特許情報解析システム（第一報）：
－フリーウェア termmi の応用－

○加藤　亮[1)]，橋本博之[1)]，辻河　登[1)]

Ryo KATO
Search
(Medicine, Biotechnology)
BA in Agrobiological Resources at University of Tsukuba.

TextMining of Patent Information
-First Report-
Application of Freeware “termmi”.

KATO Ryo [1)], HASHIMOTO Hiroyuki [1)], TSUJIKAWA Noboru [1)]
WISEL corporation [1)]
TORANOMON 30 MORI Bldg. 2-2 Toranomon
3-Chome, Minato-ku, Tokyo 105-0001 Japan
Phone: +81-3-3431-1110　Fax: +81-3-3431-1233/+81-3-3431-1221
E-mail: office@wisel.co.jp

【発表概要】
テキストマイニング専用のフリーウェア termmi を活用した特許情報解析の方法を検討した。termmi は複数の文書から用語を抽出する機能およびベクトル空間法による文書の類似度算出機能を持つ。特許を対象とした termmi の使用経験に関する報告は見られないが、今回の検討の結果、特許情報に対しても優れた解析機能を有していることが分かった。また、概念検索のツールとして利用できることも分かった。使用方法の知見の集積により、termmi の利用範囲は拡がると考える。なお、termmi の解析結果の視覚化についても併せて検討した。

【キーワード】
フリーウェア，テキストマイニング，　茶筅，termmi，ベクトル空間法，クラスタリング、特許、情報解析、視覚化、概念検索

１．はじめに

特許情報は研究開発の動向を把握し、市場の将来予測を行う上で有用な情報源である。そのため、特許情報を効率的に解析する工夫が続けられており、数年前から商用の解析システムも発売されるようになった。これらの各システムには種々の優れた特徴が

あり、価格も数百万～数十万円と幅が大きく、ユーザは目的に合わせて利用している。

一方、フリーウェアの Windows 用テキストマイニングツール「termmi」が Web 上で紹介されている[1)]。このシステムに対する関心は高く、関連ソフトも含めると 100～200 件／月の頻度でダウンロードされている[2)]。しかし、このシステムを特許に応用した報告は未だ見られないので、我々は特許への termmi の有用性を検討した。その結果、運用知識の蓄積と周辺の整備により、情報解析ツールとして利用できると判断したので報告する。

２．方法

2.1. termmi の説明

１）機能

本システムの概要について、次のような紹介が行なわれている。[1)]

(a)東京大学と横浜国立大学により共同開発されたシステムである。

(b)用語に関する複数ファイル間での重要度の数値比較を行う。

(c)システムの実行により、次の４種類の用語抽出結果をファイルとして出力する。（ ）内はファイル名を表す。

・各文書に対する用語
・各文書に固有の用語
・文書群に共通の用語（common.txt）
・文書群全体の用語（total.txt）

(d)これらの結果の比較検討により、他の論文との差異を見出す。

(e)ベクトル空間法により文書の類似度計算が行われる。

２）オペレーション

非常に簡便であり、必要な操作は次の２工程である。

(a)分析対象の文書が入ったフォルダを"termmi"のアイコンにドラッグ。直ちに個々の文書および文書群全体から用語抽出が始まる。

(b)Perl スクリプトのアイコンをダブルクリック。ベクトル空間法による文書の類似度計算が始まる。

2.2. termmi の基本的機能の検証

１）検討に使用した特許

「調光遮熱[3)]」に関する表 2.1 の特許 11 件を素材とした。技術的な内訳は電圧駆動型が７件、サーモクロミック型が４件であった。

表 2.1　termmi の基本機能検討用素材

公報番号	発明の名称	技術内容
特開2005－250119	調光材料およびこれを用いた車両	電圧駆動型
特開2005－82472	透光性積層膜、光透過性基材およびそれらの透過光制御方法	電圧駆動型
特開2005－60703	電気光学的液晶システム	電圧駆動型
特開平7－318983	電極として低輻射率被膜を持つライトバルブ	電圧駆動型
特開平5－45679	調光装置	電圧駆動型
再表03/057799	調光素子およびその製造方法	電圧駆動型
特開平5－25479	調光素子	電圧駆動型
特開平11－265005	積層体およびそれを使用した窓	サーモクロミック
特開平7－242447	自律応答積層体、その製法およびそれを使用した窓	サーモクロミック
特開平7－171926	複合複層体及びそれを使用した窓	サーモクロミック
特開平7－171925	積層体及びそれを使用した窓	サーモクロミック

２）各クラスターの特徴と視覚化

termmi ではベクトル空間法により文書間の類似度計算を行うが、類似する文書のクラスター化と各クラスターの視覚化は未だ行われていない。そのため、クラスターを判別し、視覚化する方法も併せて検討した。

(a)クラスターの始点と終点

類似度順に編集された termmi の処理結果を活用して、隣接する特許間での共通語の分布状況を調べた。そして、この要因を利用して各クラスターの

始点と終点の判別の可否を検討した。共通語の調査には、termmi の common.txt 作成機能を応用した。

(b)各クラスターに特有の用語

電圧駆動型とサーモクロミック型の各クラスターに特有の用語の順位（重要度）について、各クラスター内での順位と技術全体（total.txt）での順位との相関を調べ、クラスター設定への応用の可否を検討した。解析には Excel の回帰分析を使用した。順位の幅を 1-50、51-100、500-1000、1-1000 と変えることによる影響も調べた。

2.3. termmi の特許情報への応用

termmi での処理件数を上記 2.2 よりも多い 50 件、100 件とし、termmi での処理結果を検証した。技術分野は上記 2.2 と同じ調光遮熱とした。

１）特許 50 件に対する応用

次の検索によりヒットした特許 50 件を処理し、調査主題に該当する特許の類似度の状態を調べた。

システム　IPDL
資料　公開特許
検索項目　要約+請求項
検索期間　2004.1.1－2004.12.31
検索式　調光 and ガラス

また、処理の結果、主題に該当する特許のクラスターが形成されていないときは、クラスターの形成に必要な用語の調整法を検討した。

２）特許 100 件に対する応用

処理対象の母集団の件数を 100 件とするとともに、多種類の調光遮熱技術が混在する集合を構成した。技術別のクラスターが termmi 処理後に形成されていないときは、上記 2.3-1)と同様に用語の調整法を検討した。

表 2.2. 調光遮熱関連特許 100 件

平01- 38732	平05-307172	平09-221343	2001-310407
平01- 57242	平05-307171	平09-228763	平15-510205
平01-126629	平07-138048	平09-248874	2003- 94551
平01-138541	平07-157339	平09-256752	2003-121884
平03- 43714	平07-171926	平11- 6988	2003-140196
平03-141138	平07-171925	平11- 38455	2003-190710
平03-229218	平07-199780	平11- 38408	2003-195364
平03-266814	平07-232938	平11-131629	2003-261356
平03-276127	平07-242447	平11-157880	2003-266578
平05- 8341	平07-246366	平11-157879	2003-266577
平05- 19306	平07-274738	平11-241161	2003-267754
平05- 25479	平07-290841	平11-265005	2003-267755
平05- 25478	平07-315883	平11-265006	2003-335553
平05- 27270	平07-318983	平11-316393	2005- 31302
平05- 27271	平07-324439	平11-316394	2005- 60703
平05- 45679	平07-330336	平11-315146	2005- 62749
平05- 80310	平07-331430	WO97/041329	2005- 82472
平05- 80309	平09- 29882	2001- 19908	2005- 89244
平05- 80308	平09- 71440	2001- 75132	WO03/057799
平05-181403	平09- 80359	2001- 83554	2005-126582
平05-181401	平09-124347	2001-125151	2005-126581
平05-181402	平09-124348	2001-191441	2005-187631
平05-188353	平09-127559	2001-215456	2005-199683
平05-193040	平09-169549	2001-240434	2005-208411
平05-209022	平09-194235	2001-262144	2005-250119

3．結果

3.1. termmi によるクラスタリング

termmi での処理を次のケース１～５について行なった。

１）total.txt 未調整での termmi 処理

(a) ケース１：電圧駆動型７件とサーモクロミック型４件の計 11 件の特許に termmi 処理を行った結果を図 3.1 に示した。１件を除き、総じて各技術はクラスター化されていた。

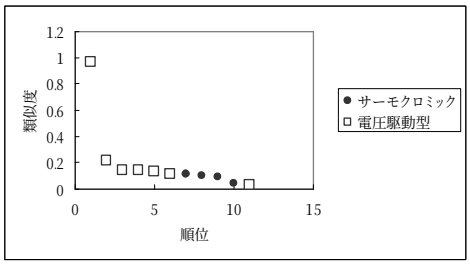


図 3.1 クラスター２種、特許 11 件の解析

(b) ケース２：IPDL から得た特許 50 件に termmi 処理を行ったが、主題に該当する特許 21 件はクラスターを形成していなかった。

(c)ケース３：termmi 処理を多種類の調光遮熱技術が混在する特許 100 件

に対して行った。着目した電圧駆動型とサーモクロミック型はともに明確なクラスターは形成していなかったが、サーモクロミックは順位 60-100 に多く、電圧駆動型は順位 50 以下に多いという傾向は見られた。

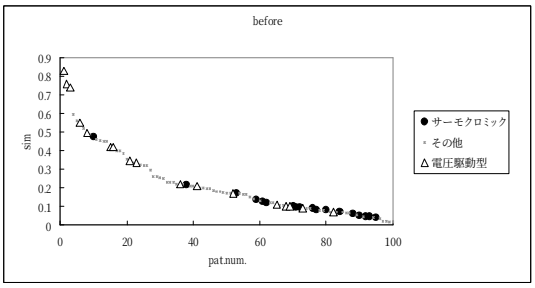


図 3.2. 特許 100 件の処理結果（調整前）

2）total.txt 調整後の termmi 処理

クラスターを形成しなかった上記のケース 2 と 3 について、クラスターを形成させる方法を検討した。total.txt から複数の調光遮熱技術に共通している重要語を中心に約 5000 語を削除した後の 16000 語を使って、再度、ケース 2 とケース 3 を処理した。

(a)ケース４：ケース２で主題に該当する特許 21 件（●）は表 3.1 に示すようにクラスターを形成し、類似度上位 27 位までに総て含まれていた。なお、ノイズは光源、撮像機等々に関する技術であり、調査主題と異なる技術が多かった。

表 3.1. 特許 50 件の termmi 処理結果

1	0.4424015572	JP2004138795	●	26	0.0062777046	JP2004363012	×
2	0.3523685254	JP2004151575	●	27	0.0058462116	JP2004114900	●
3	0.3100414652	JP2004255002	●	28	0.0051130172	JP2004108887	×
4	0.2704321692	JP2004165113	●	29	0.0050242646	JP2004165129	×
5	0.2293929399	JP2004189581	●	30	0.0048908245	JP2004325562	×
6	0.1794965718	JP2004139134	●	31	0.0048673615	JP2004219990	×
7	0.1526682560	JP20049387 3	●	32	0.0048168278	JP2004271830	×
8	0.1435928165	JP20043135	●	33	0.0047156922	JP2004252137	×
9	0.1424474397	JP200424283	×	34	0.0043682410	JP2004102105	×
10	0.1289074853	JP2004306905	●	35	0.0042209685	JP2004272096	×
11	0.1135425152	JP20043134	●	36	0.0041724954	JP2004519746W	×
12	0.0905832780	JP2004333567	×	37	0.0040564910	JP2004303573	×
13	0.0850178405	JP2004131335	●	38	0.0038574915	JP2004519745W	×
14	0.0832657524	JP200469978	●	39	0.0023527911	JP200431098	×
15	0.0821581587	JP2004182484	●	40	0.0018233490	JP2004309543	×
16	0.0675561587	JP2004302192	●	41	0.0013886250	JP2004127539	×
17	0.0623459513	JP2004109543	×	42	0.0012808411	JP2004327274	×
18	0.0583729609	JP200424534	●	43	0.0006155872	JP2004318853	×
19	0.0559743064	JP2004325497	●	44	0.0006121185	JP2004507872W	×
20	0.0556053378	JP2004123011	●	45	0.0006032203	JP2004299591	×
21	0.0410023646	JP2004150201	●	46	0.0006029439	JP2004314860	×
22	0.0248390688	JP200479221	×	47	0.0005834175	JP2004537053W	×
23	0.0157452814	JP20044795	●	48	0.0003365571	JP200493653	×
24	0.0089425224	JP2004175622	●	49	0.0000275871	JP2004288645	×
25	0.0083629754	JP2004363421	×	50	0.0000247920	JP2004311449	×

(b)ケース５：ケース３の母集団についての処理結果を図 3.3 に示した。ケース３でみられた電圧駆動型とサーモクロミック型の分離の傾向は消え、両技術とも全体に分散した。ケース５の母集団には、これら 2 つの技術以外にエレクトロクロミック、DPS、多層干渉等々の技術が含まれており、例えば高分子関連の用語など、重要な共通語が多くみられた。

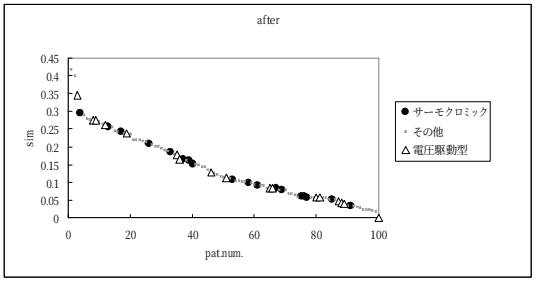


図 3.3. 特許 100 件の処理結果（調整後）

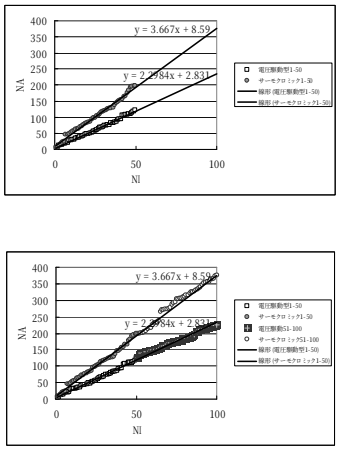
3.2. クラスターの視覚化

クラスターを視覚化するための基礎的な検討を、簡潔な例であるケース１について検討した。隣接する特許間での共通語数をまとめると表 3.2 のようになった。電圧駆動型では 37-92、サーモクロミック型では 211-318 であり、大きな差が見られた。

表 3.2 隣接する特許との共通語数

項目	用語数	技術属性
全11件の共通語	27	
JP2005060703-WO03057799	92	電圧駆動
WO03057799-JP2005250119	90	電圧駆動
JP2005250119-JP05045679	54	電圧駆動
JP05045679-JP2005082472	37	電圧駆動
JP2005082472-JP05025479	51	電圧駆動
JP05025479-JP11265005	50	電圧駆動-サーモクロミック
JP11265005-JP07242447	318	サーモクロミック
JP07242447-JP07171925	211	サーモクロミック
JP07171925-JP07171926	294	サーモクロミック
JP07171926-JP07318983	41	サーモクロミック-電圧駆動

また、ケース１の各クラスターに特有の用語について、各クラスター内での順位と処理対象の母集団全体での順位とについて回帰分析を行い、その結果を図 3.4 に示した。順位が 1-50（図 3.4.a）と 51-100（図 3.4.b）において、同じ相関がみられた。従



(a)
順位 1-50

(b)
順位 1-50
+ 51-100

図 3.4. クラスター特有語の回帰分析

って、特定の特許を適切なクラスターに配置する上で、回帰分析は有用なツールであることが示唆された。

４．考察

termmi によるクラスター形成の可否は、ケース１～５の解析結果から、母集団の特許件数と技術的錯綜状態に依存していることが窺えた。また、total.txt での用語調整においては、用語の削除によりクラスターの形成が可能となるケースと逆の場合もあることが分かった。調光遮熱技術全般をクラスターの対象とする場合には用語の削減は適しており、大きな概念を細分した個々の技術についてのクラスターを形成する場合には、種々の工夫を必要とする傾向が窺えた。

一方、上記 3.2 で示したように、各種のデータを活用することにより、クラスターの視覚化も可能になることが窺えた。例えば、隣接する特許間での共通語数が非常に少ない特許２件（JP05045679、JP2005082472)は前３組、後ろ１組を構成する他の特許４件と同じ相関係数をもつので、これら６件は同一のクラスターとして扱うことが可能となる。なお、今後のシステムの発展性としては、各ケースでみる解析の深さの点から、interactive で stepwise の特許情報解析システムが考えられる。

５．結論

termmi は有用な特許情報の解析ツールである。また、概念検索のツールとしても利用できる。

参考文献

[1]東京大学付属図書館報「図書館の窓」Vol.43 No.3, pp.61-65 (2004)
[2] ”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/
[3]特開 2006-30327