テキストマイニングによる公報間類似度マップの検討:

○高岡恵理[1)], 安藤俊幸[2)]
アジア特許情報研究会[1)], 花王株式会社[2)]
〒300-1260 茨城県つくば市西大井1733－15
TEL 090-8700-7256
E mail:patentsearch2006@yahoo.co.jp

Study of similarity map between patent publications by the text mining:

TAKAOKA Eri[1)], ANDO Toshiyuki[2)],
Asia Patent Innformation Society [1)], KAO Corporation[2)]
1733-15, Nishioi, Tsukuba-shi, Ibaraki-Ken, 300-1260 Japan
Phone: +81-90-8700-7256
E-mail: patentsearch2006@yahoo.co.jp

【発表概要】

特許公報間の技術の類似度を評価する手法として、発明の名称、要約、請求項等の文章情報をテキストマイニングし、切り出した用語の出現頻度、出現文書数から重み付けを行った上で公報間類似度を計算した。各公報間の類似度は非計量多次元尺度法により距離に換算し二次元空間にプロットする類似度マップを作成、評価した。

テスト集合に電動歯ブラシ関連公報を、用語の切り出しでは、専門用語(複数の名詞が結合した複合語)として抽出した場合と、形態素解析 MeCab にて品詞単位で用語を抽出した場合とで比較したところ、形態素解析にて作成した公報間類似度マップの方が、目視判断に近い結果となった。これは、公報文内において使用される専門用語の揺れが、形態素単位で細かく分割したことで修正された為と考えられる。

目視による結果に近い類似度マップ作成手法について、用語の統制、特許分類の利用の観点等からも考察を進めたので報告する。

【キーワード】

特許公報, 類似度, テキストマイニング, 非計量多次元尺度法, 形態素解析, 特許分類, KH Coder

1. はじめに

特許情報を活用した技術動向分析において、各種マップソフトやデータベース内蔵の統計ツールは、近年、非常に利用しやすくなってきており、マクロな動向把握であれば分析作業のハードルは低いものとなった。

一方、個々の技術の構成要件を読み込む前のセミミクロ段階における“技術の仕分け"に関しては、まだまだ人手に頼らざるを得ないのが実情であり、満足のいく仕分けを自動処理できる汎用ツールは未だ見当たらないといえる。

そこで、特許公開公報の書誌情報をマイニングし、抽出された要素(用語や技術分類等)から公報間類似度を計算して技術を仕分け、その結果を可視化できないか検討した。

書誌情報データには、和文テキスト情報を Shareresearch[1]から、 ファミリー情報を含む書誌情報及び英文テキスト情報を PatBase[2]から、そして技術分類情報を Thomson Innovation[3]の DWPI データから取得した。

また、テキストマイニングには自作 PatAnalyzer[4]の他、立命館大学産業社会学部樋口耕一准教授が制作・提供されている KH Coder[5]を使用した。

2. テスト集合について

公報間類似度の評価にあたり、テスト集合として過去 10 年間の電動歯ブラシ関連の公報で、以下の条件を具備した 46 件を選定した。

① 英文、和文の「発明の名称」、「要約」、「請求項」の各テキスト情報が揃っている。
② 技術分類コードとして、IPC、FI・F ターム、CPC の何れもが付与されている。
③ 複数(3 社以上)の出願人により継続的に出願されている。

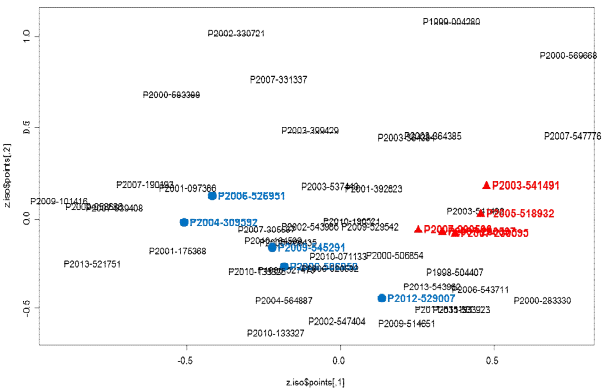
テスト集合として電動歯ブラシ分野を選定した理由としては、公報記載の文言に専門用語(複合語、外来語等)が多く、出願人間で技術表現が異なる一方で、図面を見れば技術内容を把握でき、類似度計算結果を目視で評価可能な点が挙げられる。

3. テキスト情報を利用した公報間類似度について

Shareresearch からダウンロードした「発明の名称」、「要約」、「請求の範囲(全請求項)」の和文テキスト情報から、1)専門用語を抽出した場合と、2)形態素単位で単語を抽出した場合とで、各々の用語の出現頻度等から公報間類似度を計算した。

専門用語の抽出は、PatAnalyzer にて saezuri lite[6]を介して、形態素解析ツールMeCab[7]の品詞と隣接頻度情報から求めた。また、形態素単位での単語の抽出には、MeCab の形態素(名詞)をそのまま利用した。

公報間類似度の計算には、Cosine 係数を用い、TF・IDF 値により重み付けした。また、統計ソフト R[8]を用いて公報間類似度を非計量多次元尺度法により距離に変換し、二次元空間にプロットした。図1に専門用語で抽出した場合の公報間類似度マップを示す。



【図1】専門用語で抽出した結果

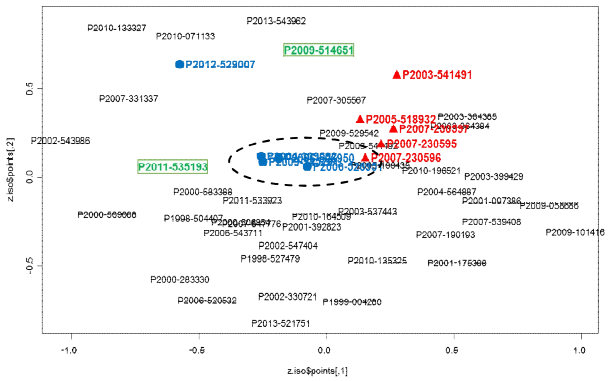
図1では、関連技術が近接してプロットされた公報群(▲を付した 5 件)がある一方、同一出願人による一連の関連技術であるにも関わらず離れてプロットされた公報群(●を付した 5 件)も認められた。

後者に関しては、出願時期が 2004 年から 2012 年に跨っており、表1に示すように、その間に使用された技術用語が同義であるにも関わらず異なっていた(訳語の相違を含む)ことが影響していると考えられる。また、特許公報で多用される「前記」や「該」を含む複合語も類似度計算結果に少なからず影響している(この点に関しては、強制排除する等により改善できる可能性がある。)。

P2004-309592		P2006-506950		P2009-545291	
歯ブラシ	68	記載	93	立上り要素	13
特徴	64	歯ブラシ	86	歯ブラシ	13
こと	64	前記ヘッド	43	:	
歯清掃要素	35	前記歯クリーニング要素	36	該歯ブラシ	3
前記歯清掃要素	32	歯クリーニング要素	33	:	
:		:		:	
タフト	7	:		:	
:		:		:	
回動	6	回転	24	旋回可能	1
回動自在	5	:		:	
:		回転自在	4	歯清掃要素	1
:		:		歯肉処置要素	1
:		ブリストルタフト	4	:	

【表1】抽出された公報別専門用語

他方、MeCab により形態素単位で単語を抽出した場合では、同義だが部分的に異なった用語を含む複合語が、形態素解析により分割され、図1で認められた前述の問題が修正されていた(図2の破線で囲んだ 4 件。P2012-529007 は、他の 4 件と技術主題を異にする。)。



【図2】形態素解析による抽出結果

しかし、ほぼ同一技術にも関わらず、着目部が異なる為に技術用語と使用語数が異なった結果、非類似と計算され、離れてプロットされた P2009-514651 と P2011-535193 のようなケースも認められた(図2、表2)。これは、公報記載のテキスト情報をコンピューターによりテキストマイニングして公報間類似度を求める場合の限界といえる。

Term	TF	DF	IDF	TF*IDF	文書
ヘッド	589	21	1.9	1100.0	**P2009-514651**|P2009-545291|***P2011-535193***他
歯ブラシ	827	41	1.2	991.1	**P2009-514651**|***P2011-535193***|P2013-521751他
:					
角度	99	17	2.1	205.8	|P2009-101416|P2011-533923|***P2011-535193***他
外側	93	15	2.2	205.0	**P2009-514651**|P2011-533923|***P2011-535193***他
上面	69	7	3.0	204.7	P2006-526951|**P2009-514651**|P2009-545291他
フィラメント	41	5	3.3	135.4	P2007-539408|**P2009-514651**|P2010-196521他
シェル	15	1	4.9	73.7	**P2009-514651**
外形	10	4	3.5	35.3	**P2009-514651**|P2010-133327他
円柱	3	2	4.2	12.7	P2007-331337|**P2009-514651**

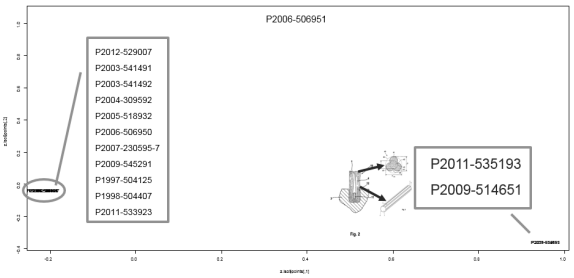
【表2】抽出語と出現文書の抜粋

4. 技術分類の説明文を利用した公報間類似度

形態素解析で判明した問題を回避するために、各公報に付与された技術分類コードの技術説明文を、付与された組み合わせに応じて統合、これをテキストマイングすることを試みた。利用した技術分類は、電動歯ブラシに関して比較的階層を多く有する CPC(Cooperative Patent Classification)を利用した。CPC 技術説明文の統合にあたっては、サブクラスの説明文から含め、分類階層の最深部の説明文までを単純に統合した。テキストマイニングは、MeCab により形態素単位で単語を抽出した。

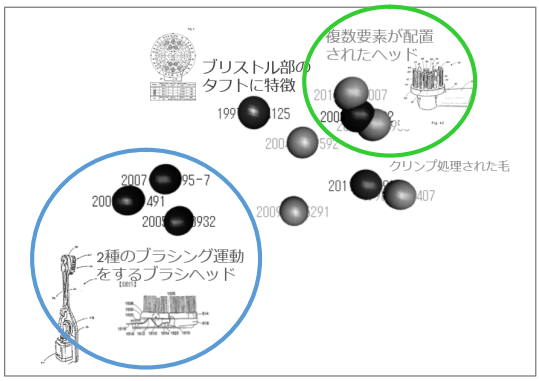
CPC 分類の技術説明文を公報のテキスト情報の代わりに援用することで、公報間で相違する用語が統制され、また審査官による技術把握の結果を利用できる。

結果は、図3に示すように技術の構成要素が異なる公報は互いに離れてプロットされ、部分的な相違を有する近接技術(P2009-514651 と P2011-535193 のケース)は重なってプロットでき期待通りであったものの、残念ながら視認性に劣るマップとなった。



【図3】CPC 分類説明文を利用した結果

図3の左側には 13 件の公報が重なってプロットされていた為、多次元尺度法で三次元空間にプロットしたところ、目視判断に近い距離で配置されていることを確認できた(図4)。



【図4】13 件の 3D 類似度マップ

5. 結論

公報記載の文章情報によるテキストマイニングより、CPC 技術分類を利用したマイニングによる方が、目視判断に近い結果となった。

6. 考察

これまでの検討結果から、他の特許分類の利用可能性についても考察した。

分類コードに基づいて複数の文書をまとめることは既に研究されており、例えば Kang ら[9]は、テキストマイニングの代わりに IPC 分類コードを利用しており、また Konishi ら[10]は、TF･IDF 値に IPC 分類コードベースの用語の重み付けを統合している。

ここでは、複数付与された各技術分類コードに基づき KH Coder を用いて、テスト集合の 46 件をクラスタリングした。

条件としては、公報間類似度の計算に Cosine 係数を用い、類似性の高い公報を 10 個のグループに分けた。クラスター間の距離の決定には、Ward 法を採用した。また、各コードの出現数は文書毎に標準化し、TF･IDF 値により重み付けした。

本検討では、Thomson Innovation から取得した分類コードを利用した。

クラスタリング結果から、CPC 分類、又は CPC 分類と IPC 分類の組み合わせにより精度が上がった(表3)ことから、目視判断に近い類似度マップの作成におけるヒントが得られた。

PatBase Family No.	IPC	CPC	IPC+CPC	FI	Fターム	FI+Fターム	IPC+FI
12200988	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20583692	0.0	0.0	0.0	0.0	0.0	0.0	0.0
20702345	0.0	0.0	0.0				0.0
21361742	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28516322	0.0	0.0	0.0	0.0	0.0	0.0	0.0
28813046	0.0	0.0	0.0	0.0	0.0	0.0	0.0
29773615	0.0	0.0	0.0	0.0			0.0
30061351	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30666713	0.0	0.0	0.0	0.0	0.0	0.0	0.0
30951420		0.0	0.0				
31037980	0.0	0.0	0.0	0.0	0.0	0.0	0.0
31058670	0.0	0.0	0.0	0.0			0.0
31410886		0.0	0.0	0.0			0.0
31770050	0.0	0.0	0.0	0.0		0.0	0.0
32017507	0.0	0.0	0.0	0.0		0.0	0.0
40970518	0.0	0.0	0.0	0.0	0.0	0.0	0.0
41657854	0.0	0.0	0.0	0.0	0.0	0.0	0.0
42066018	0.0	0.0	0.0	0.0	0.0	0.0	0.0
43352823	0.0	0.0	0.0	0.0	0.0	0.0	0.0
44976057	0.0	0.0	0.0	0.0	0.0		0.0
44989454	0.0	0.0	0.0	0.0	0.0	0.0	0.0
48463943	0.0	0.0	0.0	0.0			0.0
49983112	0.0	0.0	0.0	0.0	0.0	0.0	0.0
50550028	0.0	0.0	0.0	0.0	0.0	0.0	0.0
51342209	0.0	0.0	0.0	0.0	0.0	0.0	0.0

【表3】PatBase ファミリー別のクラスター結果(バラつきがあったファミリーに網掛け)

7. おわりに

本検討では、テキストマイニングの手法を応用して公報間類似度マップを作成した。そして、類似度計算には、いかなる変数をもって行うことが、最も目視判断に近い結果が得られるのかを検討してきた。

「発明の名称」や、「要約」等の文章情報を利用する場合においては、専門用語抽出だけでなく、形態素単位での単語の抽出も適宜取り入れることも有効である。

また、技術分類、特に技術階層が 25 万分類と細分化された CPC 分類コード

を公報間類似度計算において併用することで精度向上の可能性も認められた。

ただ、今回の検討では、目視判断との差異を確認する都合上、意図的にテスト集合を用意した為に、他の技術分野においても同様の結果が得られるのかについては、今後確認する必要がある。

[謝辞]

本報告は2015年度の「アジア特許情報研究会」のワーキングの一環として報告するものであり、テキストマイニングチームの皆様には情報の提供及び数々のアドバイスを頂きました。

ここに改めてお礼申し上げます。

8. 参考文献

[1] 株式会社日立製作所. Shareresearch. http://www.hitachi.co.jp/Prod/comp/app/tokkyo/sr/ (参照 2015-10-6).

[2] RWS Group. PatBase. http://www.patbase.com/login.asp (参照 2015-10-6).

[3] Thomson Reuters. http://ip-science.thomsonreuters.jp/products/ti/ (参照 2015-10-6).

[4] 安藤 俊幸ら.“中国特許解析・テキストマイニングによるKW分析”第 11 回情報プロフェッショナルシンポジウム https://www.jstage.jst.go.jp/article/infopro/2014/0/2014_31/_pdf (参照 2015-10-8).

[5] 立命館大学産業社会学部・樋口耕一准教授. KH Coder. Ver.2.00b. http://khc.sourceforge.net/ (参照 2015-10-6).

[6] 自然言語処理支援ライブラリ. saezuri lite. http://www.vector.co.jp/soft/winnt/prog/se495669.html (参照 2015-10-8).

[7] 京都大学情報学研究科−日本電信電話株式会社コミュニケーション科学基礎研究所 共同研究ユニットプロジェクトを通じて開発されたオープンソース 形態素解析エンジン. MeCab. http://mecab.googlecode.com/svn/trunk/mecab/doc/index.html (参照 2015-10-6).

[8] GNU プロジェクトの一つであり、AT&T(当時)のベル研究所で John M. Chambers らにより開発された S 言語・環境に似ている統計計算とグラフィックスのための言語・環境. http://www.okadajp.org/RWiki/?R%E3%81%A8%E3%81%AF (参照 2015-10-6).

[9] Kang, I.-S.; Na, S.-H.; Kim, J.; and Lee, J.-H. Cluster-based patent retrieval. Information Processing & Management. 2007, 43, 5, pp.1173-1182.

[10] Konishi, K.; Kitauchi; A. and Takaki, T. Invalidity Patent Search System of NTT DATA. Proceedings of the 4th NTCIR Workshop. 2004. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings4/PATENT/NTCIR4-PATENT-KonishiK.pdf. (参照 2015-10-9).