中国特許解析・テキストマイニングによるKW分析：
適合率を重視した特許調査支援

○安藤俊幸[1)]，桐山　勉[2)]
花王株式会社[1)]，はやぶさ国際特許事務所[2)]
〒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

Chinese patent analysis - Keywords analysis by Text mining:
Patent research support of precision-oriented

ANDO Toshiyuki[1)], KIRIYAMA Tsutomu[2)]
Kao Corporation[1)], HAYABUSA International Patent Office[2)]
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】

特許調査においては調査目的により程度の差はあれ再現率（網羅性）重視で検索集合を作ることを前提としている。ただし再現率重視だとノイズは増加する。

本報告では再現率重視の検索集合からテキストマイニング手法を応用して適合率重視で抽出/ソートを行い、スクリーニングを支援する手法として下記検討を行った。

①キーワードのネットワーク分析による重要語（特徴語）抽出
②キーワードを用いて計算した公報の類似率によるネットワーク分析
③適合率を重視した特許調査への応用

キーワード解析は iPS 細胞をテーマにして日・中文で比較検討した。中国特許公報より抽出したキーワードをネットワーク分析して重要な中国語キーワードセットを選択する。重要な中国語キーワードセット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集合の類似率ソートを用いた手法を提案する。

【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，テキストマイニング，ネットワーク分析，媒介中心性，キーワード辞書

1. はじめに

中国特許検索には英語データベースを用いて英語で検索、中国語データベースを原語（中国語）で検索する等の方法がある。最近は機械翻訳の日本語が搭載されたデータベースも増えている。Google 翻訳に代表される各種翻訳ツールや Web 上の日中、英中辞書等を使用すれば検索用の中国語キーワード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キーワードが本当に適切かは保証の限りでない。また特許の性質上翻訳ツール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キーワードを適切に抽出、利用できると上記問題は大幅に改善される。INFOPRO2012,2013 で中国語キーワード抽出して特許調査に応用した事例を紹介した [1)]。

2. 目的

適合率を重視した特許調査支援方法として調査に重要なキーワード（特徴語）、同義語、関連語を簡単に抽出する方法を最初の目的とする。重要キーワードとして発明の構成、効果等の特徴を表す発明概念特徴語の抽出を目指す。キーワード抽出は中文だけでなく日文にも対応させて汎用性を向上させる。

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）キーワード相互間の関係の解析
下記の検索の適合率（精度）向上への応用を検討する。

・特徴語の効率的抽出
・キーワード利用のブーリアン検索

（2）文書（公報）相互間の関係の解析
・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）
・2次元（平面上）での関連性の高い公報のネットワークによる可視化

本稿では（1）をメインに報告する。

3. 方法

キーワード解析は iPS 細胞を対象にして日・中文で比較検討した。
中国特許検索/データは下記データベースを使用した。IPPH CNIPR、同日本版 CNIPR、Questel 社 Orbit.com。

（1）キーワード相互間の関係の解析

キーワード抽出は下記のように2ステップで行った。

第1ステップ（入力された文書をキーワードに分解して出力）
・IKAnalyzerNet：中国語分詞ライブラリ [2)]
・ICTCLAS：中国語形態素解析ツール [3)]
・saezuri lite：C#から利用する自然言語処理支援ライブラリ [4)]（日本語処理用）
・パテントマップ EXZ（日・中文 KW 抽出）

第2ステップ（重要キーワード、同義語等の抽出）
・抽出語リスト作成/頻度解析
・頻度解析を元に R のワードクラウド表示
・抽出語の隣接頻度解析
・隣接頻度解析より語のネットワーク分析
・ネットワーク分析より特徴語抽出
・人手抽出

ネットワーク分析による検討は R と Cytoscape[5)]を使用して行った。

特定文字列をサーチするキーワード抽出と正規表現により特定文字パターンを抽出する機能も必要に応じて使用した。これらの機能は言語に依存しない。

（2）文書（公報）相互間の関係の解析

公報の類似率ソートによる検討はアイ・ピー・ファイン社 THE 調査力の深度マイニングの専門用語によるソート機能（日本語のみ）、同社 THE 調査力_クラウドによるワードソート機能、インパテック社のパテントマップ EXZ について検討した。

関連性の高い公報のネットワークによる可視化は自作類似率計算プログラムと Cytoscape を使用して検討した。

4. 結果

中国語キーワード抽出は iPS 細胞の先行技術調査を想定して確認した。

4-1. 特徴語の効率的抽出

プログラミング言語 C#で自作したキーワード解析ツール PatAnalyzer の画面コピーを図 1 に示す。中国語キーワード抽出には IKAnalyzerNet を利用している。文献 6 に IKAnalyzerNet を用いた中国語の同義語抽出の詳しい説明がある。日本語処理は saezuri lite(自然言語処理支援ライブラリ)を介して利用する形態素解析(mecab 和布蕪)、係り受け解析(Cabocha 南瓜)の機能を追加した。

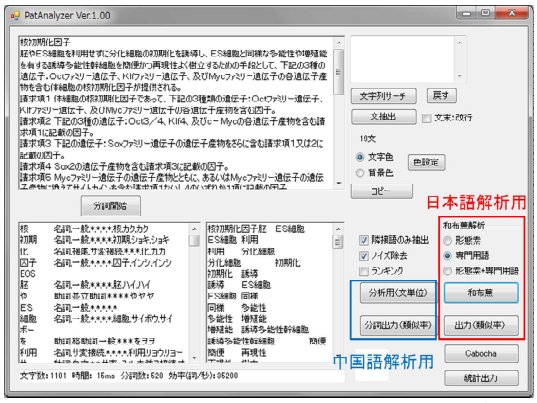

図1. 中国語/日本語キーワード解析ツール PatAnalyzer

中国語キーワード抽出の詳細に関しては INFOPRO2013 予稿[1]を参照されたい。現状では中国語の専門用語抽出は十分ではない。専門用語抽出に関しては別途考察する。

日本語公報が得られる場合には最初に日本語の専門用語を抽出して以下に述べるワードクラウドで調査に重要な特徴語を最初に把握してから中国語の専門用語を見つけると効率的である。

図 2 に WO2007069666 の請求項より RMeCab[7]+wordcloud[8]を用いて抽出した形態素の名詞とその頻度情報による形態素レベルのワードクラウドを示す。

図 2. 形態素レベルのワードクラウド

図 3. 専門用語レベルのワードクラウド

図 3 は saezuri lite を介して形態素解析ツール mecab の品詞と隣接頻度情報より求めた専門用語レベルのワードクラウドである。文字サイズは出現頻度に比例している。枠で囲んだ「誘導多能性幹細胞」、「Myc ファミリー遺伝子」、「核初期化因子」が調査に重要な特徴語である。対応する CN 公報があればその公報から該当する中国語キーワード探しに行くと効率的である。

中国語キーワード抽出には文書中の特定文字列をサーチして該当文字列があった場合に文単位で抽出する機能と正規表現により特定文字パターンを抽出する機能も実装して使用した。これらの機能は他言語にも応用可能である。

4-2. キーワード相互間の関係の解析

キーワード相互間の関係の解析ツールとして特許版事典検索システム Cyclone (サイクロン)[9]を検討した。Cyclone の収録は 1993～2007 年公開の JP 公報であるので ES 細胞をキーワードとして入力して検討した。

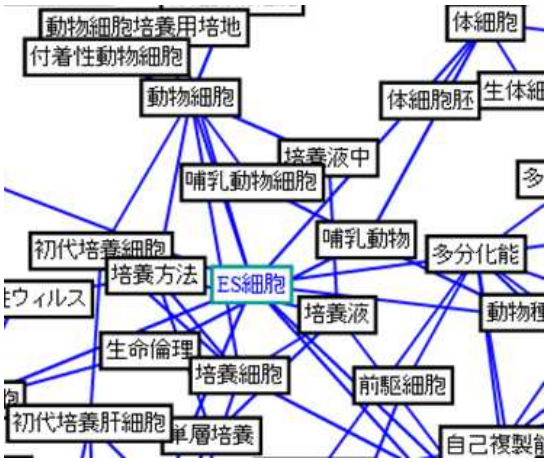

図 4. ES 細胞の Cyclone による表示

Cyclone のノイズは非常に少ない。またキーワード間の関係を動的に表示するのが特徴である。ただ Cyclone は原文の情報が不明であり詳細が確認できない等実務上で使用するには課題がある。Cyclone の表示のようにキーワード相互間の関係をネットワークとして表示/解析できるツールとして Cytoscape を使用して iPS 細胞のキーワード解析を行った。

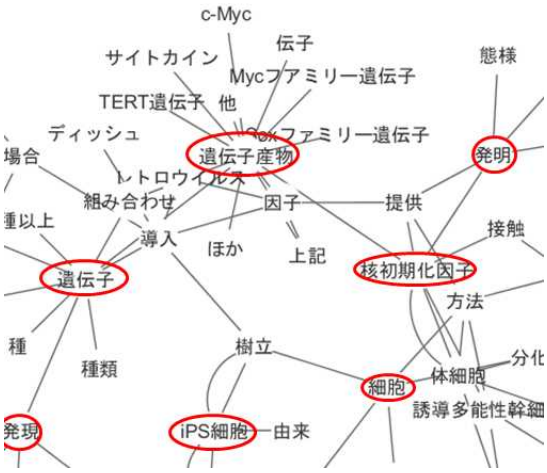

図 5. Cytoscape による語のネットワーク

図 5 は WO2007069666 明細書より抽出した日本語の専門用語上位 100 語の Cytoscape によるネットワーク表示の一部である。楕円で囲んだキーワードはネットワーク的に他のキーワードとの関係上影響が大きいものである。まだノイズ除去に課題がある。

図 6 は WO2007069666 請求項 1 の専門用語レベルの有向グラフのネットワークである。専門用語がきちんと抽出されている。

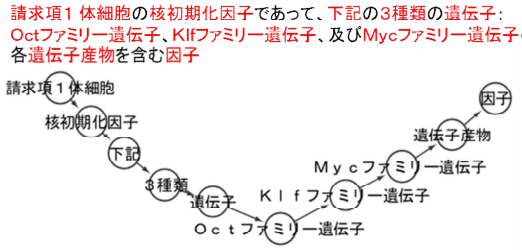

図 6. 請求項1の専門用語のネットワーク

図 7 は CN101356270 請求項 1 の形態素レベルのネットワークである。IKAnalyzerNet により分詞し各形態素の隣接頻度を求め Cytoscape でネットワーク表示した。

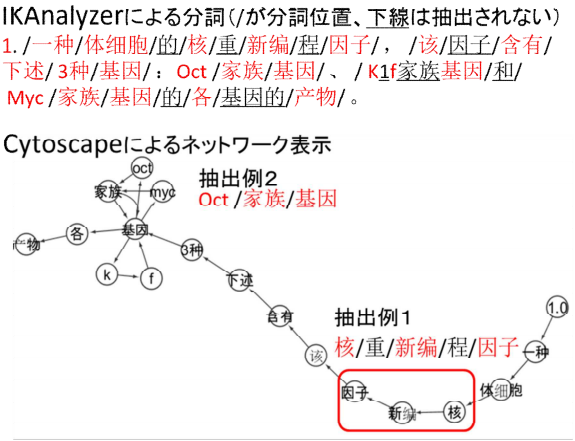

図 7. 請求項 1 形態素のネットワーク

図 6 、図 7 は同一内容の日文と中文である。中文の専門用語抽出にはまだ課題がある。更に適切な抽出に向けて検討を進める予定である。

4-3. ネットワーク分析による特徴語抽出

語のネットワーク表示の応用例として図8に示す最近公開されたWIPO Pearl のConcept Map Search[10] の語のネットワーク表示がある。図8はES細胞の中国語表示である。インタラクティブに表示されキーワードから公報にリンクされる等非常に興味深い。ただまだ収録語数が少なくiPS細胞では2語しか表示されなかった。今後の収録語数アップが望まれる。

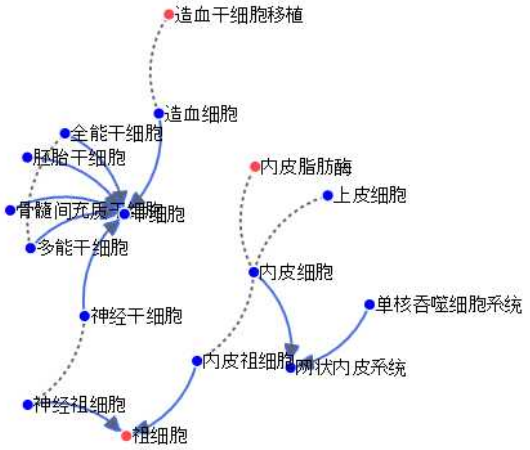

図8. Concept Map Search の例

表1にCNおよびJP公報があるiPS細胞の中国語、英語、日本語によるOrbit.com キーワード検索式を示す。近接演算、トランケーションを使用してこの段階では網羅性を高めている。この 268件をダウンロードしてネットワーク分析による特徴語抽出、公報の非類似度(距離)によるネットワーク表示検討に使用した。

表1. iPS細胞のOrbit.com 検索式

Orbit.com検索式　2014.09.27

1	(iPS 3W 細胞)/BI/CLMS AND (CN)/PN	76
2	(iPS 3W 细胞)/BI/CLMS AND (CN)/PN	138
3	(多能 3W 干细胞)/BI/CLMS AND (CN)/PN	430
4	(induced_pluripotent_stem_cell?)/BI/CLMS AND (CN)/PN	198
5	(iPS cell?)/BI/CLMS AND (CN)/PN	151
6	(iPSCs)/BI/CLMS AND (CN)/PN	25
7	(iPS 3W 細胞)/BI/CLMS AND (JP)/PN	212
8	(iPS 3W 细胞)/BI/CLMS AND (JP)/PN	51
9	(多能 3W 干细胞)/BI/CLMS AND (JP)/PN	219
10	(induced_pluripotent_stem_cell?)/BI/CLMS AND (JP)/PN	200
11	(iPS_cell?)/BI/CLMS AND (JP)/PN	218
12	(iPSCs)/BI/CLMS AND (JP)/PN	19
13	1 or 2 or 3 or 4 or 5 or 6	537
14	7 or 8 or 9 or 10 or 11 or 12	485
15	13 and 14	268

図9は表1のiPS細胞関連特許の請求項に「iPS」が含まれる177 文を最初に抽出してから専門用語を抽出して、キーワード「iPS 細胞」の周辺の専門用語をネットワーク表示したものである。ネットワークの中心性の指標として良く用いられる媒介中心性でカラーマッピングしている。媒介中心性(Betweenness Centrality)は、そのノードを通過しないと他のノードに到達できない度合、つまり、ある点がその他の2点を結ぶ最短経路である度合であり、値が大きいほど中心性が高い指標である[11]。媒介中心性は重要語(特徴語)を抽出する際の指標となる。

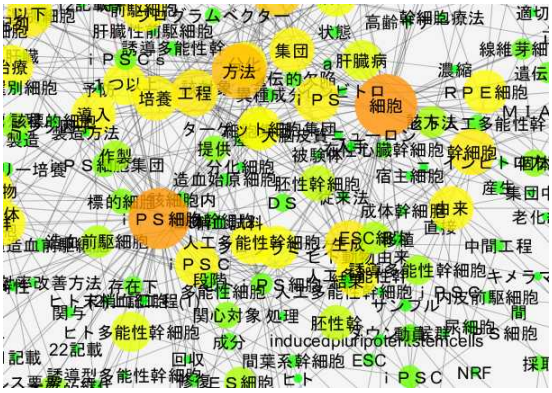

図9. iPS細胞を中心としたネットワーク図

4-4. 公報の非類似度ネットワーク表示

公報の非類似度(距離)を使用してネットワーク表示が可能である(図 10)。注目公報の近くにある公報から確認することで適合性重視の調査が可能になる。適合率向上に向けて特徴語抽出や用語の重み付け、ネットワーク分析による公報の注目特許抽出等をさらに検討する。

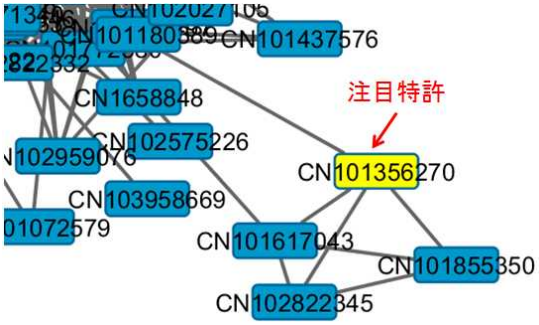

図10. 非類似度による表示(抜粋)

5. 考察

本検討は調査対象の母集団より系統的にキーワードを抽出しテキストマイニング的な手法で目的に応じた（適合率と再現率）検索キーワードを特定しようとするものである。テキストマイニングのモデルとして最初に良く使われているBag-of-words モデル [12]を試みた。このモデルは文書（公報）に単語が含まれているかどうかのみを考えて数値化する。文法、単語の順序、文構造、句読点等も無視する。形態素レベルのワードクラウドは単語の出現頻度をカウントするこのモデルである。専門用語抽出は品詞情報と単語の隣接情報を利用する単語のN-gram モデルである。単語ネットワークの特徴を生かしたモデルを試してみたい。

6. 結論

中国語キーワードを用いた特許情報解析に関して、①キーワードのネットワーク分析による重要語（特徴語）抽出、②キーワードを用いて計算した公報の非類似率によるネットワーク分析、③適合率を重視した特許調査への応用と一連の流れを検討した。

重要なキーワード群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重要キーワード抽出と選択が適合率向上のポイントである。

7. おわりに

本稿では中国語キーワードを抽出し、更に重要キーワードを選ぶ際の支援方法をテキストマイニングのモデルを考慮して検討した。また特許情報解析、中国特許調査への応用を検討した。重要キーワードの自動抽出を目標にテキストマイニング手法による特徴キーワード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2014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す。研究会のメンバーの皆様には様々な協力をしていただきました。特に昨年度の研究会メンバーである山村健一氏には貴重な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。

8. 参考文献

[1] 安藤 俊幸ら. “中国語キーワードによる中国特許情報解析” 第 10 回情報プロフェッショナルシンポジウム
[2] IKAnalyzerNet. http://www.piaoyi.org/c-sharp/IKAnalyzerNet.html accessed 2014.10.25
[3] ICTCLAS.http://ictclas.nlpir.org/ accessed 2014.10.25
[4] saezuri lite. http://www.vector.co.jp/soft/winnt/prog/se495669.html accessed 2014.10.25
[5] Cytoscape http://www.cytoscape.org/ accessed 2014.10.25
[6] 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. 63 No.12, (2013),1943-1957
[7] RMeCab http://rmecab.jp/wiki/index.php?RMeCab accessed 2014.10.25
[8] 末吉正成ら. “テキストマイニングを行う”. R ではじめるビジネス統計分析. 翔泳社, 20124, p. 304-334.
[9] 特許版事典検索システム Cyclone http://cyclone.cl.cs.titech.ac.jp/ accessed 2014.10.25
[10] WIPO Pearl http://www.wipo.int/wipopearl/search/home.html accessed 2014.10.25
[11] 語のネットワーク分析 http://www1.doshisha.ac.jp/~mjin/R/61/61.html accessed 2014.10.25
[12] Foster Provost ら. “テキスト表現とテキストマイニング”. 戦略的データサイエンス入門. オライリー・ジャパン, 2014, p. 275-306.