特許情報テキスト可視化のためのマイニング手法

Mining methods for the patent text visualization

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

PROFILE　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

特許情報の分析においては、整理・収集した情報に対して集計を行い、その結果をマッピングすることが広く行われている。特許情報をパテントマップといったビジュアルな形にまとめ上げることは、直感的な理解を促す上で非常に重要である。ここで、集計によって得られる数値情報、例えば年毎の出願件数の推移や出願人の構成比率などといったものは、棒グラフや円グラフ、もしくはバブルチャートといった種類のグラフに展開することが可能である。しかし、特許情報の「テキスト」部分の情報を可視化するためには、与えられたデータからその中に潜む傾向や法則性を抽出するデータマイニング・テキストマイニングの技術が必要となる。

本稿では、特許情報テキストの可視化を行ううえで有用であるマイニング手法を解説する。以下、２章において、テキストをいかにデータ化するか、またこれをいかに平面上のマップの配置として表現するかという点を論じる。３章で具体的な作成例を示し、最後に４章で結言としてまとめを行う。

2　特許情報テキストの可視化

特許情報テキストの可視化が目指すところは、特許情報の全体像はどのようになっているか、またその中にそのようなまとまりが存在しているか、といった情報を直感的に掴むことの助けになる情報を提供することである。ここで、特許の「まとまり」を論じるためには、特許文書同士の間に「近さ・遠さ」というものが定義できる必要がある。

この「近さ・遠さ」は、テキストデータから確固としたアルゴリズムによって機械的に計算できるものでなくてはならないが、人間の直感からかけ離れたものであってもいけない。一般には、これは１件１件の特許文書を１データ点と考え、それらの間の点の距離として近さ・遠さを表現する。そして、特許データ間の距離関係を出来るだけ保ったまま、近しい特許同士はマップ上の近い点に存在し、そうではない遠い特許同士はマップ上でも離れて位置するような配置を実現することを目指す。この模式図を図１に示す。

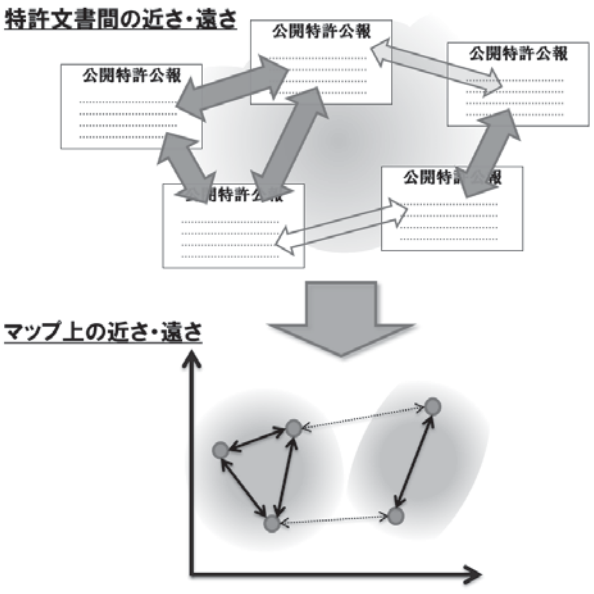

図１　文書間・マップ間の近さ・遠さ

2.1　文書のベクトル表現

１件１件の特許文書を１データ点として考えるために、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）。このベクトルを１件の特許文書と同一視することで、ベクトル同士の距離として特許文書間の距離を定義することができる。

ベクトルの各成分として用いる単語は、特許の特徴を上手く反映するものを使用することが望ましく、実際には極端な高頻度語や低頻度語を無視する、tf-idf 等の指標を用いて選択を行う、といった処理をここで行う。

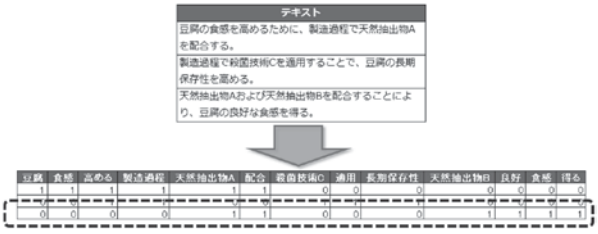

図２　特許文書のベクトル表現

2.2　次元圧縮の利用

前章で作成した特許文書のベクトル表現は、一般には極めて多次元の情報である。これを、人間が知覚可能な形で表現するため、この多次元空間内での特徴、すなわち距離関係を出来るだけ保ったまま２次元、もしくは３次元の点に落とし込む（次元を圧縮する）必要がある。

一般に、多次元空間内の距離関係を完璧に保ったまま２次元の点間の関係として表現することは不可能である。例えば、図３左は互いの距離が全て１になるように A～D の４点を配置した例である。こういった配置は３次元上では可能で、実際には正四面体の頂点をなす。しかし、「互いの距離が全て１になるような４点」を２次元上で配置させることは不可能である。例えば図３右のように配置させた場合、A-B, A-D 間の距離は１でこれが保たれているが、A-C 間の距離は$\sqrt{2}$となって元来の３次元空間よりも距離が引き伸ばされて表現されてしまっている。このように、次元を圧縮すると必ずどこかにこのような歪みが生じる。このことは避けられないため、この歪みをなるべく引き起こさせないような低次元空間での配置方法を考える必要がある。

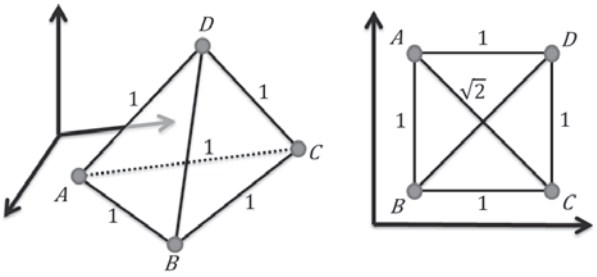

図３　２次元で表現することによる歪み

2.2.1　多次元尺度構成法

多次元尺度構成法（MDS）は、対象データ間の関係を反映するように低次元空間の点の配置を決める代表的な方法である[3],[4]。以下、多次元尺度構成法のバリエーションのうち、計量的多次元尺度構成法と呼ばれるものについて解説する。

計量的多次元尺度構成法では、多次元空間内での特許文書 i と特許文書 j との間の距離の２乗 d_{ij} を要素とする $t \times t$ 行列 D を与え、次の手続きで計算を行う。

1. D に Young-Householder 変換を行った行列 P を求める。

$$P = -\frac{1}{2} HDH,\ H = I - \frac{1}{t} ee^T$$

（ただし、e は全ての要素が１の列ベクトル）

2. P を特異値分解して、直交行列 V と対角行列 Λ を用いて次のように表す。

$$P = V \Lambda V^T$$

3. 対角行列 Λ の固有値の大きい方から d 個取出し、それ以外を０とした行列を Λ_d とし、取り出した固有値に対応する固有ベクトルを並べた行列を V_d する。$Y = \Lambda_d^{1/2} V_d^T$ とおくと、$P_d = V_d \Lambda_d V_d^T = Y^T Y$ は、P を $\mathrm{tr}((P - Y^T Y)^2)$ の意味で最小化したものになっている。

この Y を d 次元の座標としてプロットする。

2.2.2　Random Projection

x_{ij} を、特許文書 i における単語 j の出現状況を表す数値とし、その x_{ij} を要素とする行列 X を考える。X は、図２で示した特許文書ベクトルを全文書分縦に並べたものであると考えられる。

Random Projection は、(i, j) 行列である X に、Random Projection Matrix と呼ばれる (j, k) 行列

R を乗じ、得られた行列 X' の行ベクトルを、k 次元に圧縮された特許文書ベクトルとして用いる方法である[1], [2]。このイメージを図４に示す。

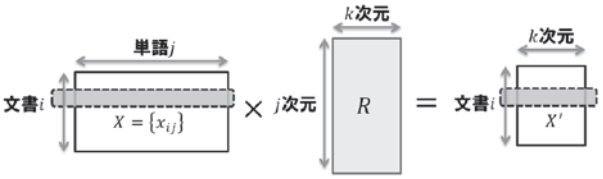

図４ Random Projection のイメージ

R の要素 $\{R_{ij}\}$ は、正規乱数を用いる、またその他に次のような構成方法が知られている。

$$R_{ij} = \begin{cases} +\sqrt{3} & \text{with probability } 1/6 \\ 0 & \text{with probability } 2/6 \\ -\sqrt{3} & \text{with probability } 1/6 \end{cases}$$

2.2.1 章で解説した計量的多次元尺度構成法のような行列の固有値計算が必要なく、計算コストが大幅に抑えられる点が特徴である。

3 実例

3.1 対象データ

２章で解説した手法を用いて、２次元上へのマップ作成を試みた。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。

得られたデータの「課題」「解決手段」部分のテキストからキーワードを抽出し、その中から特許文書特有の頻出定型表現と上記検索条件に相当する単語を除外したデータから、頻出 400 単語を用いて特許文書ベクトルを作成した（すなわち、2.2.2 章の次元数 j = 400）。この情報に対して、多次元尺度構成法及び Random Projection により、次元圧縮を試みた。

表１ サンプルデータ

件数	281 件
検索条件	要約・請求の範囲 に 『(麺 OR めん) AND (即席 OR インスタント)』
期間	出願日が 2001/01/01 ～ 2010/12/31

●多次元尺度構成法

多次元尺度構成法によって決定した 281 点の特許データの２次元上の配置を図５に示す。参考のため、データを点ではなく、出願人を表す文字でプロットしている（ただし、281 件内での出願件数の多い１位～10 位にあたる特許を文字 A ～ J、それ未満のものは Z としている）。特に点線で囲った出願人 F と H が離れている、すなわち他の特許と比較してキーワードの用いられ方が異なっているといえる。その他、出願人 D のまとまりも見て取れる。

また、多次元(400 次元)空間の距離が低次元(2 次元)空間での距離にうまく反映されているか確認を行った。特許データの全ての組合せに対して、400 次元空間と２次元平面上での距離を求め、前者を横軸・後者を縦軸にとったプロットを図６に示した。点線で囲った領域、元々のデータのうえでは離れているが２次元上では近くなってしまっているようなケース目立つが、相関係数は 0.82 であり、大まかな傾向は反映できているといえる。

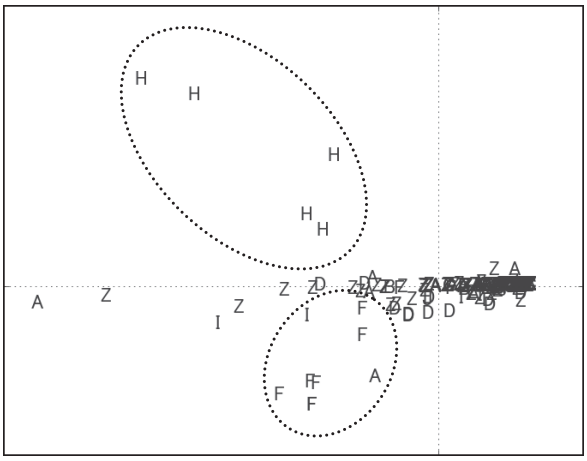

図５ 多次元尺度構成法による配置

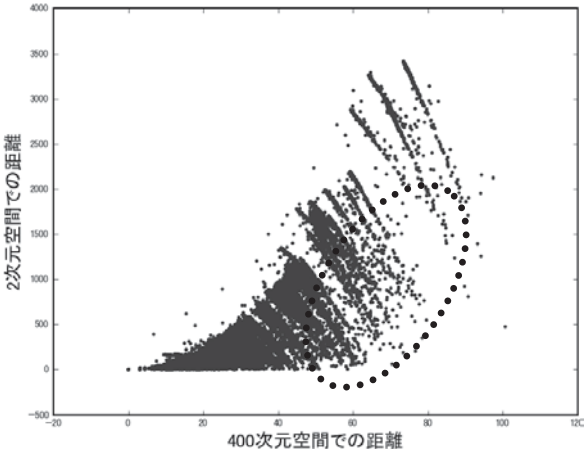

図６ 多次元尺度構成法の結果における距離関係

● Random Projection

Random Projection によって同様の図示を行ったものが図７・図８である。$\{R_{ij}\}$ は正規乱数で構成した。図７では、出願人 E,F のまとまりが緩やかにみられる。図８は、相関係数は 0.57 であり多次元尺度構成法と比較するとばらつきがみられるが、算出過程の単純さを鑑みると一定の利用価値は存在するものと考える。

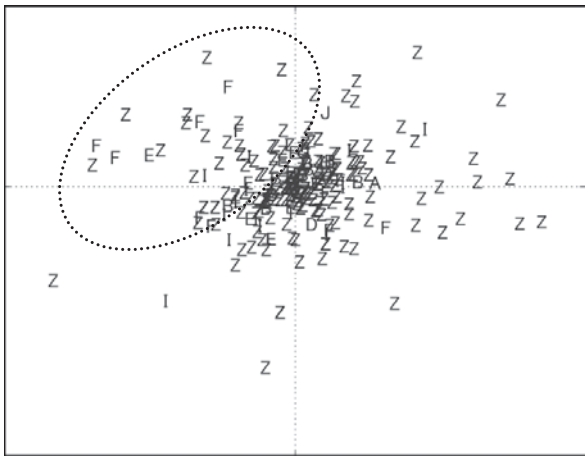

図７：Random Projection による配置

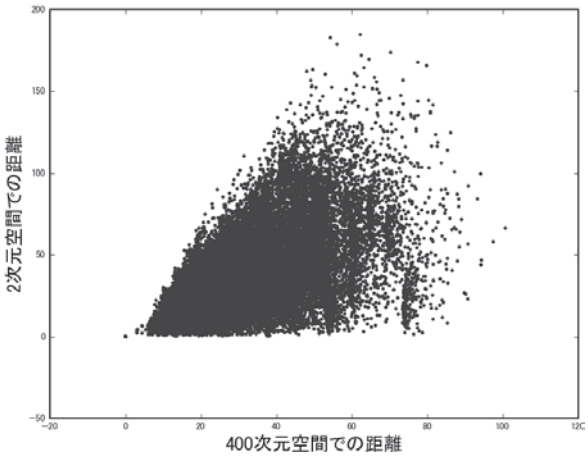

図８：Random Projection の結果における距離関係

4 まとめ

本稿では、特許文書 1 件を 1 点としてマッピングを行う方法を論じ、結果の評価を行った。筆者の経験から、テキストマイニングによる特許データの可視化はブラックボックスのようで何を示しているのか判然としないとの印象を持たれることも多いが、本稿がそういった印象を払拭することの一助となれば幸甚である。本稿の分析結果は株式会社 NTT データ数理システムの **Visual Mining Studio** 及び **Patent Mining eXpress** といった製品群によるものである。

参考文献

[1] D. Achlioptas (2001) Database-friendly random projections: Johnson-Lindenstrauss with binary coins, In Proc. of ACM Symp. on Principles of Database Systems, pages 274-281

[2] E. Bingham and H. Mannila (2001) Random projection in dimensionality reduction: Applications to image and text data, In Proc. of 7th ACM SIGKDD ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, pages 245-250

[3] 岡太彬訓 今泉忠 (1994)『パソコン多次元尺度構成法』共立出版

[4] 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