類似検索を行い類似性検索 1758 件、新規性検索 225 件、侵害性検索 1533 件ヒットした。

類似性検索＝新規性検索+侵害性検索

常に上記関係が成立しているようであり新規性と侵害性の集合の重複はない。単純に類似検索対象の CN1796780A の出願日との関係（前後）で振り分けているようである。CNIPR の類似性検索では対象 CN1796780A に対する「相関度」が表示される。この相関度を取得してパテントマップ EXZ の「全キーワード類似率」との相関関係を検討した（423 件）。CNIPR と EXZ の類似率との相関関係は直線近似の傾き 1.17 であり決定係数 R^2=0.0424 で相関は低い。CNIPR では相関度 98%を示すものもあるが 98%という数値に見合った類似性が認められない場合が多いように感じる。

表 5 のサーチ引例の中でCNIPRの新規性検索では相関度：26.4% 順位 115 で Y 文献の1つ CN1257160 が抽出された。Orbit の類似検索（英語）では A 文献 CN1454292 が1件抽出された。CNIPR の新規性検索は化粧品分野でも数例行ったが良い結果は得られなかった。CNIPR と EXZ の類似率の比較では EXZ の類似率の方が実態に近いように思われる。

	Search引例
カテゴリーX:	CN1221855, CN1651759 ,CN1619143
カテゴリーY:	CN1261128, CN1405448, CN2479242, **CN1257160**
カテゴリーA:	CN1454292

表 5. CN1796780A の PCT サーチ引例

4-4. 公報の類似率ソートの検討 2

4-3.の前セクションの検討では文書全体あるいは請求項同士の類似率でソートして検討した。CNIPR の類似検索結果に比べてパテントマップ EXZ の類似率ソートの結果が少し良かったが直ぐに実務に使える訳ではない。そこでクエリ文書(EXZ のダミー文書)の検討を行い適合率向上に有効な下記手法を提案する。

提案手法

①ターゲット公報の予備検討

・発明のポイント抽出

・重要 KW 抽出(人手)

②DB検索

・ブーリアン検索

・ダウンロード

③パテントマップ EXZ へ取り込み

・重要 KW でダミー公報設定

・類似率でソート

④確認(スクリーニング)

提案手法に沿って本願:CN1796780A を例に説明する。

表 6 に発明のポイントを抽出した。

主請求項	高度差、または温度差により発生する圧力差に起因して発生する自然対流を利用
	風車により発電
	入口と出口を有する密閉配管内に風車を用いた発電機を複数設置

表 6. 発明のポイント抽出

表 7 に重要 KW を人手で抽出した。

	日本語	中国語
A	圧力差,高度差,温度差	压差, 高差, 温差, 高度差, 温度差, 压力差, 大气梯度, 气压差
B	対流,空気流,流動	对流, 气流, 流动, 空气下降流
C	風力,発電,風車,タービン	风力, 发电, 风车, 涡轮
D	配管,ダクト	管,风道,气道
E	入口	入口,进口,进风口,进气口
F	出口	出口,排风口,排口,排气口
G	分岐,切換,切替	分枝,交换,切换,转换
H	建物,ビル	大厦,建筑,大楼,号楼
I	廃熱,排熱	余热,废热
J	竜巻	龙卷风, 旋风
K	ウインドシア	风切变

表 7. 重要 KW 抽出 (人手)

重要 KW を使用して検索した。なるべく多くの該当引例を検索集合に入れるためにIPC(F03G7/04,F03D9/00, F03G6/00)や KW を追加している。対象公報では使われていない類義語、同義語も追加した。

検索結果の特許 93 件、実案 26 件をダウンロードしてパテントマップ EXZ に全文を取り込み、重要 KW 表 7 でダミー公報(クエリ文書)を設定して類似率で降順ソートした。

提案手法				全 文			
順位	公開番号	類似率	文献	順位	公開番号	類似率	文献
0	ダミー	-		0	1796780	-	本願
1	1796780	44%	本願	3	1619143	16%	X
2	1257160	42%	Y	6	1651759	15%	X
6	1651759	32%	X	7	1257160	15%	Y
8	1619143	29%	X	8	1261128	14%	Y
9	1261128	29%	Y	38	1221855	10%	X
14	1454292	27%	A	44	1454292	10%	A
16	2479242	27%	Y	69	2479242	8%	Y
27	1221855	22%	X				

全キーワード類似率(部分一致)

表 8. EXZ による類似率ソート結果

ダミー公報を設定した場合と本願を類似元に設定した場合のソート結果を表 8 に示す。文献の順位に注目するとダミー公報は検索式でヒットした7件の引例がベスト 27 位までに入っている。請求範囲の内容に類似の特許が確実に上位に来ていると言える。本願を類似元に設定した場合ベスト 69 位まで拡散している。

4-5. クラスタリングによる可視化

2次元(平面上)でのクラスタリングによる可視化[4]を検討した。

5. 考察

CNIPR の類似性、新規性、侵害性検索は指定した公報に対する相関度を公報毎に算出するのに対して提案手法ではターゲット公報から発明概念の特徴を現す重要 KW をクエリ文書としたダミー公報を基に類似率を算出する。クエリ文書の KW 選定の良し悪しが適合率に影響する。逆に言えば KW を適切に選択することで適合率を上げることができる。原理的に自分で KW を入力するタイプの概念検索には言語に拠らず応用できる。

意図的にタイプの異なるデータベース、異なる検索方法(特許分類、KW、ブーリアン検索、概念検索等)、検索言語を変えて複

数のベターな検索集合を本提案手法でマージして類似率ソートを行いターゲットに近い公報から確認することで再現率、調査効率向上に有効である。

6. 結論

表1の方法により抽出した中国語KWを用いた特許情報解析に関して、①中国語 KW 抽出、②重要 KW 抽出、③公報の類似率によるソート、④中国語の概念検索、類似検索、⑤クエリ文書(ダミー文書)の検討、中国特許調査への応用と一連の流れを検討した。

隣接語のネットワーク分析は重要 KW の抽出に応用できる。

抽出した重要 KW の応用としてクエリ文書へ適用して調査の適合率を向上させる手法を提案した。重要な KW を選択することで中国特許調査の調査精度、特に適合率の向上に有用である。ターゲット公報を解析した発明の特徴を現す重要 KW 抽出が適合率向上のポイントである。類似率ソートにより対象に類似の公報から確認できる。文書の類似度は2次元平面上での文書の相互関係の分析にも適用できる。

7. おわりに

本稿では各種中国語 KW 抽出方法、重要 KW を選択して特許情報解析、中国特許調査への応用を検討した。重要 KW の人手抽出をモデルにテキストマイニング手法による重要(特徴)KW 抽出についてさらに検討して洗練させたい。

「謝辞」

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、報告者として名前を挙げさせていただいた他に、他テーマリーダーの皆様には様々な協力をしていただきました。ここに改めて感謝申し上げます。

8. 参考文献

[1] 知的財産情報検索委員会第2小委員会.“中国特許調査に関する研究”.知財管理 Vol. 62 No.1, (2012),67-84

[2] 安藤 俊幸ら.“中国語キーワードを用いた特許情報解析”第 9 回情報プロフェッショナルシンポジウム

[3] 間瀬 久雄. “特許概念検索における特徴語抽出に関する評価と考察”. Japio YEAR BOOK. 2011 p166-171

[4] 安藤 俊幸.“テキストマイニングと統計解析言語 R による特許情報の可視化”. 情報管理. Vol. 52, No. 1, (2009), 20-31

[5] ICTCLAS. http://ictclas.nlpir.org/ accessed 2013/07

[6] IKAnalyzerNet. http://www.piaoyi.org/c-sharp/IKAnalyzerNet.html accessed 2013/07

[7] 言選Web(中文版) http://gensen.dl.itc.u-tokyo.ac.jp/gensenweb_cn.html accessed 2013/07

[8]”Microsoft Word 文書で 1 回しか使用されていない単語の一覧を取得する方法”http://gallery.technet.microsoft.com/office/071ac83b-c21e-435d-8bca-ee21f908b280 accessed 2013/07

[9] WIPO PATENTSCOPE 多言語検索 http://patentscope.wipo.int/search/clir/clir.jsp?interfaceLanguage=ja accessed 2013/07

[10] 樋口耕一. KH Coder. http://khc.sourceforge.net/ accessed 2013/07

[11] 知的財産情報検索委員会第２小委員会.“キーワードの選定にテキストマイニングを活用した特許検索手法の提案” . 知財管理 Vol.62 No.11, (2012),1583-1597

[12] 石田基広ら.“コーパスとテキストマイニング”. コーパスとテキストマイニング. 共立出版, 2012, p. 1-14.

[13] 石田政司ら.“中国特許データベース新 CNIPR の徹底活用”第８回情報プロフェッショナルシンポジウム