

# KHCoder による中国特許の分析可能性評価

## 中国特許分析ツールの精度評価及び分析事例

○袁方<sup>1)</sup>、小林義典<sup>1)</sup>、安藤 俊幸<sup>2)</sup>

松下電器研究開発(蘇州)有限公司<sup>1)</sup>、花王株式会社<sup>2)</sup>

〒215123 中国江蘇省蘇州市工業園区鐘南街300号

Tel: +86-512-62584710

E-mail: kobayashi.yoshinori@jp.panasonic.com

## Evaluation of Chinese patent analysis using KHCoder

## Accuracy evaluation of tools for Chinese patent analysis and analysis cases

YUAN Fang<sup>1)</sup>、KOBAYASHI Yoshinori<sup>1)</sup>、ANDO Toshiyuki<sup>2)</sup>

Panasonic R&D Center Suzhou Co., Ltd. <sup>1)</sup>、Kao Co., Ltd<sup>2)</sup>

NO.300 Zhongnan Street, Suzhou Industrial Park, Suzhou 215123, P.R.China

Phone: +86-512-62584710

E-mail: kobayashi.yoshinori@jp.panasonic.com

### 【発表概要】

昨今、中国での特許出願件数は世界第一位に成長するとともに、在華企業による出願も顕著になり、かつ国内の市場規模も依然巨大であることから、中国特許に対する注目度は年々高くなってきている。

日本企業のみならず中国マーケットで事業を行う企業にとっては如何に効率的に膨大な特許情報を読み解くかが重要課題のひとつであり、そのソリューションのため「商用」も含め様々なツールの開発が加速されている。

本報では、それらの分析ツールの中から、中文テキストマイニング機能を有し、ランドスケープ図などの可視化機能を有するとともに、日文・英文での特許分析での実績があり、かつ中国語分析機能が追加された「KHCoder」にフォーカスし、その実務視点での利用可能性評価について論じる。

対比評価においては、既に全件内容を確認済みの中国特許を母集合として用い、精度、活用性等の観点で評価を行った。

また KHCoder 等を用いて特定対象分野の特許に対して分析を行い、実務面でどのような分析ができるのかについて検討した結果について紹介する。特に 1 種類の分析ツールだけではなく、その組み合わせにより、より効果的な分析が出せないか、実務上有益な情報が引き出せないか、について検討した。

### 【キーワード】

KHCoder、テキストマイニング、中国語、特許分析

## 1. はじめに

日本において中国特許を分析する際、一般的には中国特許に対する英文翻訳を用い、英文テキストマイニングを実施し、単語の抽出、分析、及び可視化等を実施している。

しかしながら、分析結果は翻訳制度に依存する可能性が高い。

また分かち書きによりキーワードを抽出した場合には、英文では熟語で構成される用語を事前に十分に登録しないと単語が分割抽出されてしまい、中文の本来の意味と異なる分析結果が出てしまう可能性もある。

従って中文特許の分析ではやはり中国語で分析することが好ましいと考える。

中文特許の分析ツールは徐々に開発されており、筆者が既に紹介した RostNat や PatAnalyzer 等や、その他に商用分析ツールもいくつか公開されている。

本報では和文、英文のテキストマイニングで広く利用されており、かつ中文分析機能が追加された KHCoder<sup>1)</sup> が中国特許テキストマイニング分析に如何に利用できるかを評価することを主題としている。

## 2. KHCoder の評価

### 2.1 分かち書き精度

中文に対するテキストマイニングツールの評価において、「文字化けの有無」と「分かち書きの正確さ」が重要であると考える。

そこで先ず、同じデータ(空調分野 A 社、257 件の中国特許出願の要約部分)に対して筆者が過去に評価し、充分利用に耐えると判断した中文対応テキストマイニングツールである RostNat<sup>2)</sup> と PatAnalyzer<sup>3)</sup> との分かち書きの比較を

行った(図1)。

PatAnalyzer		KHCoder		RostNat	
閥	261	压缩机	245	压缩机	235
压缩机	247	閥	185	节流	167
节流	182	节流	157	面板	140
面板	166	面板	155	翅片	111
冷凝	166	管	140	四通	110
板	149	蒸发器	121	蒸发器	109
管	138	冷媒	107	冷媒	107
蒸发器	121	板	106	冷凝器	81
四通	110	翅片	105	电机	79
冷媒	108	电机	97	制冷剂	78
翅片	107	冷凝器	82	温度	75
电机	97	管路	77	管路	66

図 1. 分かち書き精度比較

結果、抽出されたほとんどの形態素が三者でよく一致しており、KHCoder の分かち書きは他の 2 つのツールとほぼ同等な精度を有すると判断した。

### 2.2 可視化能力の評価

分析結果の可視化において、KHCoder、RostNat が共通して有する「共起ネットワーク分析」の機能に着目して評価を行った。なお PatAnalyzer のデータの可視化には Ucinet<sup>4)</sup> を用い、対象データは 2.1 の分かち書き精度評価時と同様である(図 2、図 3)。

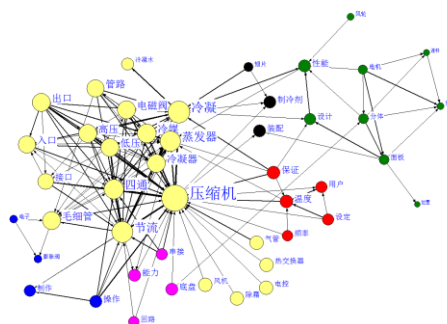


図 2. RostNat のネットワーク分析

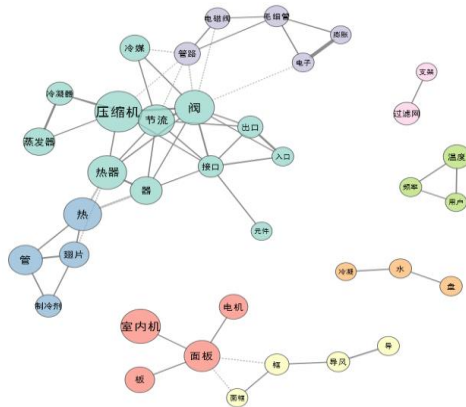


図 3. KHCoder のネットワーク分析

こちらも同様に、共起ネットワーク分析において抽出された主な「用語のネットワークのかたまり」はよく類似しており、各ツール相互に同等の評価ができるかと判断した。

### 2.3 KHCoder の長所と短所

2.1、2.2 の結果等に基づき「操作性」「可視化」「柔軟性」「処理速度」の観点で KHCoder と Rostnat、PatAnalyzer の比較を行った(図 4)。

「操作性」において、RostNat はワンクリックで分かち書き、フィルタリング(PRE 処理は必要)、キーワードの統計及び可視化を実施できる(但し、計算機の OS 言語環境は中国語必須)。

一方、KHCoder でもほぼ同様のプロセスを簡単に実現できるため操作性はよいといえる。

「可視化」において表示の種類では、RostNat、Ucinet は共起ネット分析しかできないが KHCoder では更に「多次元尺度分析」「対応分析」もでき、利用可能性が高い。

「(可視化の) 柔軟性」では、KHCoder は R 言語との組み合わせで中間データの生成をカスタマイズすることも可能ではあるが、同義語の名寄せの表示データ加工等において柔軟性が低い。

「処理速度」に関しては、数万から

数十万件のビッグデータも対応し、且つ万単位の分かち書き処理も数分で処理できる RostNat と比較すると、KHCoder と PatAnalyzer の処理速度は顕著に遅く、高速処理可能な計算機での処理が好ましい。

	操作性	可視化	柔軟性	処理速度
PatAnalyzer	△	△	◎	△
KHCoder	◎	◎	×	△
RostNat	◎	△	○	◎

図 4. 各ツールの比較

KHCoder はデータ取り扱いの柔軟性(中間データの編集等)、処理速度の面では課題はあるが、分析ツールとしての利用価値は総じて高いと考える。

## 3. 分析可能性検討

### 3.1 KHCoder による分析

2章と同じデータを用い、KHCoder で分析を実施した。なお、このデータは別途手作業で全件内容確認し、発明内容を把握している。

対応分析(図 5)では、出願は 2008 年から 2010 年の間、室内機、蒸発器、及び膨張弁に関するコンセプトが集中的に出現していることが分かる。一方、キーワードの類似度を計算した多次元尺度の結果(図 6)では、技術分類を読み取ることができた。

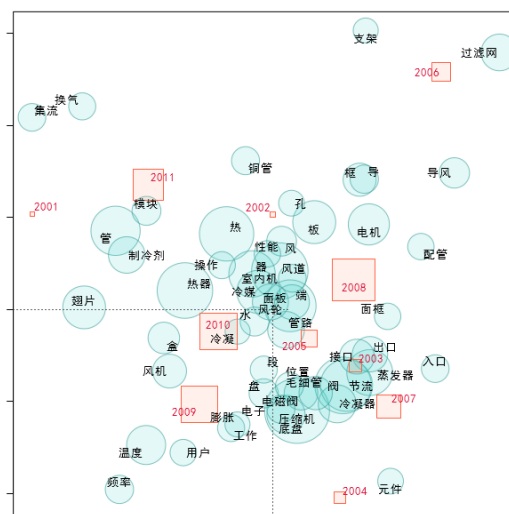


図 5. 出願年との対応分析

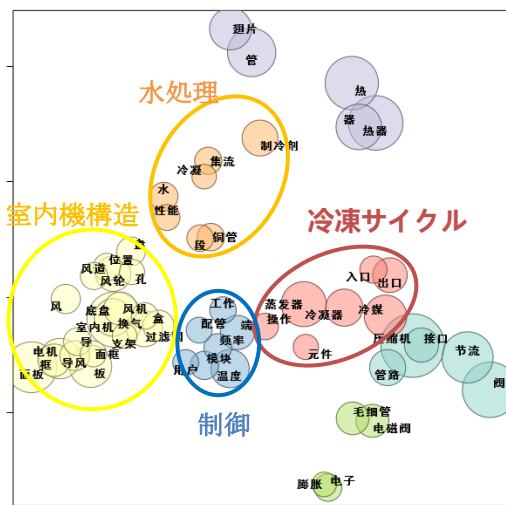


図 6. 出願年との多次元尺度分析

図 7 には同データに対する手動分析の結果を表示した。

手動分類の結果では、2008 年に風向羽根、熱交換器関連、送風関連と前面パネル、2010 年は弁関連と熱交換器の出願が最も多いことが分かった (図 8)。

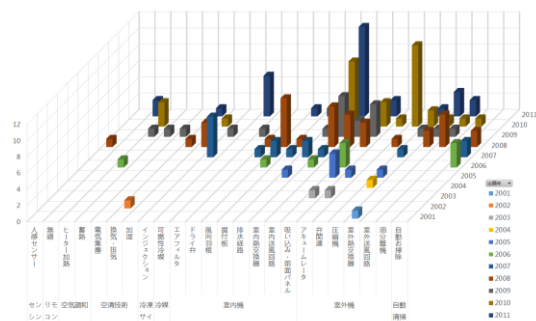


図 7. 手動分類結果の可視化

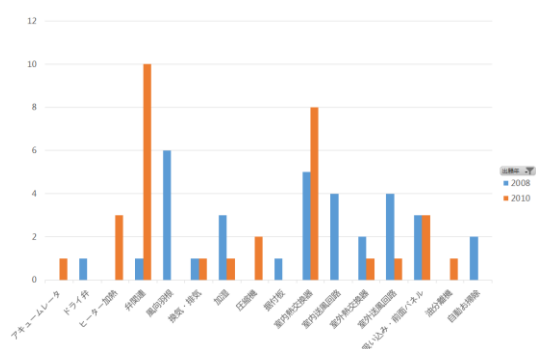


図 8. 手動分類による 08 と 10 年の出願分布

これらの結果と、KHCoder によるテキストマイニング結果を対比するために、キーワードの類似度を計算し、距離を算出して多次元尺度法で可視化を行った。

比較すると、手動分類の結果と非常に類似した特徴技術が確認できる (図 9、図 10)。

よって KHCoder により特徴技術項目を抽出できる可能性は高いと考える。

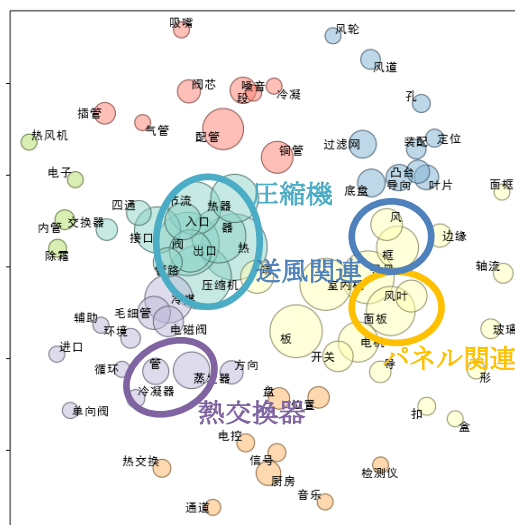


図 9. 2008 年出願のキーワード

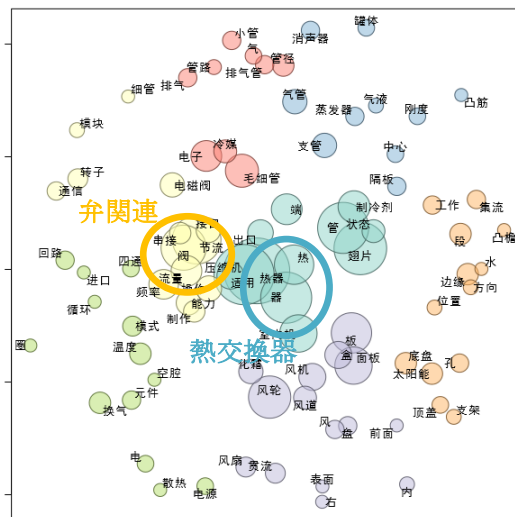


図 10. 2010 年出願のキーワード

### 3.2 ツールの組み合わせでの分析

KHCoder を利用した分析では、機能上、注目キーワードと関連する特許出願の情報（件数、具体案件等）の特定ができないため、筆者は他のツールと組み合わせでの分析がより手動分類に近い分析を出せるのではないかと考えた。

そこで incoShare 社の商用特許データベース incoPat<sup>®</sup>のテキストマイニング機能を利用し、KHCoder と組み合わせた分析を試みた（図 11）。

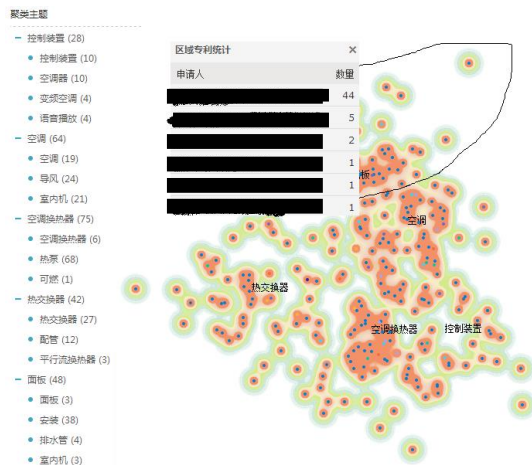


図 11. incoPat のランドスケープ分析

KHCoder で得られたキーワードを抽出し、incoPat で関連特許をリストアップした結果を図 12 に示す。

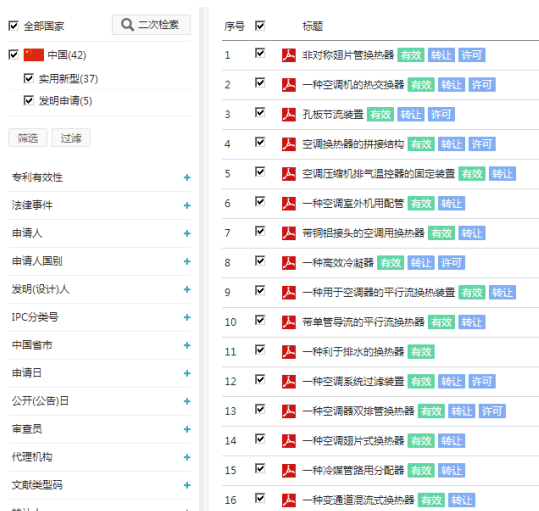


図 12. 「熱交換器」関連の特許抽出

また、KHCoder で分かち書き結果を可視化する際、最大出現数最小出現数の設定でノイズを除去したり、フォーカスする領域を絞り込んだりすることができる（図 13）。

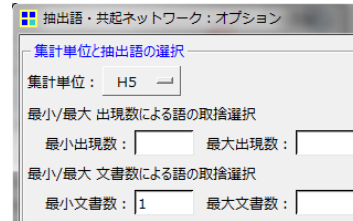


図 13. KHCoder の出現数設定

中心となるキーワードとの関連性を確認すると同時に、ノイズを除去したゾーン(図 14)は、萌芽キーワードを含んでいる可能性が高いと考える。



図 14. 抽出語の絞り込み

この考え方に基づき、PatAnalyzer でランキングを組みなおし、可視化を実施した（図 15）。

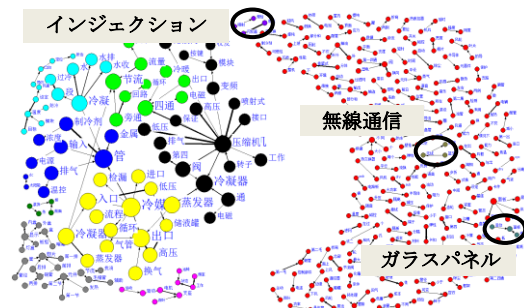


図 15. 萌芽キーワードの抽出

この結果、インジェクション、無線通信、ガラスパネルという3つの萌芽キーワードを抽出し、それぞれについて母集合中の出願傾向を調査した。

トータル件数で全体に占める割合は少ないが、出願は継続的且つ毎年増えていることが読み取れる（図 16）。



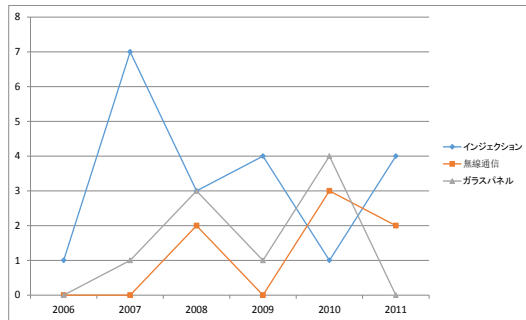


図 16. キーワード関連の出願傾向

実際、出願人は3つのキーワードと関連する商品を市場投入している。

このように、現在中文特許分析が可能なツールを組み合わせることで、特許出願から見た技術動向分析を行うことができる可能性は高いと考えている。

#### 4. おわりに

中国語特許分析の可能性検証した結果、KHCoderは分かち書きの精度、操作性、及び可視化の種類においては、現在使われているツールと比べて同等以上の機能があることを検証することができた。

また手動分類したデータとの比較検証結果からも、傾向などを把握するには十分利用可能であると認識した。

また、KHCoderと他のツールの組み合わせにより、より実用的な分析に資すると認識した。

今後は「萌芽技術の探索方法」について、事例を追加して検証し、その利用可能性を検討していきたい。

#### [謝辞]

本報告は「アジア特許情報研究会」でのワーキングの一環であり、会員の方々には貴重なご助言、ご協力を頂いた。ここに改めて感謝申し上げる。

#### 5. 参考文献

- [1] KHCoder  
<http://khc.sourceforge.net/>  
 accessed 2016.09.13
- [2] RostNat  
[http://www.tsjc.tsinghua.edu.cn/publish/jc/250/2014/20140616145239930572386/20140616145239930572386\\_.html](http://www.tsjc.tsinghua.edu.cn/publish/jc/250/2014/20140616145239930572386/20140616145239930572386_.html) accessed 2016.09.13
- [3] PatAnalyzer  
[http://www.geocities.jp/patentsearch2006/INFOPRO2014CN\\_KW\\_yokou.pdf](http://www.geocities.jp/patentsearch2006/INFOPRO2014CN_KW_yokou.pdf) accessed 2016.09.13
- [4] Ucinet  
<http://www.analytictech.com/>  
 accessed 2016.09.13
- [5] incoPat  
<http://www.incopat.com/>  
 accessed 2016.09.13