機械翻訳文を利用した中国特許文献の分析研究
中国特許文献原文と複数の機械和訳文を併用する分析法

○桐山　勉[1)], 藤城　享[2)], 栗原　健一[3)], 川島　順[1)], 長谷川正好, 田中宣郎[4)], 渡邊　彩[5)],
はやぶさ国際特許事務所[1)], 一般財団法人日本特許情報機構[2)], 田中貴金属グループ TANAKA ホールディングス株式会社[3)], 日科情報株式会社[4)], 銀龍専利東京事務所[5)],
〒101-0025　東京都千代田区神田佐久間町 1-14　第 2 東ビル 512 号室
Tel: 不掲載[1)] FAX: 不掲載[1)]
E-mail: 不掲載[1)]

Studies on Chinese Patent by plural Machine Translated Japanese documents
Combined Analysis on original Chinese Patent Information by plural Machine Translated Japanese documents

KIRIYAMA Tsutomu[1)], FUJISHIRO Akira[2)], KURIHARA Ken-ichi[3)], KAWASHIMA Jun[1)], HASEGAWA Masayoshi , TANAKA Nobuo[4)], WATANABE Aya[5)],
HAYABUSA INTERNATIONAL PATENT OFFICE[1)], Japan Patent Information Organization (Japio)[2)], TANAKA PRECIOUS METALS TANAKA HOLDINGS Co.,Ltd.[3)] , Nikka Joho[4)], Dragon Intellectual Property Law Firm (DRAGON)[5)],
1-14 KANDA-SAKUMACYO, CHIYODA-KU, TOKYO 101-0025,　JAPAN
Phone: not cited[1)] Fax: not cited[1)]
E-mail: not cited[1)]

【発表概要】
中国特許庁の第十二次五カ年計画により中国特許文献（特許・実用新案）の件数は世界一になったので、複数の機械和訳文を検索・評価・分析など各種ツールを併用して解析せざるを得ない。 ①中国特許文献の要約文・請求の範囲文だけでなくフルテキスト和訳文も取扱う。 ②この和訳文で形態素解析と類似度を調べる。
③複数の可視化ツール（市販の特許マップ・無料の Cytoscape など）を併用して解析を行う。④私的索引を併用して分析する。
このような分析研究を行ったので、INFOSTA-SIG-パテントドクメンテーション部会活動報告として発表する。

【キーワード】
中国特許調査, 中国語, 機械翻訳,和訳文,単語切り出し, 形態素解析, BX 的な連携活用,CSV データ・ハンドリング, 三カ国言語, シソーラス整備,全図参照,私的索引

1. はじめに

中国国家としての特許戦略に第十二次五ヶ年計画が公表され、2011 年の統計で中国特許情報の量は米国を抜き世界第一位である。しかも、中国語という言語問題がある。早急に、中国人の中国特許情報の調査専門家レベルに自分達を育成する必要がある。2012 年に当パテントドクメンテーション部会にて立てた3年計画の最終年として、日本で使える Tool を有効活用して中国特許マップ図を活用する方法と EXCEL 型の評価管理 Tool に注目して、複数の機械翻訳和訳文を参照しながら中国特許文献を分析することを研究した。

2. 目的

言語と制度の壁を乗り越えて、日本の特許情報調査専門家が「中国特許情報の中国人専門家レベル(プロサーチャー)」に到達することが目的である。中国の急激な環境変化に対応するために、スケジュール設定を 2012 年 1 月に3年間とした。2012 年を自己啓発期として、先ず三カ国語のシソーラス辞書を準備し、複数の DB を使い、SDI は中国語の CNIPR を利用する研究をした。2013 年を第2年目として社内教育者育成期と位置付けた。そして、日本にて有効活用できる Tool の駆使を想定し、東芝製の Eiplaza/DA による特長語の切り出しと、レイテック社の PAT-LIST-GLS による特許マップ作成を検討した。第3年目の 2014 年を社内教育の実践期と位置付けした。そして、中国言語が殆ど理解できない研究・技術者の視線からオリジナル中国語特許調査をするには、やはり機械翻訳全文和訳文と全図参照と私的索引が欠かせないと判断した。そこで、中国特許を全文でデジタル化して検索できる DB を探し、中国特許全文を機械翻訳している DB を探した。次に、中国特許を評価整理する方法として EXCEL 型のテンプレートをデフォルトとして準備されている Tool が使い易いと判断して全図参照も併せてできる Tool を探した。更に、EXCEL 型のテンプレートだけに左右されずに中国語特許のマップ軸が X 軸・Y 軸とも自由に選ぶことができる市販マップ Tool を補完的な Tool として使わざるを得なかった。また、俯瞰可視化 Tool としては無料の Cytoscape[1]を選び、それを使いこなすことに専念した。複数の機械翻訳和訳文をデフォルトで扱える評価整理 Tool としてアイ・ピー・ファイン社の THE 調査力_クラウド[2]を見つけることができた。依頼交渉して1か月間の試行研究を重ねて、評価整理の EXCEL 型の Tool の一つとして THE 調査力_クラウドなら研究者・技術者でも問題なく使えるという実感を得たので、更に試行研究期間の延長依頼交渉を行ってこの研究を継続した。

本研究の狙いは、複数の検索 DB を駆使し、中間プロセスにおいて EXCEL 型の評価整理 Tool を使いコメントを記入する際のスクリーニングの効率化と品質向上とスピードアップ化を図るものである。

その為に、市販の特許マップ Tool と俯瞰可視化 Tool を駆使する手法である。この中国特許の分析プロセスの概要を図1に示す。

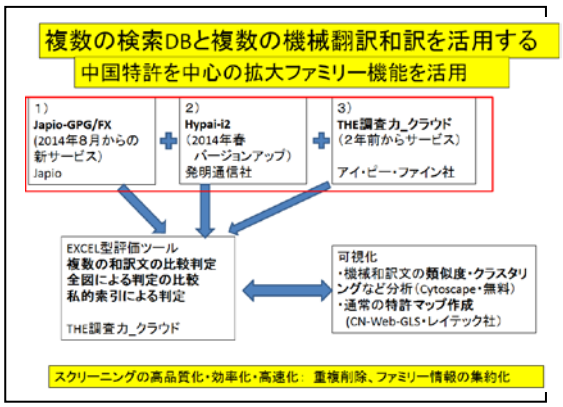

図 1　中国特許の分析プロセス

3. 方法

4つの研究テーマに関してそれぞれの担当者を決めて分担したため、各社の中国特許の検索システムは異なる。 それらから中国特許の書誌データを収集した。今年は、検索 DB システムと単語切り出し機能を有するシステム間で CSV 形式（Unicode、UTF8）にてデータをやり取りすることになり、IT 技術のリテラシーの義務的な教育がおのずと必要になった。少なくとも、データ・ハンドリング加工のアレルギーだけは無くさないと実践できないレベルに達した。

表 1　研究テーマと検索 DB システム.

No	研究テーマ	検索 DB	備考
1	電動車椅子	HAYPAT-i 他	予稿集
2	耐震免震	J-GPG/FX 他	当日発表
3	Ag 接点スイッチ品	Patent-SQ 他	当日発表
4	CO2 固定化	J-GPG/FX 他	当日発表

今年の研究では、データの流れが特に重要なプロセスである。4つのテーマを研究する中国特許検索システムが異なるために、ダウンロードの項目と形式は微妙にことなり、受け皿となる THE 調査力_クラウドとの BX 的なデータのやり取りが重要なプロセスとなった。

今年の研究では IT リテラシー教育を重要視した。その概念として三人のバーチャル・アシスタントを活用することを考えた。

第一のバーチャル・アシスタント はインターネットに接続された高性能PCとマルチ画面である。10 万円以上の 15.6 インチ画面のノート型PCと、19 インチ以上のディスプレーを接続して二画面で特許情報解析を行っている。

第二のバーチャル・アシスタント はインターネット接続された環境下で、特許検索システムと特許管理評価ツールと特許マップ・ツールの3つのITリテラシー道具を Windows のマルチ画面にて BX 的に連携活用する方法（筆者はテトラヘドラル法と呼ぶ）である。

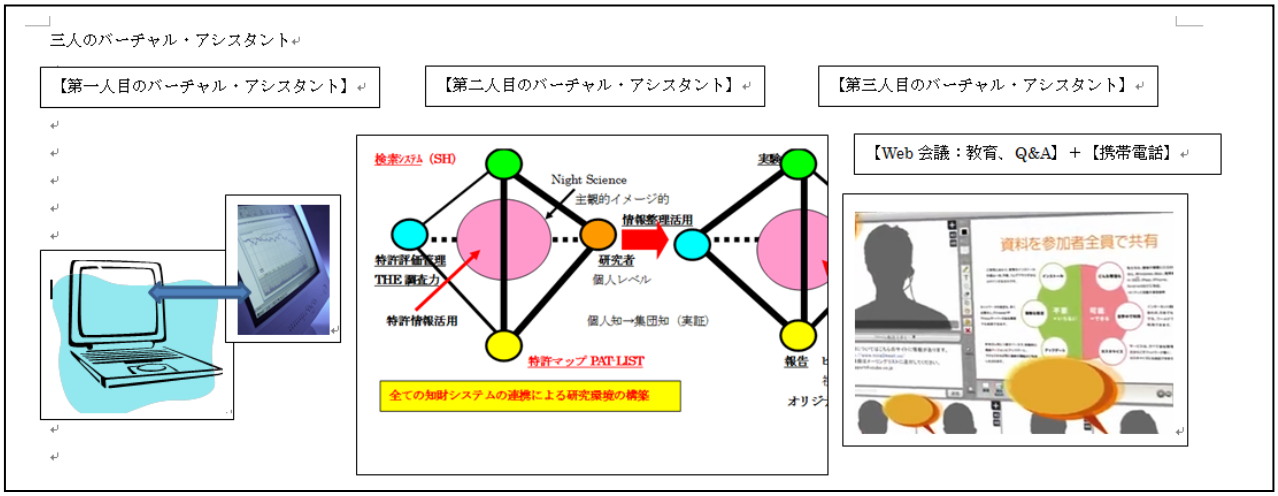

図 2　三人のバーチャル・アシスタント活用法

第三のバーチャル・アシスタントは、Web 会議と称してマルチ・ウインドウズと複数の PC と電話回線を同時に活用する方法である。一台の PC に負担をかけないように2台の PC を併用し、切替 SW にて拡張画面の表示を筆者は切り替えている。一つの PC で電子メール、PPT、WORD、検索システムを使い、もう一台の PC で EXCEL 型管理評価システム、特許マップシステム、Skype などの Web 会議に使っている。PC の発熱と負担を極力押さえるためである。2台の PC は無線 WiFi にて同時にインターネットに接続させている。

4. 結果

この予稿集では4つの研究テーマの内で、もっとも研究が進んでいる「電動車椅子」の結果を中心に記載する。当日の INFOPRO 発表の中では残りの3つの研究テーマの結果を含めて発表する。最初の三カ国語の技術用語のシソーラス整備の段階では、Japio-GPG/FX[3]の技術用語支援機能が役立った。日本語で技術用語を入力すると対応する英語技術用語を一覧で示してくれた。それらをインターネットの Google 翻訳を用いて、日本語→中国語（簡体語）および英語→中国語（簡体語）の検索を行い、それらから得られた中国語（簡体語）の技術用語を、自分の研究テーマの三カ国語のシソーラス整備一覧表にした。この最初の三カ国語の技術用語のシソーラス整備はどのような研究テーマを調査する場合にも最初にすべき中国特許調査のプロセスである。

中国特許の検索 DB として特許データ仕様が THE 調査力_クラウドに良く似ている発明通信社の Hypat-i[4]を選んで良かった。

複数の機械翻訳和訳文としては表1に掲載の4種類を選んだ。

表2　4種類の機械翻訳和訳文

No	機械翻訳和訳文の種類
1	Hypat-i 全文和訳文
2	Japio-GPG/FX 全文和訳文（力技）
3	アイ・ピー・ファイン全文和訳文
4	Google 翻訳全文和訳文（力技）

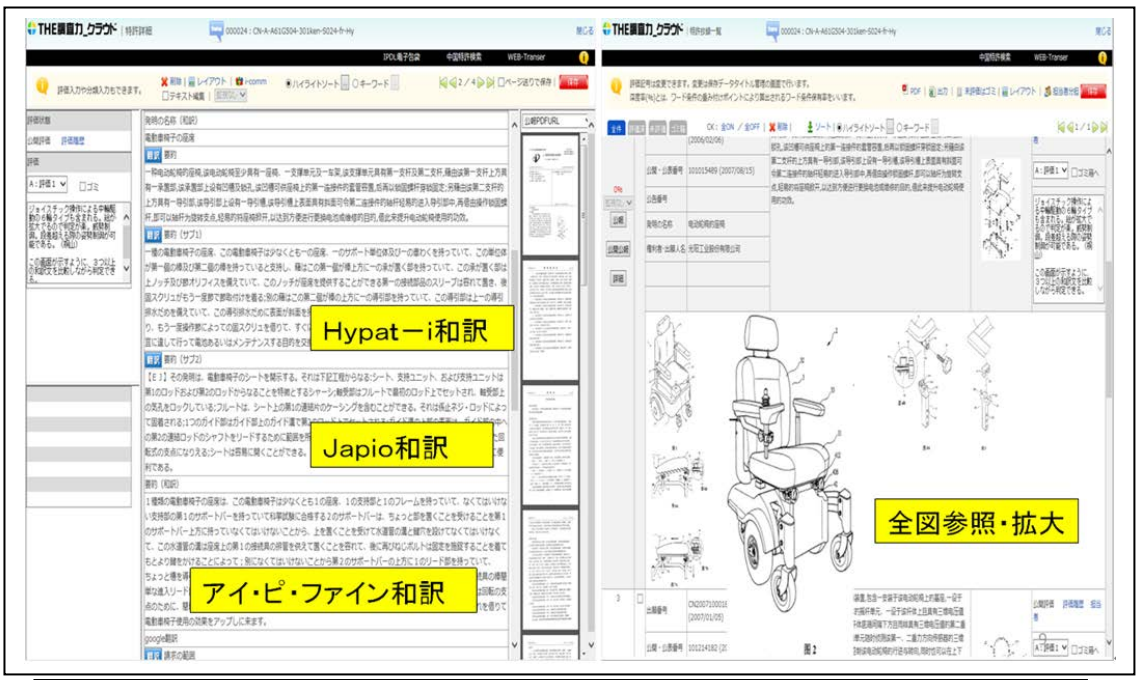

図３　マルチウインドー画面に表示された複数種の和訳文群と動的な全図群

機械翻訳和訳文のデータの取り込みに関しては、Hypat-i 和訳文とアイ・ピー・ファイン和訳文にて実施した。差異に疑問を感じたデータを100件以内選び、更に二種類の和訳文を追加した。その理由は、特許データ仕様が細かい部分で異なる Japio-GPG/FX の機械翻訳和訳文を取り込むのと、Google 翻訳和訳文のデータを準備する作業は力技で行うからだ。和訳文を4種類も比較すれば概略の問題点は把握できた。更に、権利情報の詳細が必要な場合にのみ専門翻訳者の人手翻訳を依頼することにした。

三カ国語の技術用語シソーラス表を最初に整備しても、中国語言語が殆ど理解できない研究者・技術者の立場に立つと、この全図参照と4種類の機械翻訳和訳文を参照しながら自分の評価判断コメントを記載入力することが必要という考えに到達した。電動車椅子の約1500件の特許の内、機械翻訳和訳文にて極めておかしいと気づいたのは約 50 件であった。

個人索引にて、電動車椅子を(a)クローラ型、(b)遊星輪型、(c)四輪以上駆動の太いタイヤ型、(d)レール型、(e)動物ロバ型、(f)動輪脚型、(g)汎用2輪駆動型、(h)その他の8種類に分け、その次に、出願人別にチェックしていくのが、今回の研究で得たスクリーニングの通常の最適方法であった。

SDI 機能は、THE 調査力_クラウドにもあり、事前登録を中国語言語「电动轮椅」と IPC 分類「A61G5/04+A61G5/06」の両方で行った。研究した期間では IPPH で検索され、THE 調査力_クラウドで漏れた電動車椅子の事例はなかった。

THE 調査力_クラウドは EXCEL 型の評価整理 Tool であるためテンプレートの横軸（X 軸）は時系列だけがデフォルトとして用意され、縦軸（Y 軸が）は出願人別、発明者別、IPC 分類別がデフォルトで用意され、オプションとして個人索引別などが設定できた。しかし、あくまで横軸は時系列しかないため、中国特許のマップ作成として別の市販の

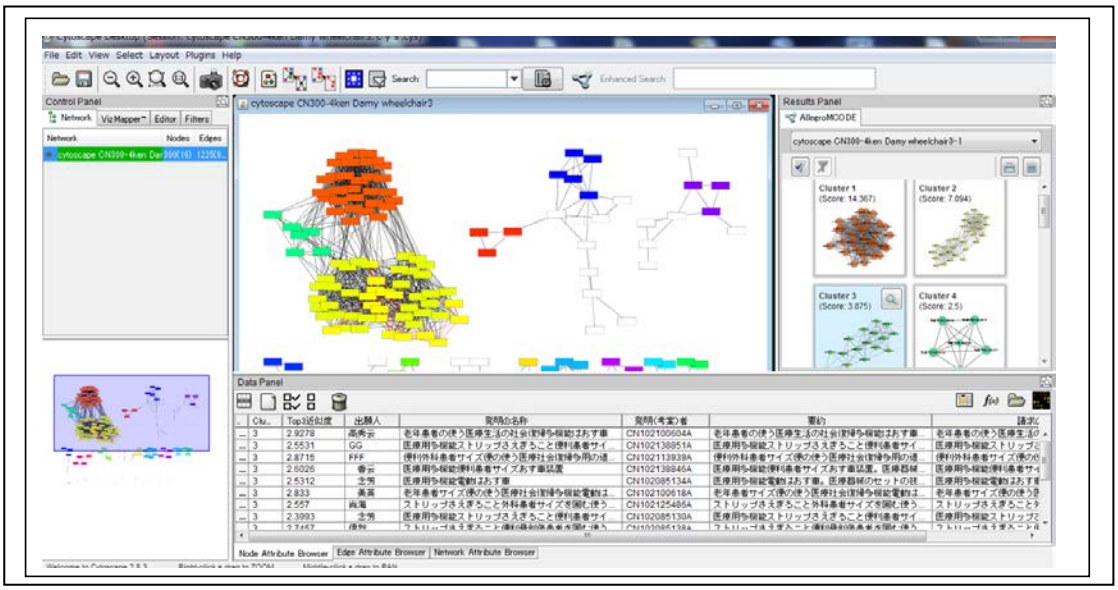

図４　中国特許文献の和訳文を利用した Cytoscape 図作成とクラスター分析図

特許マップ Tool にて補完せざるを得なかった。特に、横軸と縦軸に例えば機能・効果とそれを実現する具体的な手段を設定する際には、例えば、PAT-LIST-GLS [5]を活用せざるを得なかった。この研究では数か月間にわたり研究分野別の残り3人も、レイテック社のご好意により PAT-LIST-GLS が試行できた。

俯瞰可視化に関しては、無料の Cytoscape を駆使した。図4にその一例を示した。俯瞰可視化に関して、中国特許だけを扱うのであれば、この機械翻訳和訳文を取り扱い Cytoscpae 図が作成できた。日本特許だけ、または米国特許だけを取り扱うのであれば、この Cytoscape にて俯瞰可視化図を作成できる点まで研究できた。しかし、世界の複数国の特許群を複数言語で一度に扱うレベルまでは、残念ながら Cytoscape でできていない。

現時点では世界の電動車椅子を同時に解析する場合には、PatBase 検索→Bizlnt Smart Chart 取込→Search Technology 社の Vandage Point および Aduna Map にて分析を筆者は行っている。将来は、全て英語で扱えるようにしたいが、各国の特許データ仕様の細かな INID Code 仕様の違いによる理由から苦戦した。

5. 考察

機械翻訳和訳文の解析による Cytoscape 図と中国語言語のままでの解析による Cytoscape 図に違いが発生するのではないかという疑問に対して研究をしたので、その一部を紹介する。現実として、機械翻訳和訳文と中国語原文の類似度解析で微妙に差異が生じて、(a)クラスター分析が完全に一致する場合、(b)機械翻訳和訳文の分析でクラスターに入る場合、(c) 中国語言語の解析にてクラスターに入る場合が生じることが判っている。コンピュータに負担を著しくかけるので現時点では「免震耐震」の部分集合にてテスト研究を行っている。その一部を図5に掲載する。

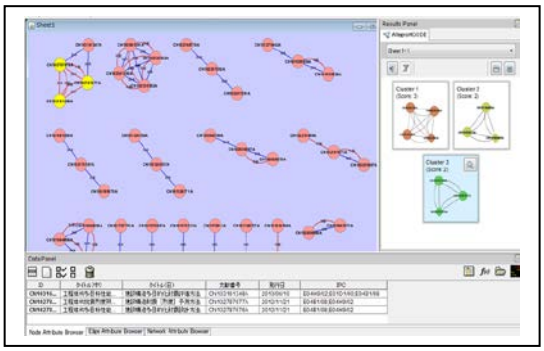

図5　機械翻訳和訳文と中国原文の Cytoscape 図の上での比較テスト

6. 結論

中国特許の全図面を参照する機能と複数の機械翻訳和訳文を並列表記できる機能と技術用語選択判断評価機能をデフォルトで有する THE 調査力_クラウドと、特許マップと俯瞰可視化の両 Tool を併用することにより、中国特許分析に対する「中国語言語が全く解らない」という恐怖感と中国言語アレルギーが殆どなくなった。本研究を達成するに際して、大垣 Cytoscape 勉強会の今津均氏にご指導とご支援を全面的に頂いた。また、THE 調査力_クラウドと PAT-LIST-GLS を三カ月にわたり試用させて頂いた。ここに謝意を心から表します。

7. 参考文献

(Web 参照日は全て 2014.8.30 です。)
[1]http://www.cytoscape.org/index.html
[2]http://www.ipfine.com/TIP_Cloud/
[3]https://gpgfx.japio.or.jp/
[4]http://www.hatsumei.co.jp/hypat_i/
[5]http://www.raytec.co.jp/products/patlist/gls.htm