論　説

テキスト分析ツールを活用した特許分類業務の効率化

知的財産情報検索委員会
第　1　小　委　員　会*

抄　録　テキスト分析に基づいて特許文献を分類する自動分類ツールおよびテキストマイニングツールに注目し，特許分類業務を効率化するにはこれらのツールをどのように使うべきか，特にツールの性能を引き出すにはどのように使うべきかを検証した。まず特徴語でカテゴリーを生成する自動分類ツールについて検証し，IPCやFIを用いて分析集合の内容レベルを揃えると分類業務にとって有用なカテゴリーが多く生成されるとの知見を得た。次にタネ文献を利用した自動分類ツールについて検証し，ツールの分析条件と分類性能の関係性から導き出される効率的な使い方および外国語文献に対する分類性能について考察した。またテキストマイニング・マップを特許分類業務に用いる視点で検証した。

目　次

1．はじめに

技術動向調査，出願前調査，SDI[1)]，審査請求前調査，侵害調査，無効化資料調査など特許調査担当者の業務は多岐に及ぶ。この中でも特に，知財部門が企画部門や技術部門と，或いは調査会社が顧客企業と情報交換を繰り返しながら行う技術動向調査とSDIは，各企業が研究・開発の実態に即した独自の戦略的データベースを構築するために欠かせないものとなっている。そのため，調査担当者にはIPCやFIといった特許分類とは別に各企業の戦略に合わせた独自視点の分類による技術動向調査，SDIが求められる。さらにはグローバル企業においては，日本語文献のみならず外国語文献を対象とした調査も強く求められる。

技術動向調査は一般に調査対象の文献数が多く，またSDIはその頻度が高いため，その解析作業には多大な労力を要する。このような背景から，テキスト分析に基づいて特許文献を分類するツールに対して調査担当者の期待は大きい。期待されるテキスト分析ツールには，自動分類ツール，テキストマイニングツールなどがある。

しかしながら，期待の一方で，これらのテキ

* 2012年度　The First Subcommittee, Intellectual Property Information Search Committee

スト分析ツールから有用な情報を効率的に得る手法を確立できている調査担当者は少ないと思われる。

そこで当小委員会の第１ワーキンググループ（WG）では，技術動向調査やSDIにおいて行う特許分類業務を効率化するにはこれらのテキスト分析ツールをどのように使うべきか，特にツールの性能を引き出すにはどのように使うべきかを検証した。

２章では，特徴語でカテゴリーを生成する方式の自動分類ツールを題材にして検証した結果を述べる。具体的には，ツールに有用な分類を行わせるにはどのような分析集合を与えればよいかについて述べる。

３章では，タネ文献を利用する方式の自動分類ツールを題材にして検証した結果を述べる。具体的には，ツールの分析条件と分類性能の関係性，その関係性から導き出される効率的な使い方，および外国語文献に対する分類性能について述べる。

４章では，テキストマイニングツールを題材にして検証した結果を述べる。具体的には，審査引用情報とテキストマイニング・マップの関係について述べる。

２．特徴語でカテゴリーを生成する自動分類

自動分類ツールにおいては，単語およびその使用頻度を分析の基礎情報として用い，各文献中で使用頻度の高い単語をその文献を特徴づける特徴語であるとして重視する分析方法が主流である。またそれとともに，分析対象の集合において使用している文献が多い単語は“ありふれている”として重視しないよう調整されることが多い。

このような特徴語を利用する自動分類方式のひとつに，特徴語でカテゴリーを生成する方式がある。この方式では，集合中の各文献から抽出した特徴語を使用文献数等に基づきランク付けし，上位ランクの特徴語それぞれによるキーワード検索を行って集合を分類する。つまり，集合から“ほどよく高い使用頻度”の特徴語を選出して，その特徴語が記載された文献を集めることによりカテゴライズするのである。

当WGでは，特徴語でカテゴリーを生成する自動分類ツールの性能を引き出すためには，まず，特許分類業務にとって有用な特徴語を選出させる手法を確立することが必要と考え，どのようにすれば有用な特徴語が選出されるかを検証した。

なお２章における各検証は東芝ソリューション株式会社製のEiplaza/DAを用いて行った。

２．１　特許分類業務に有用な特徴語

検証に先立ち，調査対象の母集合を用意するとともに，各集合からベンチマークデータとして先願と後願の関係にある文献（先後願ペア）を抽出し，先後願ペアのテキストを目視確認して特許分類業務に有用な単語を選出した。

以下，詳細について説明する。

(1) 検証に用いた母集合

１章で述べたように，効率化したい業務のひとつは技術動向調査である。そこで，技術動向

表１　検証に用いた母集合

母集合名	検索式	文献数
LED照明集合（LED集合）	H01L33/＊（照明＋電球）＋F21＊LED	6,645
家庭用浄水器集合（浄水器集合）	C02F1/＊（家庭＋飲み水＋飲料）	9,047
履物集合	4F050＋4F051	9,739

調査を想定して，表１に示す5,000～10,000件規模の３集合を検証の題材とした。

また，各文献の分析テキストとして「発明の名称」「課題（要約中の課題記載部分）」「解決手段（要約中の解決手段記載部分）」「請求項（全請求項）」を用意した（パナソニック ソリューションテクノロジー株式会社製のPatent SQUAREで取得）。

(2) 先後願ペア

ベンチマークデータ，すなわち同一カテゴリーに分類されるべき文献として，特許法29条の２に基づき拒絶理由が通知された後願とその引例とされた先願のペア（先後願ペア）を各集合から抽出した。

先願と後願は，それぞれが実質同一とされた技術の記載を含みながらも，出願人が異なる。テキスト分析ツールでは，出願人による単語使用傾向の影響を少なからず受けるとされているが，先後願ペアに注目すれば，確実に出願人が異なるため単語使用傾向に対して一定の客観性を担保できると考えられるのである。

表１の母集合には38件の後願が含まれ（４章にて利用），そのうち先願と後願がともに母集合に含まれる18組20ペア（１つの後願に対して複数の先願が引用されている場合を含む。18組20ペアは後願18件と先願20件からなる）を２章の検証で使用した。18組20ペアの内訳は，LED集合から７組８ペア，浄水器集合から６組７ペア，履物集合から５組５ペアである。

(3) 主題語

各先後願ペアの分析テキストを目視確認し，分析テキスト中のどの単語が特許分類業務に有用であるかを当WGのメンバーで議論した。その結果，まず第一に母集合が発明の主題で分類されることが望ましく，抽出されるべき特徴語は“発明の主題を表す単語（主題語）”であるべきとの結論に至った。

例えば，図１のNo.3の浄水器の先後願ペアである特願平9-139032と特願平9-226070は，いずれも電解処理により水中のリンを除去する発明に関するものである。浄水器の母集合を分類するならば，第一に先後願ペアNo.3をその主題である「リン（の除去）」カテゴリーに分類したいと感じられた。なおその場合「電解」カテゴリーを作るとすれば「リン（の除去）」カテゴリーのサブカテゴリーとするのが妥当である。

このような考えの下，先後願ペア全18組のそれぞれに対して主題語を選出した（図２）。

2. 2 大規模集合から選出される特徴語

自動分類ツールのユーザーである調査担当者にとっては，大規模な母集合をそのままツールに入力するだけで分類が完了することが理想である。そこで，まず，表１のLED／浄水器／履物集合をそのままツールに入力して特徴語を選出させ，先後願ペアの主題語が選出されるか確認した。分析テキストには請求項を用いた。

(1) 結　果

図１に，浄水器集合から選出された特徴語を示す。同一の「No.」で括られている文献が先後願ペアである。横軸の「供給」～「陽極」は選出された特徴語のうち先後願ペアに含まれる

No	出願番号	共通特徴語数	供給	可能	浄水	配置	イオン	膜	流	容器	濾過	電気	分離	電極	ポンプ	電解	位置	タンク	端	塩	カートリッジ	陽極
1	特願平4-318471	5	○		○		○		○			○				○	○					
1	特願平6-54681		○		○		○	○	○					○		○				○		○
2	実願平6-5343	3			○			○			○				○						○	
2	特願平6-247538		○							○	○				○		○				○	
3	特願平9-139032	6				○						○		○		○				○		○
3	特願平9-226070					○	○		○			○		○		○				○		○
4	特願2000-28454	5 (8)		○			○	○			○											
4	特願2001-146274							○			○	○	○					○		○		
4	特願2001-1593			○			○	○	○			○	○	○		○		○	○	○		○
5	特願2001-217262	3		○			○			○							○		○			
5	特願2002-340614					○				○							○		○			
6	特願2008-141798	5		○							○				○			○	○			
6	特願2009-135608			○		○					○				○			○	○			

図１　共通特徴語（浄水器集合）

特徴語を抜粋したものである。

図中の○印は各文献の分析テキストである請求項が横軸の特徴語を含むことを示している。そのうち色付けした○印は先後願ペアに共通する特徴語（共通特徴語）であり,「共通特徴語数」はその数を示している。例えば，先後願ペアNo.3の共通特徴語は「配置」「電気」「電極」「電解」「塩」「陽極」の６つである。

図２に，浄水器集合から選出された特徴語と主題語の関係を示す。「全請求項」欄の○×印は先後願ペアに設定した「主題語」が当該ペアの共通特徴語として選出されたか否かを表しており，○印は選出されたことを示している。LED／履物集合の結果については図３を参照されたい。

No.	出願番号	主題語	全請求項	特徴語共通数
1	特願平4-318471	イオン	○	5
1	特願平6-54681			
2	実願平6-5343	濾過	○	3
2	特願平6-247538			
3	特願平9-139032	リン リン酸	×	6
3	特願平9-226070			
4	特願2000-28454	ミネラル	×	5 (8)
4	特願2001-146274			
4	特願2001-1593			
5	特願2001-217262	容器	○	3
5	特願2002-340614			
6	特願2008-141798	汚水 汚染	×	5
6	特願2009-135608			

図２　主題語と共通特徴語の関係（浄水器）

図２に示した通り，浄水器集合では６組中３組で主題語が選出されなかった。また，図３に示される通り，LED集合では７組中１組，履物集合では５組中５組全ての先後願ペアに対して主題語が選出されず，全体では半数の18組中９組で主題語が選出されなかった。

以上より，共通特徴語が多くとも必ずしもその中に主題語が含まれないこと，分野による差が大きいことが分かる。

このように5,000～10,000文献の大規模集合をそのまま分析しても，特許分類業務に有用な主題語が特徴語として選出されないことが多いことが確認された。

(2) 考　察

主題語が特徴語として選出されない要因のひとつとして集合の大きさが考えられる。つまり，選出されなかった主題語は，母集合が大きすぎたために“ほどよく高い使用頻度”の単語と認識されなかったと考えられる。

例えば，浄水器の先後願ペアNo.3については，母集合を構成する文献数に対して，請求項に「リン」または「リン酸」が記載されている文献数が少ないために「リン」や「リン酸」が“ほどよく高い使用頻度”の単語と認識されなかったと考えられる。

また，もうひとつの要因として集合を構成する文献の多様性が考えられる。つまり，集合を大きくすると技術的記載内容が多様となり，本来なら“ありふれている”とされるべき単語のランクが下がらずに“ほどよく高い使用頻度”の単語と認識される。そのために主題語がランク外になったと考えられる。

特許分類業務に有用な主題語が特徴語として選出されないことが多いと，自ずと自動分類の結果も有用なものとならない。

そこで，集合の大きさおよび内容の多様性が要因との仮説を立てて，この要因を取り除く簡便な工夫によって主題語を選出させる改善が可能か検証した。

2. 3　IPC／FI／Fタームによる一次整理

上述した要因を取り除くには，集合をほどよ

く小さくし，集合内の文献の内容レベルを揃えればよいと考えられる。

当WGでは，これを実現する方法として，母集合をIPC／FI／Fタームで一次整理することに可能性を見いだし，その効果を検証した。

IPC／FI／Fタームを利用すれば機械的に一次整理できるため効率的である。

(1) 検証方法

母集合を一次整理することによって先後願ペアの主題語が選出されやすくなるか否かを確認した。手順を以下に示す。

1）各先後願ペアに共通付与されたIPC／FI／Fタームを1つずつ当該ペアの一次整理用コードに設定する。

ただし共通付与されたコードが複数ある場合は，そのうち内容が主題語に近いコード，階層が深いコードを優先する。

2）各先後願ペアに対して当該ペアの一次整理用コードが付与された文献の集合（一次整理集合）を作成する。

3）各先後願ペアの一次整理集合から特徴語を抽出させ，当該ペアの主題語が共通特徴語として選出されたか確認する。

本検証では3通りの分析テキスト「解決手段」「請求項」「課題」で選出と確認を行った。

4）一次整理用コードの階層を1段上げてステップ1）～3）を実行する。

(2) 検証結果

図3に，一次整理の前後で特徴語の選出状況がどのように変化したかを示す。

図でNo.ごとにまとめられた出願番号は各先後願ペアである。「一次整理前」は，表1の母集合全体を分析テキスト「全請求項」について分析したときに各ペアの「主題語」が共通特徴語として選出されたか否かを表しており，○印は選出されたことを示している。「一次整理後」は，「一次整理コード」で整理した集合を分析テキスト「解決手段」「全請求項」「課題」について分析したときに各ペアの「主題語」が共通特徴語として選出されたか否かを示している。

母集合全体を分析したときに選出されなかった主題語が，IPC／FI／Fタームで一次整理した集合を分析したときは選出されるようになる改善傾向が見られた。但し，Fタームで一次整理した場合は，IPCやFIで一次整理した場合に比べて主題語がやや選出されにくい傾向がある。また，改悪事例は無かった。

図4に，一次整理用コードの階層を1段上げたときの影響を示す。

LED集合においては階層をなるべく下げると主題語が選出されやすくなり，履物集合においては階層をサブクラスにすると主題語が選出されやすくなった。

2. 4 効率的な使い方

上記結果より，特徴語でカテゴリーを生成する自動分類ツールを使うときは，大規模な母集合をそのまま自動分類させるのではなく，IPCまたはFIで小規模な集合に分けてから自動分類させるのが効率的であると結論づけられる。

さらに，このように使用すれば生成されたカテゴリーの特徴語が持つ意味をIPCやFIの説明と併せて読み取ることができる。そのため，同じ意味のカテゴリーをまとめるなど，分類結果に対する操作も躊躇なく実行でき，ツールを活用しやすくなるであろう。

一方で，IPCやFIの階層の違い，および分析テキストの違いに対し，選出される特徴語の変化はややセンシティブであるとの印象を受けた。ユーザーとしては，これらの違いに左右されにくいツール，或いはこれらの違いを簡単な操作で試せるツールの登場に期待したい。

	No.	出願番号	主題語	一次整理前		一次整理後					
				全請求項	共通特徴語数	一次整理コード			解決手段	全請求項	課 題
LED	6	特願2001-152783 特願2001-542388 特願平11-82580	青，緑，赤 （セット）	○	6 (8)	IPC	H01L 33/00	光の放出に適用される半導体	○	○	○
						FI	H01L 33/00,410	波長変換要素	○	○	○
						FT		FT該当なし	－	－	－
	12	特願2004-194509 特願2004-41502	吸収	×	6	IPC	C09K 11/08	無機発光性物質を含有	×	×	×
						FI	C09K 11/08J	蛍光体の混合又は組合せ	○	○	×
						FT	4D024 AA02	水道水，飲料用水	○	○	×
	18	特願2004-162279 特願2005-186322	入射	○	3	IPC	F21V 5/04	レンズ形状	○	○	×
						FI	F21V 5/04	レンズ形状	○	○	×
						FT	5F041 EE11	レンズとの結合，組み合わせ	○	○	×
	23	特願2004-332656 特願2006-84793	樹脂	○	2	IPC	H01L 33/00	光の放出に適用される半導体	○	○	○
						FI	H01L 33/00,400	半導体素子本体のパッケージ	○	○	○
						FT	5F041DA20	パッケージング-その他	○	○	○
	25	特願2005-130536 特願2006-279583	放熱	○	4	IPC	F21V 29/00	冷却または加熱手段	○	×	○
						FI	F21V 29/00A	放熱，遮熱	○	○	○
						FT	3K014 LA01	光源の種類-点	○	○	○
	27	特願2006-345206 特願2007-100646	放熱	○	2	IPC	F21Y 101/02	点状光源-小型のもの	○	○	○
						FI	F21S 2/00,100	モジュール式構造	○	○	○
						FT	3K243 MA01	点状光源を用いる照明装置	○	○	○
	32	特願2009-259312 特願2011-87828	蛍光	○	4	IPC		IPC該当なし	－	－	－
						FI		FI該当なし	－	－	－
						FT		FT該当なし	－	－	－
浄水器	1	特願平4-318471 特願平6-54681	イオン	○	5	IPC	C02F 1/46	電気化学的方法	○	○	○
						FI	C02F 1/46,A	電解水の製造	○	○	○
						FT	4D061 DB07	電気・磁気で水処理	○	○	○
	8	実願平6-5343 特願平6-247538	濾過	○	3	IPC	C02F 1/28	収着によるもの	○	○	○
						FI	C02F 1/28,G	簡易浄水器	○	○	○
						FT	4D024 AA02	水道水，飲料用水	○	○	○
	17	特願平9-139032 特願平9-226070	リン リン酸	×	6	IPC	C02F 1/58	特定溶存化合物除去	○	○	○
						FI	C02F 1/58,R	無機リン化合物	○	○	○
						FT	4D061 EB05	電・磁気水処理転換	×	×	○
	24	特願2000-28454 特願2001-146274 特願2001-1593	ミネラル	×	5 (8)	IPC	A23L 2/00	非アルコール性飲料	○	○	○
						FI	A23L 2/00,V	アルカリイオン水	○	○	○
						FT	4B017 LK02	非アルコール性飲料	○	○	○
	31	特願2001-217262 特願2002-340614	容器	○	3	IPC	C02F 1/68	飲料水改良の特定物質	○	○	○
						FI	C02F 1/68,510B	飲料水	○	○	○
						FT		FT該当なし	－	－	－
	38	特願2008-141798 特願2009-135608	汚水 汚染	×	5	IPC	B01D 29/00	加圧，吸引ろ過機	○	○	○
						FI	C02F 1/00,L	水,廃水,下水の処理	○	×	×
						FT	4D041 AA01	重力濾過機,下向流濾過	○	×	○
履物	3	特願平8-149814 特願平8-174898	スノーボード	×	8	IPC	A43B 5/04	スキー靴，それに類似したもの	○	○	○
						FI	A43B 5/00,310	その他の特定スポーツ用の靴	○	○	○
						FT	4F050 BC07	靴の甲被-踵部	×	×	×
	6	特願2001-86879 特願2002-564965	ジエン	×	5	IPC	C08K 3/00	無機配合成分の使用	○	○	×
						FI	C08L 15/00	ゴム誘導体の組成物	○	○	×
						FT	4J100 CA04	全体構造-2元共重合体	○	○	×
	7	特願2001-224411 特願平11-206381	開口 孔	×	2	IPC		IPC該当なし	－	－	－
						FI		FI該当なし	－	－	－
						FT	4F050 BA02	中底	×	×	×
	8	特願2003-297179 特願2004-188191	アウトソール・ ミッドソール （セット）	×	1	IPC	A43B 5/00	スポーツ用の履物	○	○	×
						FI	A43B5/00,303	ゴルフ靴	○	△	△
						FT	4F050BA10	底-スパイク	○	△	△
	9	特願2004-59071 特願2005-120737	釣	×	3	IPC	A43B 13/22	滑り止め底または耐摩耗の底	×	×	○
						FI	A43B13/22A	接地面の形状，構造	×	×	○
						FT	4F050HA27	材料-フェルト	○	○	○

図３　一次整理による特徴語選出の改善

3．タネ文献を利用する自動分類

次に，タネ文献を利用する自動分類ツールの検証について述べる。

タネ文献は予め用意する分類済みの文献であり，教師データなどと呼ばれる。自動分類ツー

	No.	出願番号	主題語	一次整理前	一次整理後				
				全請求項	一次整理コード		解決手段	全請求項	課　題
LED	1	特願2001-152783 特願2001-542388 特願平11-82580	青，緑，赤（セット）	○	FI	H01L 33/00,410 波長変換要素	○	○	○
					⇩ 階層を上げる				
					FI	H01L 33/00,400 パッケージに特徴のあるもの	○	○	△
	2	特願2004-194509 特願2004-41502	吸収	×	IPC	C09K 11/64 アルミニウムを含むもの	○	○	○
					⇩ 階層を上げる				
					IPC	C09K 11/08 無機発光性物質を含有するもの	×	×	×
	5	特願2005-130536 特願2006-279583	放熱	○	FI	F21V 29/00A 放熱，遮熱	○	○	○
					⇩ 階層を上げる				
					FI	F21V 29/00 冷却装置	×	×	×
履物	4	特願2003-297179 特願2004-188191	アウトソール・ミッドソール（セット）	×	FI	A43B 5/00,303 ゴルフ靴	○	△	△
					⇩ 階層を上げる				
					FI	A43B 5/00 スポーツ用の履き物	○	○	○
	5	特願2004-59071 特願2005-120737	釣	×	FI	A43B 13/22A 接地面の形状，構造	×	×	○
					⇩ 階層を上げる				
					FI	A43B 13/22 滑り止め底または耐摩耗の底	×	○	○

図４　IPC／FIの階層による影響

ルは，未分類の文献（テスト文献）の中からタネ文献に類似するテスト文献を選び出してタネ文献と同じ分類を自動付与する。

自動分類ツールへの期待は次のようなものである。例えば，各社が独自に分類してきた分類済み公報をタネ文献にしてSDIで得た新着公報をさらに自動分類すれば，技術部門へのきめ細かな情報提供が可能となる。また，技術動向調査において，文献群の一部を人手で分類した後に，それらをタネ文献にセットして残りの文献群を自動分類すれば，知財部門から企画部門に，或いは調査会社から顧客企業にスピーディーなレスポンスが可能となる。さらには，外国語文献をも自動分類できれば，有意な文献の解析に注力できて業務効率が大幅に効率化される。

3.1節では日本語文献について特にその分類精度について検証する。3.2節では同様の検証を外国語文献（中国語文献）で行い，日本語文献の検証結果との差異を中心に考察する。

なお３章における各検証は２章に引き続きEiplaza/DAを用いて行った。

3. 1　日本語文献の自動分類

まず，日本語文献の自動分類精度について検証した。

(1) 検証方法

検証は，自動分類によるFIの再現性を評価することで行った。FIが再現できれば各社の独自分類も再現できると思われる。Fタームではなくとしたのは，２章で述べたように発明の主題で分類された方が有用だと感じたことが理由である。IPCではなくFIとしたのは次節で述べる中国語文献を細分類できる可能性について知

見を得たいことが理由である。

図５の例を参照しつつ検証手順を説明する。

１）母集合からタネ文献用集合とテスト文献集合を作る。

SDIを想定し，出願年1993～2003年のタネ文献用集合と出願年2004～2011年のテスト文献集合を作成した。

２）タネ文献用集合の各文献を付与されているFIごとに分けて，それぞれを当該FIのタネ文献としてセットする。

各母集合から付与数の多いグループを1つずつ選び，その下位グループのFIと分冊識別記号を使用した（表２）。図５の例ではH01L33/00,184に176件のタネ文献がセットされている。

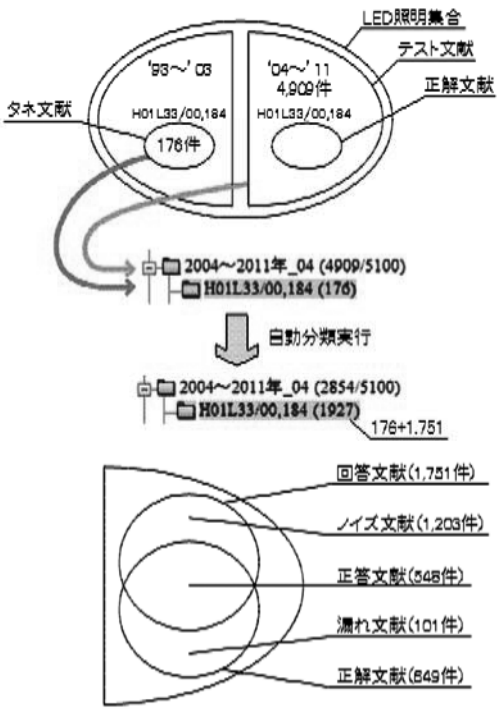

図５　SDI検証の方法

表２　検証に使用したFI

集合	FI範囲（FI数※）	テスト文献数
LED	H01L33/00以下　（35）	4,909
浄水器	C02F1/28以下　（20）	2,536
履物	A43B13/00以下　（50）	3,200

※タネ文献またはテスト文献が０件のFIは検証から除外してある

３）テスト文献集合を自動分類させる。

自動分類されたテスト文献を回答文献と呼ぶことにする。図５の例では4,909件のテスト文献のうち1,751件の回答文献がH01L33/00,184に分類され，ツール上ではタネ文献と合わせて1,927件の部分集合ができている。またタネ文献のFIと同じFIが付与されているテスト文献を正解文献と呼ぶことにする。

４）再現率とノイズ率を算出する。

FIごとに回答文献と正解文献を突き合わせ，テスト文献を，付与通りに自動分類された正答文献，付与があるにもかかわらず自動分類されなかった漏れ文献，付与がないのに自動分類されたノイズ文献に仕分け，再現率とノイズ率を算出する。

再現率（％）

＝正答文献数÷正解文献数×100

ノイズ率（％）

＝ノイズ文献数÷回答文献数×100

図５の例では，再現率84.4％，ノイズ率68.7％となる。

５）分析テキスト，分析パラメータを変更してステップ１）～４）を繰り返す。

分析テキストは「発明の名称＋要約」「発明の名称＋要約＋請求項」の２通りとした。分析パラメータについては類似度に対するしきい値の設定を６段階で変更した。

(2) 検証結果

図６，７にタネ文献数と再現率の関係を，図８にタネ文献数とノイズ率の関係をそれぞれ示す。図７は図６の横軸を一部拡大したものである。図６～８ではLED／浄水器／履物集合の結果が区別できるように，それぞれ異なる記号で示している。なお，後述する結果を踏まえて分析テキストには請求項を含めている（発明の名称＋要約＋請求項）。また，分析パラメータはツールの標準値（類似度しきい値0.15）としている。

この結果から，タネ文献数が少ないと再現率の振れ幅が大きく，統計的には再現率が不定な状態になることが分かる。一方，タネ文献数が充分な数になると再現率の振れ幅は小さくなり，高めの再現率で安定した状態になる。今回利用した自動分類ツールでは，タネ文献数が100件を超えると再現率が安定した。

ノイズ率はタネ文献数を増やすほど減少する傾向にあるが，その減少の速さには技術分野間で差が開いた。図８の結果からは，LED集合においてノイズ率の減少が最も速く良好であり，浄水器／履物集合においてLED集合と同等に低いノイズ率を達成するには２倍以上のタネ文献数が必要と推定される。

以上から分かるように，自動分類の精度を上げるにはタネ文献を多く与えることが重要である。

図９に，分析テキストを変えてLED／浄水器／履物集合のそれぞれを自動分類したときの再現率を３分野の平均値と併せて示す。

なお，図９の各値はタネ文献数100件以上を確保できたFIでの平均値を示している。また分析パラメータにはツールの標準値を用いた。

図９によれば，再現率は，いずれの技術分野でも分析テキストに請求項を含めたときの方が高いことが分かる。３分野平均で約９ポイントの差がついた。一方，技術分野間での差は分析テキストでの差よりも大きい。請求項を含ませたときの数値で比較すると，LED集合での再現率81％に対して履物集合での再現率は57％であり，その差は24ポイントに及んだ。

また図10に，分析テキストを変えたときのノイズ率を示す。ノイズ率に関しては，分析テキストの違いによる有意な差は見られなかった。

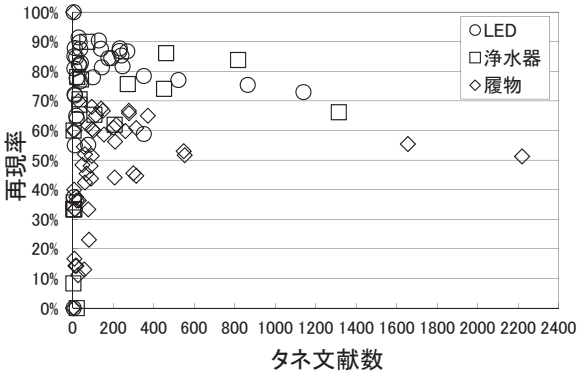

図６　タネ文献数と再現率の関係（１）

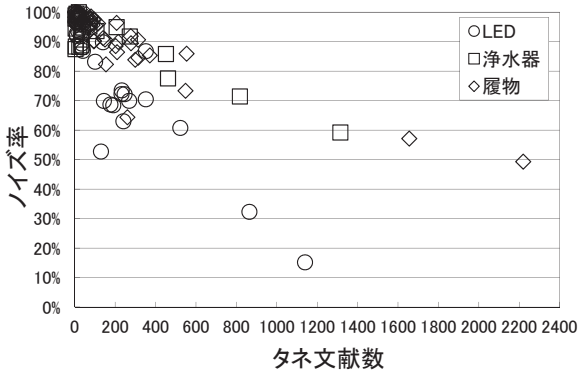

図８　タネ文献数とノイズ率の関係

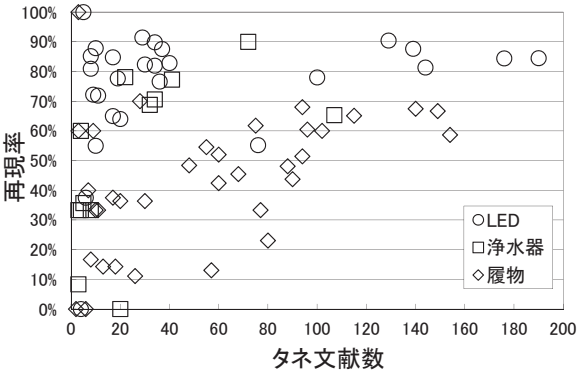

図７　タネ文献数と再現率の関係（２）

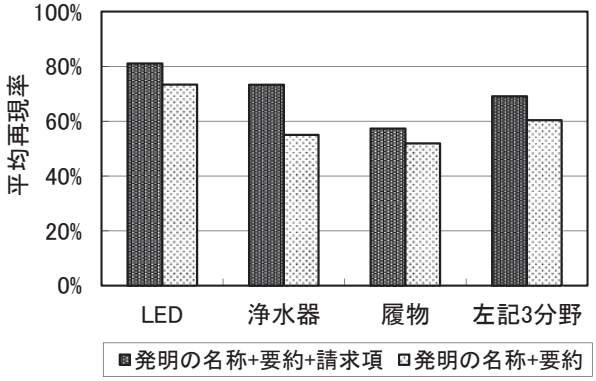

図９　技術分野，分析テキストと再現率の関係

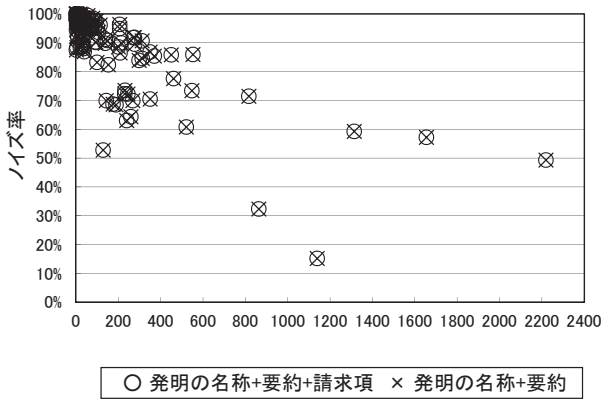

図10　分析テキストとノイズ率の関係

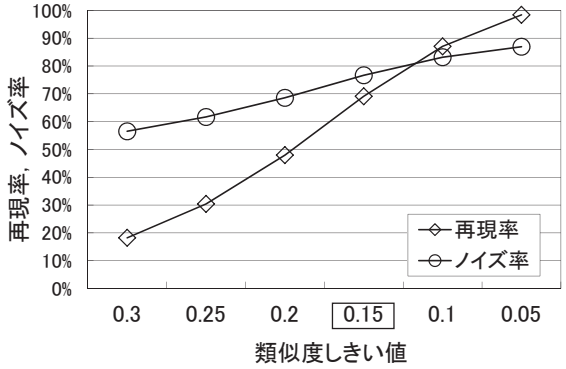

図11　類似度しきい値と再現率，ノイズ率の関係

総合的にみて，分析テキストを発明の名称と要約のみとするよりも請求項を含ませた方が自動分類精度は向上すると言える。また，技術分野による自動分類精度の差が大きい。

図11に，ツールの分析パラメータである類似度しきい値と再現率，ノイズ率の関係を示す。図11の値は，分析テキストに請求項を含ませ，タネ文献数100件以上を確保できたFIでの３分野平均値である。

グラフの左側では類似度に対する判定が厳しくなる。そのため再現率，ノイズ率ともに低くなっている。逆にグラフの右側ほど判定は甘くなり，再現率，ノイズ率ともに上昇する。

このように再現率とノイズ率はトレードオフの関係にあり，正答文献を多く得ようとすればノイズ文献も多くなる。

3. 2　効率的な使い方

実務では，自動分類を実行して得た回答文献が正答文献かノイズ文献かを人が読んで確認しなくてはならず，ノイズ文献を読む回数を減らすことが効率化に直結する。

ここで再度図11に注目する。類似度しきい値0.05の設定は，再現率がほぼ100％であり，しかも再現率がノイズ率を上回っていることから一見すると最善の設定に見える。しかし，これを文献数で捉えると視界は一変する。具体例としてH01L33/00,184を自動分類したときの正答文献数，ノイズ文献数，回答文献数の関係を図12に示す。

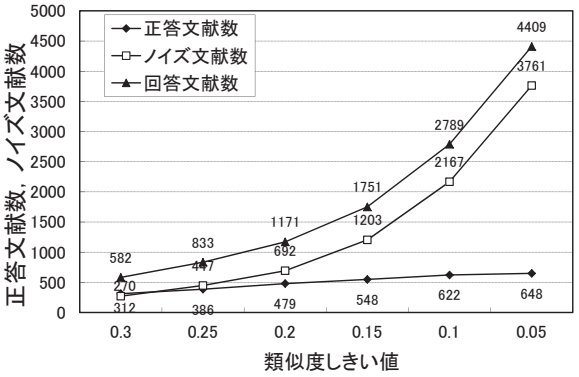

図12　H01L33/00,184を自動分類したときの数値例

図12の例で類似度しきい値を0.05に設定すると（正解文献数は649件であるから漏れ文献数はたったの１件ではあるが）実に3,761件のノイズ文献を読まなくてはならなくなるのである。

他方，類似度しきい値を0.3に設定すれば，正答文献数は半分以下となるが，読むべきノイズ文献数は14分の１以下となる。この場合，読んで精査した正答文献をタネ文献に加えて再度自動分類を実行すれば，自動分類の精度は徐々に高まり，作業は効率的に進捗していくであろう。さらには，精査により確定したノイズ文献をノイズのタネ文献とし，再度の自動分類において正答文献のタネ文献とノイズのタネ文献に

テスト文献の引っ張り合いをさせることで精度向上を図ることも考えられる。

なお，テスト文献数が少ない場合は，類似度しきい値を甘めに設定して一気に正答文献を集めた方が効率的である。

以上から実務においてタネ文献を利用する場合，テスト文献数が多いときは分析パラメータを厳しめに設定してノイズ文献数を減らし，テスト文献数が少ないときは分析パラメータを甘めに設定して正答文献数を増やすことが効率化につながると考えられる。

なお，実務を考えた場合，安定した結果を得るためにタネ文献数100件を要することはやや多いと感じる。ベンダーにはこの件数を減らす改善に期待する。例えば，予めIPCなどの特許分類ごとに“ありふれた単語”を分析しておき，この分析結果をタネ文献に付与されているIPCに応じて適用することで，タネ文献の多寡によらず安定した結果を得ることができるのではなかろうか。

3. 3　外国語文献の自動分類

次に，外国語文献の自動分類精度について検証する。

外国語文献は，言語の壁があるため一般に日本語文献よりも特許分類業務にかかる負荷が大きい。そのため，外国語文献を精度よく自動分類することができ，内容を精査する件数を減らすことができれば，業務効率の向上に大きく寄与することになると考える。

当WGでは，注目度が高い言語のひとつである中国語の文献を題材にして検証を行った。

(1) 検証方法

検証は，中国語文献に対して日本語文献と同様にFIの再現率とノイズ率を算出し，中国語文献での結果を日本語文献での結果と比較することで行った。

中国語文献には元々FIが付与されていないため，ファミリーの関係（INPADOC Family）にある日本語文献からの転記によって中国語文献にFIを擬似付与することで検証用データを作成した。そのため，検証用データは必然的に対応日本特許が存在するもののみに限定され，3.1節の検証と比較して検証用データは少なくなっている。

なお，検証に用いた中国語文献の分析テキストは「発明の名称＋要約＋第一請求項（メインクレーム）」である（株式会社 発明通信社製のHYPAT-iで取得）。日本語文献の分析テキストは「発明の名称＋要約＋請求項（全請求項）」とした。また，分析パラメータはツールの標準値とした。

(2) 結　果

図13に中国語文献でのタネ文献数と再現率の関係，図14にこれと対応する日本語文献でのタネ文献数と再現率の関係をそれぞれ示す。

また，図15に中国語文献でのタネ文献数とノイズ率の関係，図16にこれと対応する日本語文献でのタネ文献数とノイズ率の関係をそれぞれ示す。

日中を比較すると，再現率は中国語文献の方が若干高く，ノイズ率は日本語文献の方がタネ文献数に対する低下が若干早い。

しかしながら，日中で決定的な差はないと考

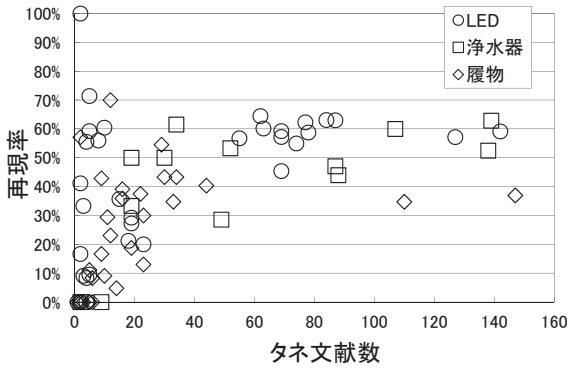

図13　中国語文献でのタネ文献数と再現率

えられる。単語抽出などの前処理が日中ともに同レベルで行われているとすれば，後段の自動分類ロジックに対する言語の違いの影響は少ないと考えられる。このことからツールの多言語対応に期待を感じた。

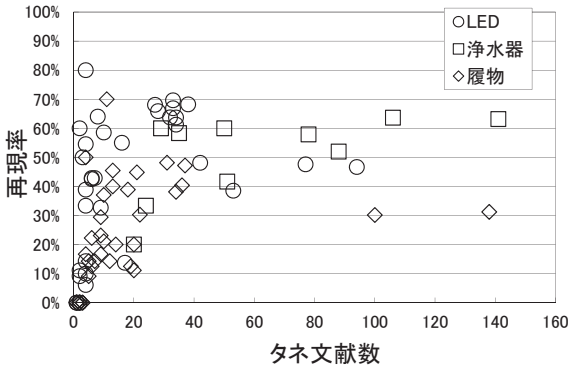

図14　日本語文献でのタネ文献数と再現率

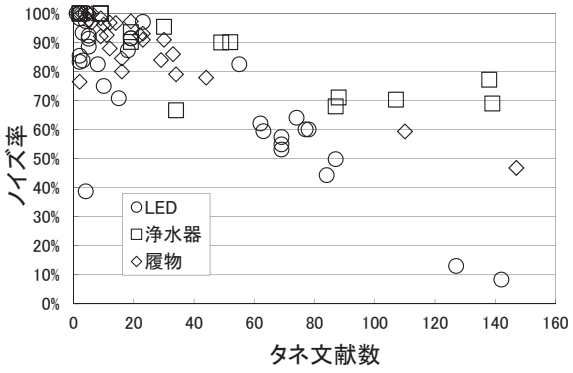

図15　中国語文献でのタネ文献数とノイズ率

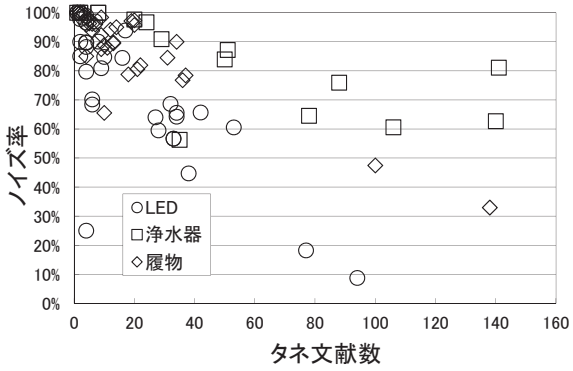

図16　日本語文献でのタネ文献数とノイズ率

4．テキストマイニング・マップを用いた分類

続いて，テキストマイニングツールを特許分類業務に活用する視点で検証する。

テキストマイニングツールは，比較的文献数の多い集合を分析できる特質を活かして，文献集合を俯瞰するツール或いはユーザーに気づきを与えるツールとして提供されることが多い。そのため，テキストマイニングツールの多くは文献の類似関係をマップ状にプロットする機能を備えている。このマップ状のプロット結果をテキストマイニング・マップと呼ぶことにする。

上述したように，テキストマイニングツールは必ずしも分類に特化したものではない。また，テキストマイニング・マップでは多次元の文献情報を２次元に縮退させているため，マップ上での文献間の距離は必ずしも正しいとは言えないとされている。さらには，テキストマイニングツールの分析結果は出願人による単語使用傾向の影響を少なからず受けるとされている。

とはいえ，少なくとも類似関係にある文献はテキストマイニング・マップ上で近接しているのであるから，マップ上で適度な枠内に収まっている文献群を取り出せば特許分類業務に活用できると思われる。

そこで４章では，前出した先後願ペアを利用することで，テキストマイニング・マップから類似文献群を取り出す範囲の目安を得ることができるか否かを検証した。

なお４章における検証は，株式会社アモティ製のPAT-ReSergeを用いて行った。

(1) 検証方法

本検証で使用したテキストマイニングツールはクラスタリング方式を採用しており，一定以上の類似性を有する文献をクラスタとしてまとめ，さらにクラスタ間の類似性を分析する。こ

のツールが提供するテキストマイニング・マップでは，各クラスタを表す円がクラスタ間の類似性を表す配置でプロットされる。

本検証では，第一にテキストマイニング・マップにおいて先後願ペアがどの程度同一クラスタ内にプロットされるか検証し，第二に同一クラスタ内にプロットされない先後願ペアはどのような位置関係になるかを検証した。

１）LED／浄水器／履物集合の分野それぞれについて複数の先後願ペアを抽出する。

本検証では，ステップ２）の条件調整を勘案しつつ，2.1節（2）項で述べた後願38件のいずれかを有する21ペアを抽出した。内訳はLED集合から８ペア，浄水器集合から９ペア，履物集合から４ペアである。

２）先後願ペアを包含する集合を作成する。

本検証では，作業の容易化のために，出願人・出願日・IPC等の条件を調整して各集合を500件以下の規模とした。また，各集合ができるだけ多くの先後願ペアを含むよう勘案して上記条件調整を行った。

３）各集合に対してテキストマイニング・マップを作成させ，マップ上に先後願ペアをプロットする。なお，検証に用いたツールの分析テキストは明細書全文である。

(2) 検証結果

同一クラスタに入った先後願ペアは，LED集合で８ペア中２ペア，浄水器集合で９ペア中１ペア，履物集合で４ペア中１ペアであった。平均すると先後願ペアの約２割が同一クラスタに入り，約８割が異なるクラスタに分かれたことになる。

図17に履物集合の例を示す。濃い円が先後願ペアの一方または両方を含んだクラスタ，淡い円がその他のクラスタである。ペアＡは同一クラスタに入り，ペアＢ～Ｄは異なるクラスタに分かれた。但し，異なったクラスタに分かれた先後願ペアであっても，比較的近くにプロットされた。具体的には，図17の例のように49ブロックに分割されている画面上で，４ブロック内のクラスタにプロットされた。

以上のように，テキストマイニング・マップにおいては，類似文献は必ずしも同一クラスタに入らないが，比較的近くにプロットされることが確認できた。

従って，マップから類似文献群を取り出す際には，周辺の複数クラスタをまとめて取り出す必要がある。そして，先後願ペアのプロットはマップから類似文献群を取り出す範囲の目安になると思われる。すなわち，先後願ペアは出願人の単語使用傾向に対して一定の客観性を担保できるため，テキストマイニング・マップを利用する上で先後願ペアのプロットから得られる距離感は有用な情報と考えられるのである。

なお，図17の例から分かるようにペア間の距離は一定ではないため，複数の先後願ペアをプロットして範囲の大きさを吟味すべきである。

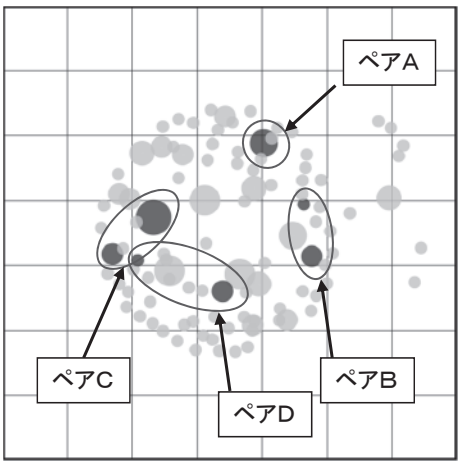

図17　７年間に出願された出願人９社での履物集合164件に含まれる先後願ペアの例

一例として先後願ペアＣ付近にプロットされたクラスタの内訳を図18に示す。

図中で実線の円がクラスタを表している。ペ

アＣの先願，後願，ペアＤの先願が入ったクラスタをそれぞれクラスタＣ１，Ｃ２，Ｄ１とし，クラスタＣ１とＣ２の間にプロットされた２つのクラスタをそれぞれクラスタＸ１，Ｘ２としている。

また，引き出し線で示した文字は目視確認した各文献の主題である。同一クラスタ内で主題が共通している文献群を点線の楕円で囲っている。

先後願ペアＣの間にプロットされたクラスタＸ１とＸ２のうち，クラスタＸ１の文献は先後願ペアＣと同じ主題“アウトソールとミッドソールの接合”に関するものであった。類似文献を取り出すには付近のクラスタも確認する必要があることが分かる。

一方，出願人に注目すると，クラスタＣ１は全てＢ社，クラスタＣ２は全てＳ社，クラスタＸ１はＣ社であった（不図示のクラスタには複数出願人が混在するクラスタもある）。同じ主題の文献が異なるクラスタに分かれる要因のひとつが出願人による単語使用傾向であることは否めない。

このことから，マップから類似文献群を取り出す際には，出願人のバリエーションを確保するようにして周辺の複数クラスタを取り出すべきであると考えられる。例えばクラスタＣ１のＢ社先願と類似する文献を集める場合，直近クラスタＸ１にてＣ社文献を見つけたとしても，さらに範囲を広げてＡ社文献が入ったクラスタＸ２とＤ１，Ｓ社文献が入ったクラスタＣ２を確認すべきである。

以上より，テキストマイニング・マップを利用して類似文献を集める際の確認範囲に関し，分析集合に含まれる先後願ペアのプロットが確認範囲の目安となり得るとの知見，および出願人のバリエーションを確保できる範囲に適宜広げるべきであるとの知見が得られた。

なお，テキストマイニングツールのユーザーが期待するのはあくまでも類似文献が同一クラスタに集まっているマップである。今後もツールの分析精度向上に期待したい。

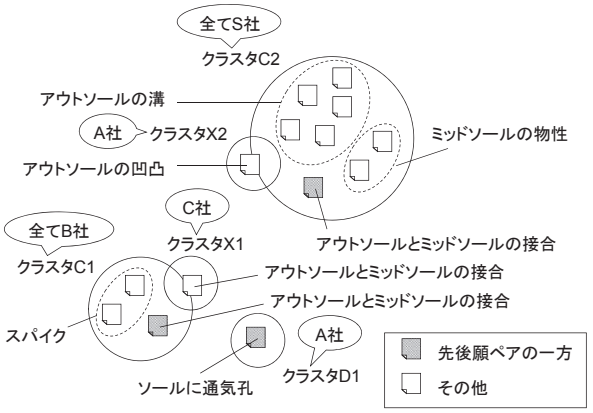

図18　クラスタの内訳

5．ま と め

以上，当WGがテキスト分析ツールについて行った検証を２～４章で紹介した。

２章では，特徴語でカテゴリーを生成する自動分類ツールに関し，特許分類業務にとって有用な特徴語を選出させる使い方を検証した。この検証を通じ，特許分類業務に有用な特徴語を選出させるには分析する集合の内容を揃えると良いことが分かった。そして，このような集合をIPC／FI単位で用意すれば効率的であることを述べた。

３章では，タネ文献を利用した自動分類について性能を引き出すための使い方，日本語文献と中国語文献に対する性能差を検証した。この検証を通じ，タネ文献とテキスト量を多く与え，文献数および再現率とノイズ率のトレードオフを考慮して分析パラメータを選ぶことにより性能を引き出せることが分かった。また，中国語文献に対しても日本語文献と概ね同等の精度で自動分類できることを確認した。

４章では，テキストマイニング・マップを特

許分類業務に活用する視点で検証した。この検証を通じ，先後願ペアをマップ上にプロットすることが分類作業の一助となり得るとの知見，出願人のバリエーションに注意して確認範囲を適宜設定すべきとの知見が得られた。

6．おわりに

当検証を通じ，使い方次第でテキスト分析ツールから得られる情報が随分変わる印象を受けた。

特許文献のテキスト情報は増加する一方であるため，今後もテキスト分析ツールによる業務効率化への期待はますます高くなるであろう。

調査担当者は，ツールの特質を理解してその性能を引き出す使い方を導き出すことで業務への組み込み方を的確に判断できると思われる。

本稿がその一助となれば幸いである。

なお，本稿は2012年度知的財産情報検索委員会第１小委員会第１ワーキンググループのメンバーである高井史比古（セコム，小委員長），市川敬子（中外製薬），白石達弥（三菱重工業），竹内旭（ニプロ），武島正治（積水化学工業），松原誉実（アイピックス），毛利克輝（ルネサスエレクトロニクス）が担当した。

注　記

1）SDIとは，Selective Dissemination of Information の略。
利用者に応じて選択的に情報を提供すること。典型的には，利用者ごとに検索式と配信先を予め関連付けて登録しておき，データベースに追加・更新された情報の中から検索式に適合した情報を配信先に配信すること。新着情報を継続的・定期的に調査したいときなどに利用される。

（原稿受領日　2013年６月11日）