特許文献から技術動向を把握するためのマイニング手法

Mining Method for grasping technical trends from patent literature

株式会社 NTT データ数理システム　データマイニング部主任研究員　岩本　圭介

1999 年株式会社数理システム（現：株式会社 NTT データ数理システム）入社。同社のテキストマイニング事業の立ち上げ時より、一貫してツール開発・手法開発及び分析業務に携わる。現職はデータマイニング部主任研究員。

iwamoto@msi.co.jp

1 はじめに

特許文献から世の技術動向を把握し、そこで得られた情報をもとに研究開発の方向付けや事業戦略の組み立てを行うということが、各方面で広く行われている。そのようなプロセスにおいては、大量の情報を分析して何らかの有益な知見を導き出すことができる手段が必要である。特に、特許情報の主要な部分は文書（テキスト）データであるため、この目的に貢献する分析手法としてテキストマイニングの利用が大いに認知されてきている [1]。

本稿では、ある特許データの集合に対して、一般に書誌情報に明示されている特許文献の分類コードから一旦離れ、テキスト情報のみを用いて、人間の直感にできるだけ合致した分類項目を自動的に作成して技術動向把握の助けとなる情報を抽出することを試みた。以下、2 章において分析に用いた手法について論じ、3 章でその実践内容を解説する。最後に 4 章で結言としてまとめを行う。

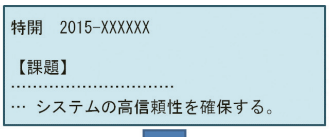
本稿の分析は、株式会社 NTT データ数理システムのマイニング製品群である **Visual Mining Studio** 及び **Text Mining Studio** を用いて行った。

2 手法

テキストマイニングは、テキストを自然言語処理技術により自動的に単語へ分解してそれらの間の文法的構造を求めるステップと、得られた単語群の情報に対して分析を行う実際に傾向把握を行うステップからなる。最初のステップは、いわばテキスト情報を分析可能な「データ」に変換するために必要なものである。そして、次のステップで、得られた「データ」に対して統計やマイニングの諸手法を適用していく。

2.1 特許文献への自然言語処理技術の適用

計算機にとっては元来文字の羅列であるテキスト情報を単語単位に分割する形態素解析、及び単語間の修飾関係を求め文章としての構造を明らかにする構文解析それぞれの機能は、それぞれオープンもしくはプロプライエタリで多数ツールが存在しておりそれらを用いることも可能である。特許文献のテキストにこれらを適用させ、「データ化」を行った結果のイメージを **Text Mining Studio** のアウトプットに準じた形で示す（図 1）。



公開番号	単語ID	見出し語	原形	品詞	品詞詳細	係り先
:	:	:		:	:	:
特開2015-XXXXXX	1	システムの	システム	名詞	一般	2
特開2015-XXXXXX	2	高信頼性を	高信頼性	名詞	一般	3
特開2015-XXXXXX	3	確保する。	確保	名詞	サ変可能	-1
:	:	:		:	:	:

図 1　特許文献情報データ化のイメージ

図 1 では、形態素解析の結果は「システムの高信頼性を確保する」という文字列が、単語に分解された上で ID や品詞情報が付与されるという形で反映されている。また、構文解析の結果は「係り先」という列において、

その単語が他のどの単語を修飾するか、という ID を別途付与する形で反映されている。

ここで、文節末尾に付される助詞や助動詞、及び活用による表記の違いは吸収･統一しておくことが望ましい。図 1 では、原文そのままの表記「見出し語」列の他に、これらの統一を行った「原形」列を与えることで解消している。

また、特許文献を扱う際には、以下の特別な事情が存在するため、これらを勘案する必要がある。

- 名詞が複数繋がることで 1 つの意味ある技術用語が形成されることが非常に多い。更に、こういった用語は技術の革新とともに常に新たな語の組合せ・パターンが生じるため、逐一列挙することはほぼ不可能である。適切に連語を連語として扱える仕組みが必要である。
- 特許文献特有の定型文や定型語句が多数あり、クリーニングを行わなければ、これらは本来注目すべき技術用語を埋もれさせてしまうノイズとなる。例えば「本発明」「請求項」等の語句は、頻出するものであるが技術動向の把握において意味のあるものではない。ユーザの指定により、こういった語を分析対象から除外することができる機能が必要である。

2.2　文書のベクトル表現

前節図 1 のデータは、公開番号との紐付けがなされたうえで、「1 レコード 1 単語」の形式で単語の列挙を表形式で保持しているものといえる。しかし、1 件 1 件の特許文献を 1 件のデータとしてマイニング手法を適用させる場合は、「1 レコード 1 特許文献」の形式の数値データにこの情報を落とし込むことが望ましい。そのために、出現した単語を各成分に取り、単語出現の状況を数値に落とし込んだベクトル表現を作成する（図 2）。この行ベクトルを 1 件の特許文献と同一視することができる。

2.3　分類項目作成のためのクラスタリング

図 2 の下表は、行で表されるそれぞれの特許文献に対して、列で表されるそれぞれの単語が出現しているか否か、を示している。この行ベクトルに対し、クラスタリングによって類似したベクトルのまとめ上げを行うことで特許文献の分類を試みる。

公開番号	原形
:	
特開2015-XXXXXX	システム
特開2015-XXXXXX	高信頼性
特開2015-XXXXXX	確保
:	
特開2014-YYYYYY	システム
特開2014-YYYYYY	安定性
特開2014-YYYYYY	確保
:	

公開番号	...	システム	高信頼性	確保	安定性	...
特開2015-XXXXXX		1	1	1	0	
特開2014-YYYYYY		1	0	1	1	

図 2　特許文書のベクトル表現

ここで、クラスタリング手法としては、二項ソフトクラスタリングを用いた。二項ソフトクラスタリングは、株式会社 NTT データ数理システムのデータマイニングツール **Visual Mining Studio** の機能である。以下、二項ソフトクラスタリングの解説を行う [2]。

特許文献を d_i として特許文献の集合を $D=\{d_1, d_2, ..., d_N\}$、またある単語を w_j として単語の集合を $W=\{w_1, w_2, ..., w_M\}$ とする。二項ソフトクラスタリングは、D と W の間に潜在的なクラスタ $Z=\{z_1, z_2, ..., z_L\}$ が介在しており、d_i と w_j の同時発生確率 $P(d_i, w_j)$ が次式

$$P(d_i, w_j)=\sum_{k=1}^{L} P(z_k)P(d_i|z_k)P(w_j|z_k)$$

で表されるものと考え、実データに対しての当てはまりが良くなるようこの $P(z_k)$, $P(d_i|z_k)$, $P(w_j|z_k)$ を算出する手法である。これらが算出できれば、文献 d_i がクラスタ z_k に所属する確率 $P(z_k|d_i)$ も求めることができる。広く用いられている k-means 法などは分類対象である要素を、常にどれか 1 つのクラスタに割り振るタイプのクラスタリング（ハードクラスタリング）であるが、二項ソフトクラスタリングは分類対象である要素が複数のクラスタに所属することを許容する（ソフトクラスタリング）ものである。その場合、量 $P(z_k|d_i)$ は d_i の各クラスタへの帰属度合いとみなすことができる。本稿では、d_i の所属先の重なり具合を分析することで、クラスタ間の関係について評価することも目的としている。また、所属先のクラスタを $\arg\max_k P(z_k|d_i)$ なる（最も所属確率の高い）クラスタ k に決定すると、これはハードクラスタリングとしても用いることができる。

二項ソフトクラスタリングでは、文献 d_i 側だけでな

く、単語 w_j 側についても同時にクラスタリングを行う。特に、値 $P(w_j|z_k)$ はあるクラスタ z_k における単語 w_j の出現確率を表しており、この確率値が高い単語を観察することで、それぞれのクラスタがどのような観点でまとめられたものであるかという情報を得ることができる。

上記は PLSA（確率的潜在意味解析：Probabilistic Latent Semantic Analysis）として知られる手法と類似しているが、**Visual Mining Studio** の二項ソフトクラスタリングでは、PLSA と NMF（非負値行列分解）とのハイブリッドによる手法の適用により精度向上が試みられており、初期値設定や収束判定等に工夫を施し大規模データに耐えられる実装となっている [3]。

3 実例

3.1 対象データ

今回の分析課題を、『スマートフォンや携帯端末の位置情報は、どのような場面で利用が試みられているのか、特許文献から把握する』こととした。サンプルとして使用したデータについての情報を表 1 に示す。

表 1　サンプルデータ

件数	675 件
検索条件	要約 に 『(スマートフォン OR 携帯端末) AND 位置情報』
期間	出願日が 2010 年～ 2015 年

得られたデータの「課題」部分のテキストに対して形態素解析･構文解析を行い、表 2 に示す単語を除外した。

表 2　削除した単語

種別	例
特許文書特有の 頻出定型表現	本発明、提供、請求項、など 約 70 単語
単独の英数かな文字	1、2、A、B、あ、い、う など
データ取得時の検索 条件に含まれる単語	スマートフォン 携帯端末 位置情報

3.2 データの概観

2.3 章で示したクラスタリングを試行する前に、特許文献のテキストに出現する単語や係り受け（構文解析の結果求まった、修飾 - 被修飾 の関係にある語の組）の状況を概観しておく。

「課題」部分のテキストに出現した単語のランキングを図 3 左に示す。全体的に「容易」「利便性」などユーザにとっての使い易さを訴える単語が上位を占めており、具体的な適用場面はこの時点ではあまり見えてこない。図 3 右は、特に「～性」で終わる単語のみに限ってランキングを作成したものである。「利便性」以外にも「信頼性」「セキュリティ性」等の観点が存在することがわかる。

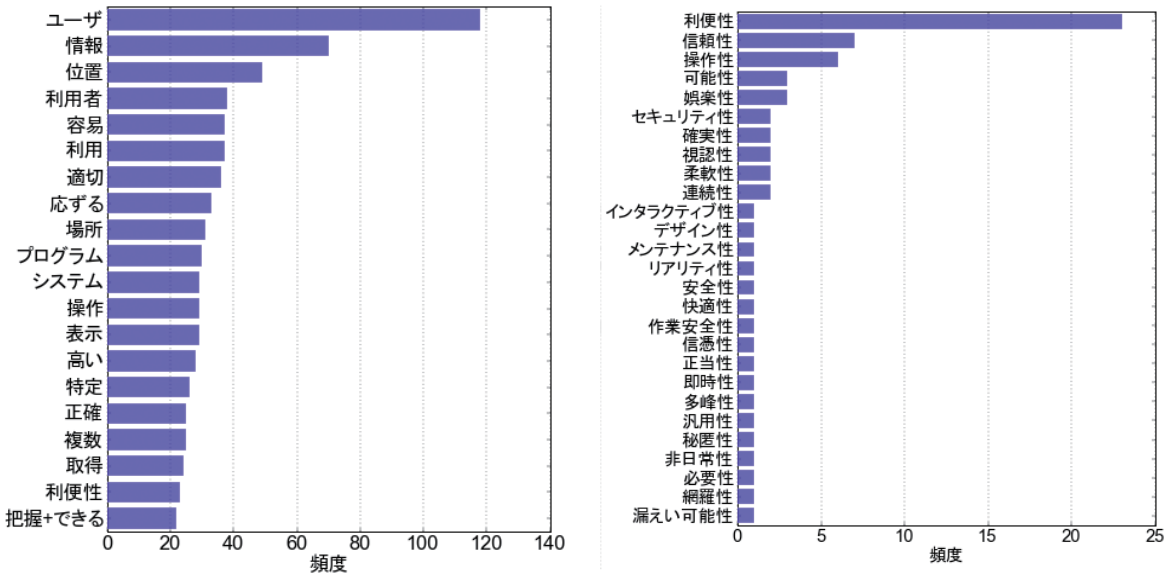


図 3　単語出現頻度（左）、特に「～性」に限ったもの（右）

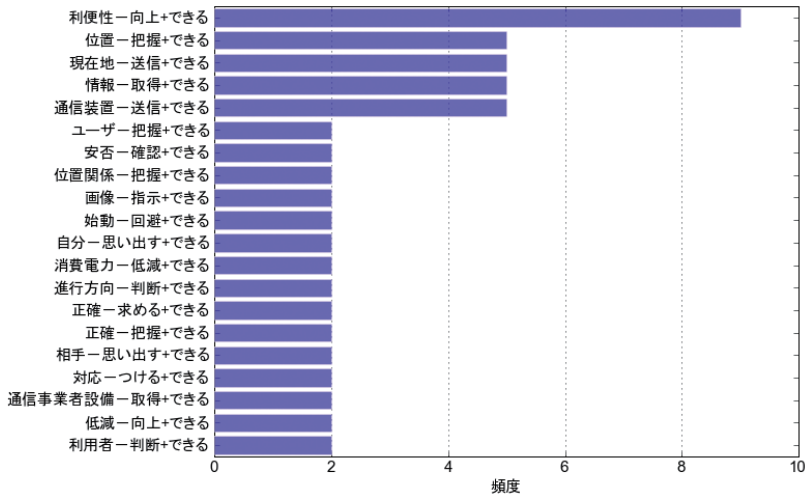


図４　係り受け頻度、「～できる」という表現に限ったもの

図４に、係り受け表現のランキングを示す。係り元－係り先の組を集計しているが、係り先の単語の語尾が「～できる」「～可能」というように、記述者が『可能である』という態度を表明しているような表現のみに絞っている。これにより、その発明が何を可能とするか、そのメリットを抽出することができる。利便性の向上の他に、安否の確認、消費電力の削減などへの言及がみられる。

3.3　クラスタリングによる把握

続いて、2.2 章～ 2.3 章に示した手法を用いて、「課題」のテキストをクラスタに分類し、表３の結果を得た。クラスタ $Z=\{z_1, z_2, ..., z_L\}$ の個数は、$L=40$ とした。表３の『クラスタ ID』の列には各クラスタで $P(w_j|z_k)$ の値が最も大きかった単語、『代表語』にはその他の $P(w_j|z_k)$ 上位単語をあわせて示した。『件数』には、特許文献 d_i 毎に $\arg\max_k P(z_k|d_i)$ となるクラスタ k を採用した場合にそのクラスタに割り振られる特許文献の件数を示した。図２のベクトル表現を作成する際に、列を構成する単語が１つも含まれていない特許文献は零ベクトルとなってしまい分類の対象にならないため（表の「その他」にあたる）、ある程度の量の単語を用いてこの列を構成する必要がある。しかし、これが多過ぎても非常に瑣末な情報が含まれることになるため、「全データの 90% がカバーできるだけの名詞」でベクトル表現の列（次元）を作成するようにした。

表３を概観すると、図３、図４の単語や係り受けの頻度集計と比較して、話題の全体がカバーされており、スマートフォンや携帯端末の位置情報が利用されているであろう状況が俯瞰的に見渡せるようになっていることがわかる。

ここを掘り下げの基点として、更に絞り込んで元々のテキストを参照することで具体的な理解につなげることができる。例えば「No.13 駐車位置」のクラスタに関しては、位置情報を利用して駐車位置を忘れても自分自身が駐車した車の位置を知ることが可能なシステムの実現について多く述べられている。また、「No.26 患者」については、同時に「看護師」も代表語として挙げられているが、ナースコールによる患者からの呼出し際に、看護師の実際の居場所を考慮して制御を行う手段について述べられている。

3.4　クラスタ間の関係

前節で試行したクラスタリングはソフトクラスタリングであるため、実際は１件の特許明細が確率値 をもって複数のクラスタに分類される。ここで、同時に割り振られやすいクラスタ同士の関係を図５に図示した。図５は、丸で表されているノードがそれぞれのクラスタ、線で結びついているもの同士は１件の特許に同時に付与されやすいクラスタを示している。

先述の「No.13 駐車位置」は、「No.31 車両」「No.17 目的地」と同時に発生することが多く、車両の制御関連という大きな括りでの適用が存在することが分かる。「No.8 顧客」「No.15 店舗」は、原文を参照すると、店舗内での顧客行動の取得や、顧客端末を利用しての注文形態がトピックとなっている。

表３　クラスタ毎の代表単語

クラスタ ID	代表語	件数
No.1 アプリケーション	アプリケーション	25
No.2 コンテンツ	コンテンツ	25
No.3 自分	自分 線区	22
No.4 画像	画像 範囲 測位機能 通信事業者設備	22
No.5 消費電力	消費電力	20
No.6 地域	地域 番組 災害	19
No.7 画像表示装置	画像表示装置 移動中 ＮＡＭ発言者 現実世界 無線ＬＡＮ通信部	19
No.8 顧客	顧客 両者 保守員 エレベーター 点検作業	18
No.9 サーバ	サーバ	18
No.10 対象	対象 広告効果	18
No.11 精度	精度 自機	18
No.12 画像形成装置	画像形成装置	17
No.13 駐車位置	駐車位置 作業負担 路線	17
No.14 効率	効率 文字入力＋できる 携帯電話機 指一本	17
No.15 店舗	店舗	17
No.16 周囲	周囲 セキュリティ 投稿者	16
No.17 目的地	目的地 環境	16
No.18 管理サーバ	管理サーバ 避難経路 要請者 通信障害 部品準備作業	16
No.19 画像処理装置	画像処理装置 作業者 外出先	16
No.20 ＲＦＩＤタグ	ＲＦＩＤタグ 携帯電話 従業員 要員 内容	15
No.21 乗車中	乗車中 鉄道車両 地点 乗員 経路	15
No.22 高精度	高精度 方向	15
No.23 エリア	エリア	14
No.24 文書蓄積サーバ内	文書蓄積サーバ内 度合い 文書蓄積サーバ 管理対象物 行動履歴	14
No.25 操作性	操作性 通信タグ	14
No.26 患者	患者 運転者 看護師	14
No.27 制御方法	制御方法 制御プログラム	13
No.28 現在地	現在地 移動手段	13
No.29 作業員	作業員 表示装置 大型化	13
No.30 信頼性	信頼性 遊技者	13
No.31 車両	車両	13
No.32 広告毎	広告毎 回数 街頭広告 前方	12
No.33 屋内	屋内 手間 放射線量 シームレス 屋外	12
No.34 移動経路	移動経路 タイミング データ	12
No.35 メッセージ アプリ	メッセージ アプリ	10
No.36 基地局 使用者	基地局 使用者	10
No.37 道路上	道路上 被保護者 確実性 緊急通報方法 空調機制御装置	9
No.38 通信履歴	通信履歴 データ 基地局識別子 １項 生活パターン	9
No.39 基地局	基地局 各分割領域 確率分布 移動確率 移動対象物毎	7
No.40 参照点	参照点 観測点 地図上	5
その他	代表語がありません	67

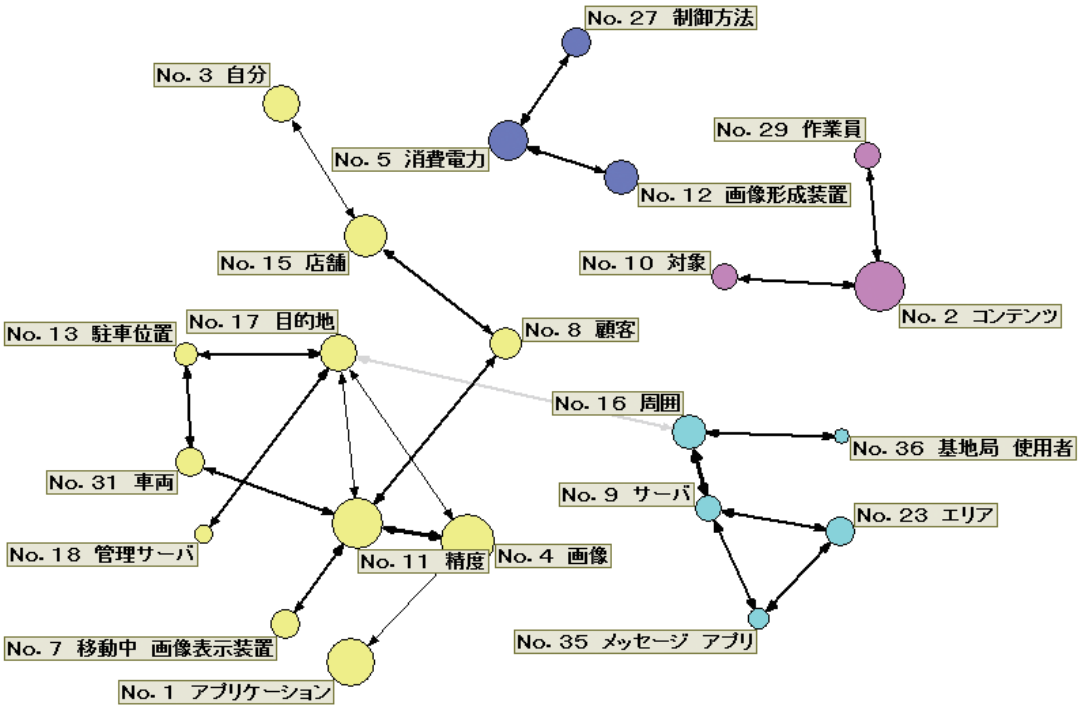


図５　同時に割り振られやすいクラスタ間の関係

4 まとめ

本稿では、特許データにソフトクラスタリング手法を適用して、データ全体を概観するための分類を作成するための手法を論じた。分析例では、自社の技術がどういった方面に活用できるか、ひいてはビジネスとして展開できるか、という点を抽出することを意識している。今回、書誌情報は用いずに「課題」部分のテキストのみを扱ったが、生成されたクラスタと出願人・発明者といった情報とを掛け合わせることで、競合他社の動向把握などより踏み込んだ解析を行うことも可能である。

参考文献

[1] 豊田裕貴 菰田文男 編著（2011）『特許情報のテキストマイニング』ミネルヴァ書房

[2] 株式会社 NTT データ数理システム（2016）『**Visual Mining Studio** マニュアル バージョン 8.2』

[3] 若杉徹 高橋勲男（2014）「医薬品調剤履歴に関する確率的構造解析に基づく適応症の推定」2014 年度人工知能学会全国大会論文集