アジア特許情報のテキストマイニングによる解析：
自動テキスト分類への挑戦

○安藤俊幸[1)]，中西昌弘[2)]，道中孝徳[3)]，多田幸輔[4)]
花王株式会社[1)]，オリンパスメディカルシステムズ株式会社[2)]，ユーエムジー・エービーエス株式会社[3)]，富士フイルム株式会社[4)]
〒131-8501　東京都墨田区文花 2-1-3
Tel: 03-5630-9377　FAX: 03-5630-9712
E-mail: ando.t@kao.co.jp

Text mining analysis of Asian patent information.: Challenge to automatic Text Classification.

ANDO Toshiyuki[1)], NAKANISHI Masahiro[2)], MICHINAKA Takanori[3)] , TADA Kousuke[4)]
Kao Corporation [1)], Olympus Medical Systems Corporation[2)] , UMG ABS Ltd.[3)] , FUJIFILM Corporation[4)]
2-1-3, Bunka, Sumida-ku, Tokyo 131-8501 Japan
Phone: +81-3-5630-9377　Fax: +81-3-5630-9712
E-mail: ando.t@kao.co.jp

【発表概要】

アジア特許情報研究会でのテキストマイニングによる特許情報の解析、自動分類に関する研究結果を発表する。

アジアの特許情報活用に関して大別して下記2種類の検討を行った。

①フリーのツール類を用いたテキストマイニングの要素技術のシーズ面からの検討。
テキストマイニングによる各種解析、自動分類用の分類体系構築、特許の自動分類、調査・解析用ユーザー辞書（日本語、英語、中国語）作成支援、関心特許のスクリーニング。自動分類はLUT（参照テーブル）使用、自己組織化マップ、学習ベクトル量子化について検討した。

②テキストマイニング解析が可能な商用データベース PATENT INTEGRATION を用いたニーズ面から検討。
調査対象のテキストマイニング機能を使用した概要把握。

上記①②について報告する。

【キーワード】

テキストマイニング，テキスト自動分類，クラスタリング，termmi，統計解析言語 R，多次元尺度法，自己組織化マップ，学習ベクトル量子化，ベクトル空間法，可視化

1. はじめに

近年、中国を始めとしてアジアの特許情報活用の重要性はますます高まっている。またビジネスのグローバル化にともない特許調査に使用する言語も多言語化してきている。ようやく調査･解析ツールの多言語対応（ユニコード使用）が始まりつつある。現在普及している複数のパテントマップソフトの中国語対応も始まっていると聞いているが本予稿執筆時点ではまだ市販されていない。

パテントマップソフトを使用すると自分で入力した分類評価を利用して必要な情報をパテントマップとして視覚化できるが、そのための分析、評価作業に多くの労力がかかるという問題がある。アジアの特許の場合特に言語に関係する様々な問題が存在する。調査対象の検索集合（母集団）を一から読み込んで人手で分類するのではなくあらかじめ決めた分類体系に自動的に分類されるとその後の調査、解析が非常に楽である。また実務上、あらかじめ分類体系を決めるのも特許件数が多くなると難しくなるという課題も存在する。

そこでテキストマイニング解析を行いあらかじめ分類体系を決めて、決めた分類体系に自動分類する検討を行った。

2. 目的

膨大な特許情報の中からエンドユーザーにとって必要な情報を迅速に抽出･活用できる特許情報の分析･評価支援手法の開発を目的とした。インフォプロや研究員が調査テーマに関してパソコンで手軽かつ探索的に利用できるように特に下記3点を重視した。

(1) 検索結果の大まかな把握
（全体像の俯瞰、分類体系構築支援）
(2) 特許調査上の重要性に基づいた
必要な観点に自動分類支援
(3) 自動分類結果の戦略的利用
（商用ＡＳＰ型特許ＤＢ、市販
パテントマップソフトとの連携）。

調査･解析用ユーザー辞書（日本語、英語、中国語）はデータベース検索にも有用である。

3. 方法

解析、自動分類操作の概要を説明する。（解析操作は一部順不同）

①調査対象文書集合作成

特許データベース（NRI,QPAT, Thomson Innovation 等）で検索して調査対象文書集合を CSV ファイル（ヘッダ付）でダウンロードする。特許文書の全文を分析対象にする場合は全文テキストをダウンロードする。

②特許文書の分析対象部分を、要約、課題、解決手段、請求項、全文から選択し、分析対象の文書群を作成する。（EXCEL マクロ）

③termmi を使用して各文書毎に専門用語を重要度付きで抽出する。[2]

④-1 必要に応じて専門用語の重み付け（重要度）を調整する。（EXCEL マクロ）

④-2 必要に応じてワード（形態素：形容詞、動詞等）を重み付きで追記する。（EXCEL マクロ）

④-3 必要に応じて特許分類（IPC,FI,F ターム）に重み付けして追記する。（EXCEL マクロ）

⑤各文書間相互の非類似度（距離）を計算してファイルに出力する。（自作 VB.Net プログラム）[1]

⑥-1 各文書間相互の非類似度マトリックスを読み込み多次元尺度法により各文書間の相対的位置関係を可視化する。（統計解析言語R+MASS パッケージ）[1]

⑥-2PATENT INTEGRATION のクラスタリング機能を利用して検索集合の概要を把握する。[10]

⑦-1 抽出した専門用語－文書の統計デ

ータを集計する。(自作 VB.Net プログラム)
⑦-2 必要に応じて抽出した専門用語間の関係を解析、可視化する(対応分析、ネットワーク分析)。[1][9]
⑧各請求項より発明のカテゴリーを抽出して集計する。(EXCEL マクロ)
⑨自動分類用の LUT (参照テーブル)を使用して自動分類する。(EXCEL マクロ)
⑩-1 自己組織化マップ (Self-Organizing Map 以下 SOM と略記)で自動分類する。(R、自己組織化マップツール) [4][5][6][7]
⑩-2 学習ベクトル量子化 (Learning Vector Quantization 以下 LVQ と略記)で自動分類する。[7][8]
⑪自動分類結果をカラーマッピングして 3D表示する。(R+rgl パッケージ)
⑫必要により商用 ASP 型特許データベース、市販パテントマップソフト等へ分類結果をフィードバックする。
⑬解析、分類結果の検証

4. 結果

4－1. 検討用テスト集合

特許情報の分析・評価支援手法の検討用テスト集合としてインクジェットカートリッジ分野の下記検索集合をダウンロードして用いた。JP の F ターム 2C056KC** (インクタンク) 8846 件の集合で US 公開 and CN がある公報を抽出し CN の重複を除去、ファミリー数が異常に多い2件を除去した。JP,US 公開,CN 各々766 件を母集団とした。
使用 DB：Thomson Innovation,NRI
検索期間：1993.01.01～2010.12.31

4－2. 専門用語－文書行列作成

termmi を使用して各特許文書毎に専門用語を抽出する。専門用語は各文書ごとに重要度付きで抽出されファイルに出力される。文書により同じ専門用語でも重要度 (重み) が異なる。そこで専門用語－文書の集計機能を自作 VB.Net プログラムに追加した。ハッシュテーブルを使用して専門用語－文書の統計情報を高速に集計できる。集計結果より特定の専門用語が含まれる文書数を参考にして特徴 (重要) 専門用語を選択できる。termmi を使用することで日本語だけでなく英語の専門用語も抽出できる。termmi に付属のツール "termdocument.pl"で「用語・文書行列」を出力することができる。中国語の専門用語は参考文献[4]の「言選Web」(中文版) で抽出した。

4－3. テキストマイニングによる解析

多次元尺度法(MDS: multi-dimensional scaling)は、個体 (文書) 間の親近性データをもとに低次元 (2次元あるいは3次元) 空間に類似したものを近く、そうでないものを遠くに配置する方法である。

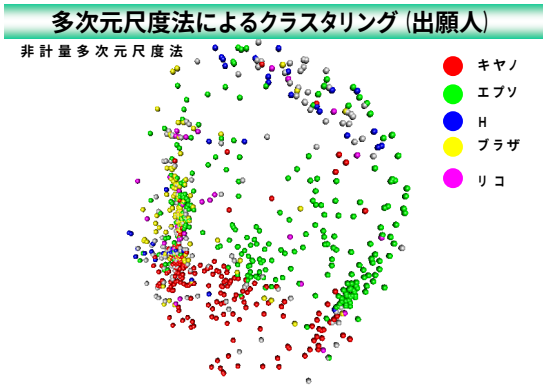

図1. 多次元尺度法によるクラスタリング

また PATENT INTEGRATION の「クラスタ」機能でクラスタリング検討を行った。PATENT INTEGRATION のクラスタリングは専門用語 (複合語) ではなく形態素解析結果の名詞レベルで行っている。10 個のクラスタに色分けされ各クラスタの重要語が表示され、チェックボ

タンで選択したクラスタの表示をオン/オフできる。またクラスタ上で出願人、発明者で色分け表示可能である。特許データベースで検索してそのままデータをダウンロードすることなくテキストマイニング解析を行うことができ便利である。マップ表示された個々の文献番号、発明の名称、出願人、発明者、重要語、座標データは表示だけでなくダウンロード可能である。ただクラスタ番号はダウンロードできない。PATENT INTEGRATION のクラスタリングアルゴリズムは公開されていないようである。

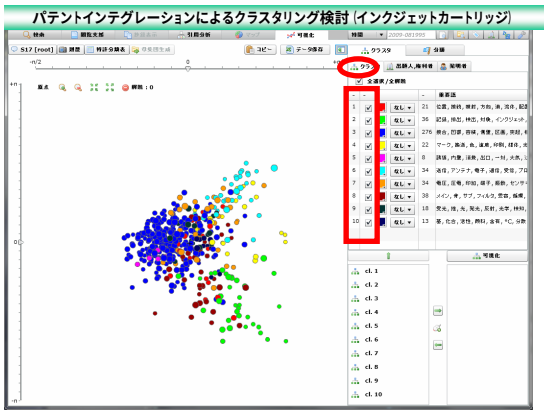

図2. PATENT INTEGRATION

4－4. テキスト自動分類

(1) LUT (参照テーブル) 使用

発明のカテゴリー (請求項の最後のターム) マッチングによる自動分類の流れを下記に示す。

①請求の範囲 (全請求項) を各請求項に分割
②各請求項を文に分割 (「。」で分割)
③最初の文を末尾からスキャンして漢字、カタカナ部分を抽出する
⑤抽出したカテゴリーランキングを作成
⑥カテゴリー自動分類用 LUT (参照テーブル) 編集
⑦分類表読込 (Excel マクロ)
⑧自動分類実行 (Excel マクロ)

簡単なアルゴリズムの割には発明の「物」の分類には良い精度が得られる。「方法」の場合は、方法の発明ということは分かるがより詳しく発明を特定するには更に工夫が必要である。要約の「課題」「解決手段」の専門用語のマッチングによる自動分類も可能である。ただし分類精度は請求項の文末のマッチングによる発明のカテゴリーの方が良い。

(2) 自己組織化マップ

コホネン(T.Kohonen)により連想記憶という人工ニューラルネットワークの研究の中で提案されたアルゴリズム。高次元データを2次元平面上へ非線形射影してマップを描く。類似したもの同士が近くに配置される。

①自己組織化マップの入力データ

専門用語－文書行列の行と列を入れ替えた (転置)、文書－専門用語行列を入力とした。専門用語上位200語からノイズを除いた 190 語 (次元) を使用してマップ化した。

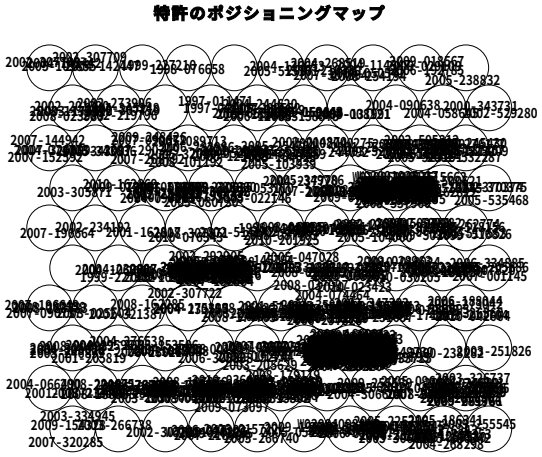

図3. 特許の自己組織化マップ

(3) 学習ベクトル量子化

学習ベクトル量子化[7][8]は教師あり学習法であり、SOM を発展させ統計的クラス分けのためにコホネンによって考案されたものである。R では class パッケージ でサポートしている。

5. 考察

多次元尺度法による非類似度（距離）を用いた文書間のクラスタリングの特性として予め決められた分類体系に当てはめる自動分類への応用には限界があると考えられる。

自己組織化マップ、学習ベクトル量子化による自動分類では複数の観点をそれぞれ考慮した分類が可能である。

LUT（参照テーブル）使用のテキスト自動分類はマッチングが成立して分類できると発明のカテゴリーの分類では分類目的にも依存するが満足のいく分類結果が得られる。ただしマッチングが成立しないと分類不能である。その点自己組織化マップ、学習ベクトル量子化は類似の分類項目に分類される。

クラスタリング解析と自己組織化マップ、学習ベクトル量子化による自動分類を補完的に使いこなすと良いと思われる。

6. 結論

テキストマイニングによる解析から分類体系構築、特許の自動分類まで一連の流れを検討した。

専門用語－文書の統計情報、専門用語間の関係を解析、可視化することで調査･解析用ユーザー辞書（日本語、英語、中国語）作成支援を行うことができる。ユーザー辞書を用いて文書の自動分類が可能である。用語の選択が重要である。

7. おわりに

本稿ではテキストマイニングによる各種解析から3種類の方法でテキストを自動分類することを試みた。自己組織化マップ、学習ベクトル量子化については学習パラメータの検討等が不十分であり手法の性能をまだ充分に引き出しきれていない。今後実務データを用いてさらに最適化検討を進めたい。自己組織化マップはトーラス型SOM、球面SOM、高速球面 SOM 等更に発展しているのでこれらについても検討してみたい。

8. 参考文献

[1] 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 .

[2] 前田朗. ”専門用語（キーワード）自動抽出システムのページへようこそ”. http://gensen.dl.itc.u-tokyo.ac.jp/ (参照 2011-08-05)

[3] 金明哲. “多次元尺度法”. Rによるデータサイエンス. 森北出版,2007, pp97-106

[4] 金明哲. “自己組織化マップ”. Rによるデータサイエンス. 森北出版,2007, pp127-133

[5] 豊田秀樹編著. データマイニング入門:R で学ぶ最新データ解析. 東京図書, 2008.

[6] 津高信一郎. “自己組織化マップを用いたテキスト自動分類の試み”情報処理学会第 46 回全国大会 (4).1993-03-01. pp.187-188

[7] 大北 正昭編著. 自己組織化マップとそのツール. シュプリンガージャパン,2008.

[8] 金明哲編. “学習ベクトル量子化”. R で学ぶデータサイエンス5 パターン認識. 協立出版, 2009, pp. 100-106.

[9] 石田基広. “RMeCab によるテキスト解析”. R によるテキストマイニング入門. 森北出版, 2008, pp. 51-82.

[10] 山下佳之. “テキストマイニング技術の特許分析･特許検索実務への活用 特許検索･分析サービス「パテント･インテグレーション」”. 情報管理. Vol. 52, No. 10, (2010), 581-591 .