INFOPRO2015 A21

テキストマイニングによる公報間類似度マップの検討

第12回情報プロフェッショナルシンポジウム
2015年12月11日

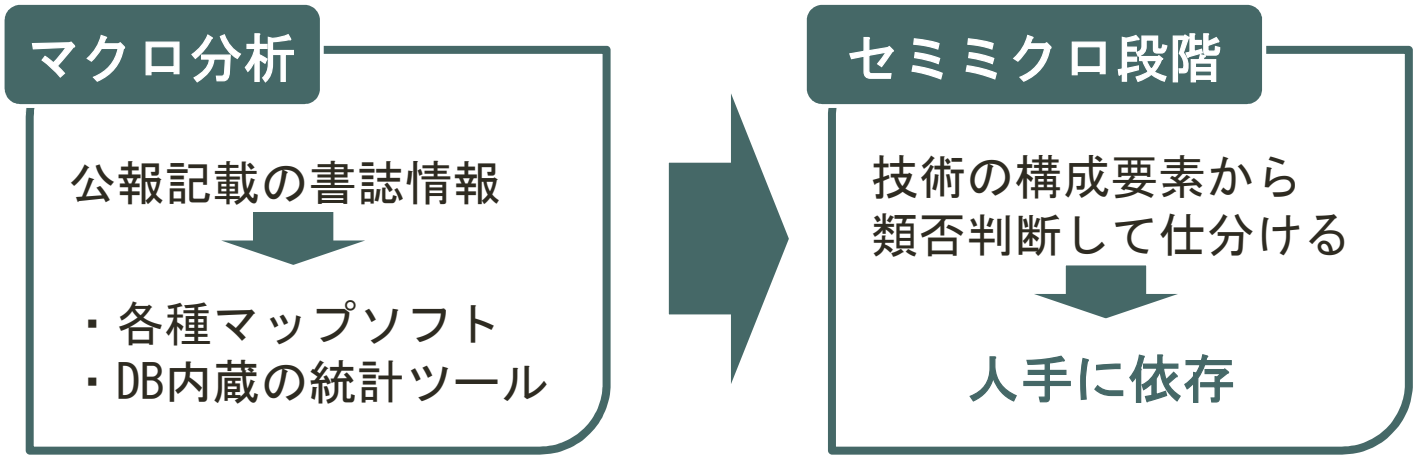
○アジア特許情報研究会　高岡恵理
花王株式会社　安藤俊幸

発表内容

- 検討に至った背景
- 用いたテスト集合
- テキスト情報による公報間類似度
- 技術分類を併用した場合の類似度
- まとめ
- 今後の課題

検討に至った背景

特許情報を活用した技術分析において、



そこで、特許公開公報を **テキストマイニング**し、抽出された

技術用語から**公報間類似度**を計算し、技術を仕分けられないか？

用いたテスト集合

：過去10年間の**電動歯ブラシ**関連の公開公報46件

選定理由

●公報記載の文言に専門用語（複合語、外来語等）が多い
●出願人間で技術用語が異なる＝異表記同義語が多い

また、
　分類結果の妥当性を図面から比較的容易に判断ができる。

選定条件

●タイトル、要約、請求項の和文テキスト情報が揃っている
●IPC、FI・Fターム、CPCの技術分類が付与されている
●複数の出願人（3社以上）により継続的に出願されている

データソース＆使用ツール

使用データ	データソース
和文テキスト	Shareresearch/日立製作所
DWPIデータ	Thomson Innovation/トムソン・ロイター
ファミリー情報	PatBase（拡張ファミリーNo.)/RWSグループ

データ加工	使用ツール
和文テキストマイニング	PatAnalyzer/花王㈱安藤さん制作
クラスタリング	KH Coder/立命館大学樋口耕一准教授制作
類似度計算	SimCalc/花王㈱安藤さん制作
類似度の距離変換	R：library(MASS)
多次元空間へのプロット	R：library(rgl)

テキストマイニング処理の流れ

テスト集合

出願番号(西暦表示)	発明の名称	要約	請求の範囲(全)
P2004-309592	歯ブラシ	(修正有)【課題】歯	【1】ヘッドと、このヘッドの上面か
P2006-506950	歯ブラシヘッド	【課題】 歯クリーニ	【1】歯ブラシヘッドにおいて、ヘッ
P2007-054703	歯ブラシヘッド	【課題】歯クリーニン	【1】歯ブラシヘッドにおいて、ヘッ
P2007-054715	歯ブラシヘッド	【課題】歯クリーニン	【1】歯ブラシヘッドにおいて、ヘッ
P2009-529542	電動歯ブラシと、電動歯	本発明の伝動装置	【1】電動歯ブラシ(12)に用いら
P2006-543711	流体方向づけ部材を備え	本発明は、電動歯ブ	【1】電動歯ブラシ用のリムを備え
P2011-148292	流体方向づけ部材を備え	【課題】本発明は、電	【1】電動歯ブラシ用のブラシヘッ
P2012-529007	口腔ケア製品並びにそれ	口腔ケア器具は、口	【1】口腔への挿入のために寸法
P2013-235956	口腔ケア製品並びにそれ	(修正有)【課題】ベ	【1】口腔への挿入のために寸法
P2009-514651	歯ブラシおよび歯ブラシ	本発明は、モノフィラ	【1】モノフィラメントとして構成さ
P2011-535193	歯ブラシ並びに歯ブラシ	本発明は、歯ブラシ	【1】歯ブラシ用ブリストルであっ
P2014-105537	歯ブラシ並びに歯ブラシ	(修正有)【課題】歯	【1】歯ブラシ用ブリストルであっ
P2011-533923	電気歯ブラシ、及び電気	本発明は、概して、	【1】電気歯ブラシ(1)用のブラシ
P2011-533925	電気歯ブラシ、及び電気	本発明は、概して、	【1】電気歯ブラシ(1)用のブラシ
P2013-543962	効果的な洗浄のための	歯ブラシ10又はマ	【1】柄と、基部部材及び歯と接触
P2003-541489	複雑な運動の歯ブラシ	電気で駆動する歯ブ	【1】第1カム及び第2カムを含む
P2003-541491	複数運動歯ブラシ	電動歯ブラシを開示	【1】第1端部、第1端部の反対側
P2003-541492	複数動作の歯ブラシ	電動歯ブラシが提供	【1】内部にモータが配置されたハ
P2003-541493	複雑運動の歯ブラシ	電動歯ブラシが、遠	【1】電動歯ブラシであって、歯ブ
P2007-180265	電動歯ブラシ	【課題】製造するの	【1】電動歯ブラシであって、歯ブ

テキスト情報：Excelファイル

抽出処理
PatAnalyzer

・専門用語
・形態素単位

抽出語　頻度

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	:	

類似度計算

・TF（Term頻度）
・DF（文書頻度）
TF・IDFで重み付け

全公報間類似度　Cosine係数使用/SimCalc

公報番号	P1998-504407	P1998-527479	P2000-506854	P2000-569668	P2000-583388	P2001-175368	P2002-543985	P2002-543986
P1998-504407	0	0	0					
P1998-527479	0.9546223	0	0					
P2000-506854	0.833376	0.7953697	0					
P2000-569668	0.9714311	0.9447964	0.9786615					
P2000-583388	0.5888051	0.945308	0.9365913					
P2001-175368	0.9651633	0.8626221	0.9277118	0.9930673	0.8970379	0	0	0
P2002-543985	0.9607788	0.9986628	0.9939299	0.9934112	0.983174	0.9911553	0	0
P2002-543986	0.8939807	0.8687187	0.8430547	0.9736708	0.9543828	0.9458383	0.8821606	0

統計ソフトR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

分析対象のテキスト情報

テキスト情報

- タイトル
- 要約
- 全請求項

用語抽出方法

- 専門用語の抽出
- 形態素単位での単語抽出
- ノイズ除去後の専門用語

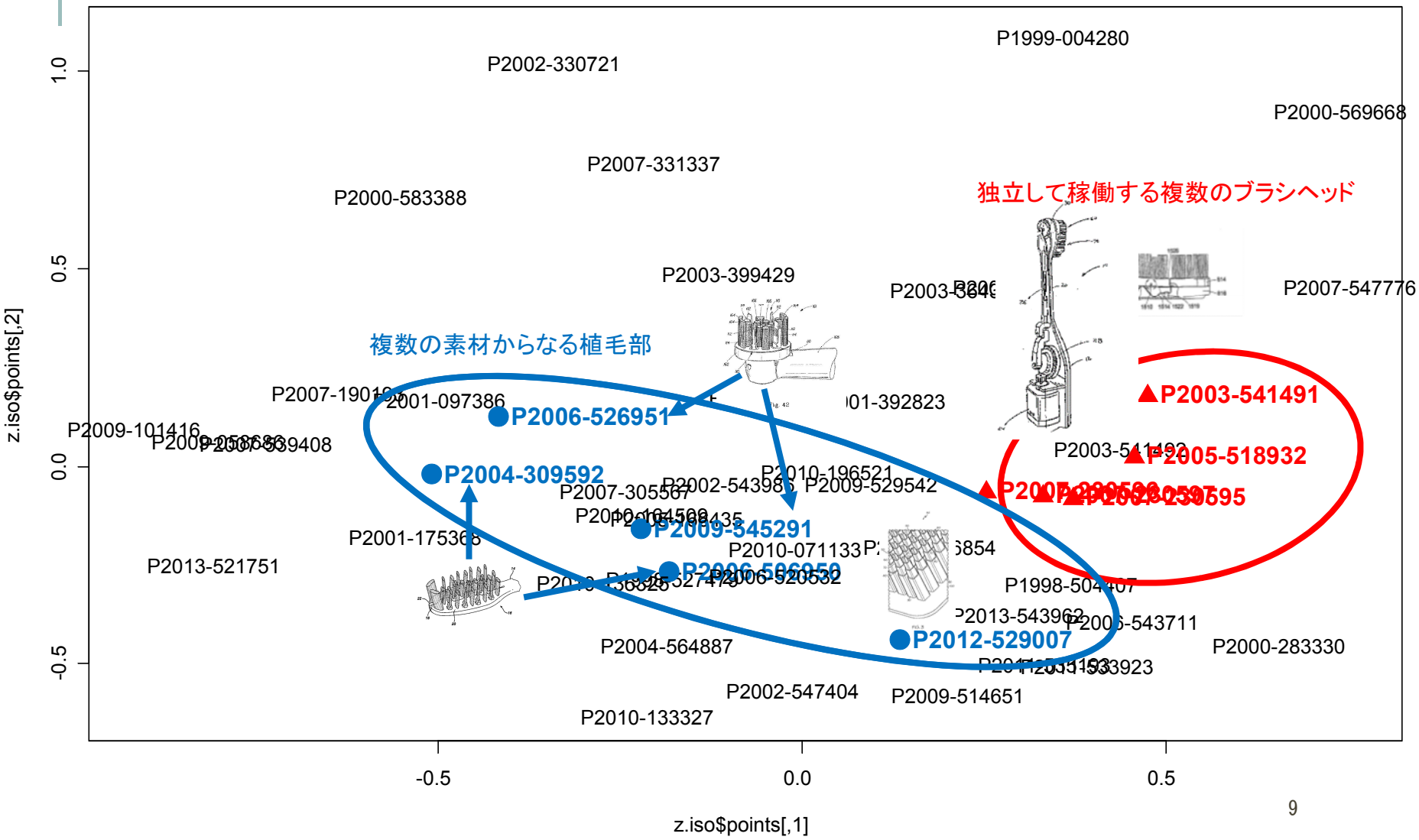
専門用語の抽出は、・・・・・
自作PatAnalyzerにてsaezuri liteを介して、形態素解析ツールMeCabの品詞と隣接頻度情報から求めた。

形態素単位での単語抽出には、MeCabの形態素（名詞）を　そのまま利用した。

発表内容

- 検討に至った背景
- 用いたテスト集合
- **テキスト情報による公報間類似度**
- 技術分類を併用した場合の類似度
- まとめ
- 今後の課題

専門用語による抽出結果



関連技術が非類似となった理由

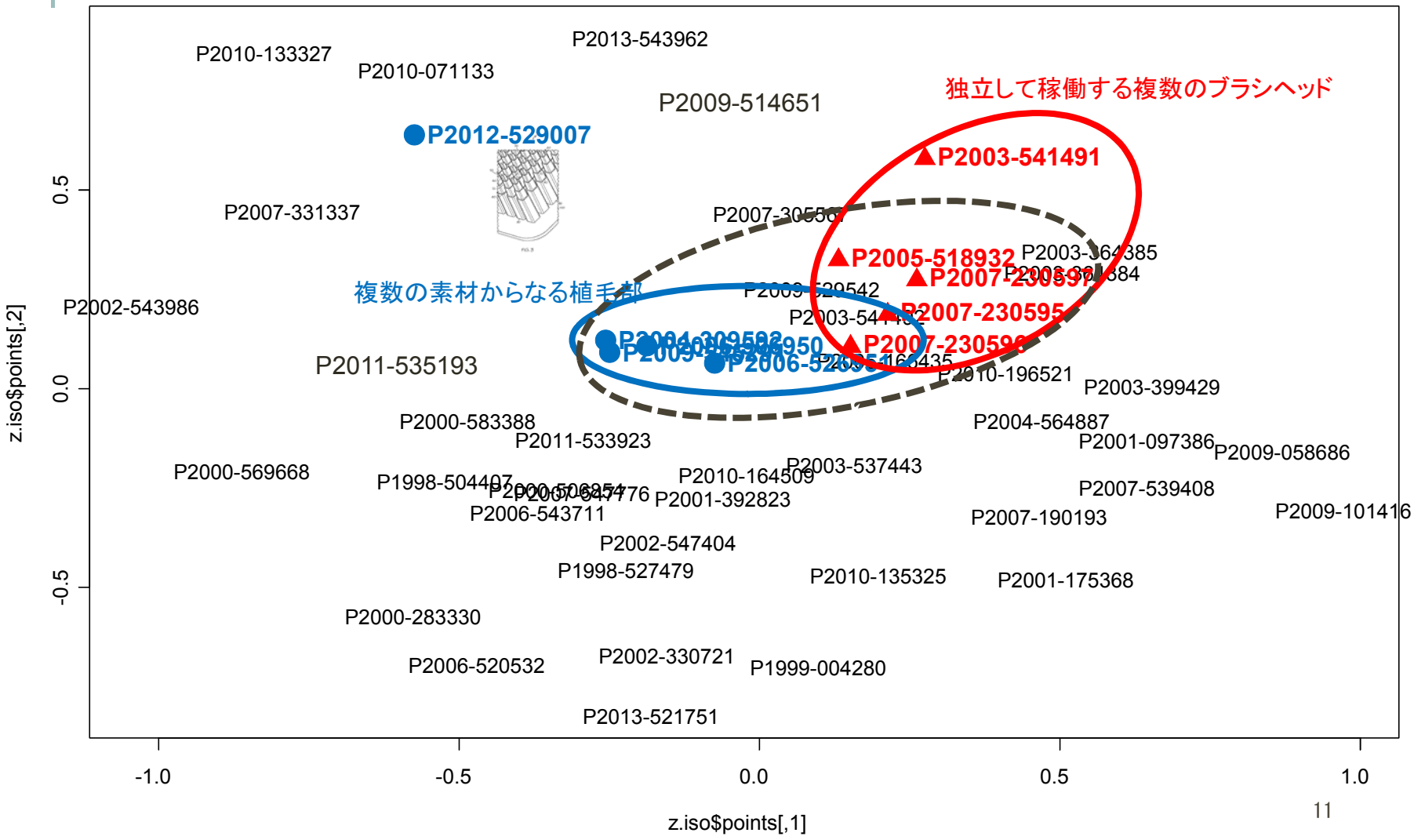
P2004-309592	
歯ブラシ	68
特徴	64
こと	64
歯清掃要素	35
前記歯清掃要素	32
:	
タフト	7
:	
回動	6
回動自在	5
:	
:	
:	

P2006-506950	
記載	93
歯ブラシ	86
前記ヘッド	43
前記歯クリーニング要素	36
歯クリーニング要素	33
:	
:	
:	
回転	24
:	
回転自在	4
:	
ブリストルタフト	4

P2009-545291	
立上り要素	13
歯ブラシ	13
:	
該歯ブラシ	3
:	
:	
:	
:	
旋回可能	1
:	
歯清掃要素	1
歯肉処置要素	1
:	

専門用語（複合語、外来語等）による抽出でなく、形態素単位の方が良いのでは？

形態素単位での単語抽出



専門用語のノイズによる影響

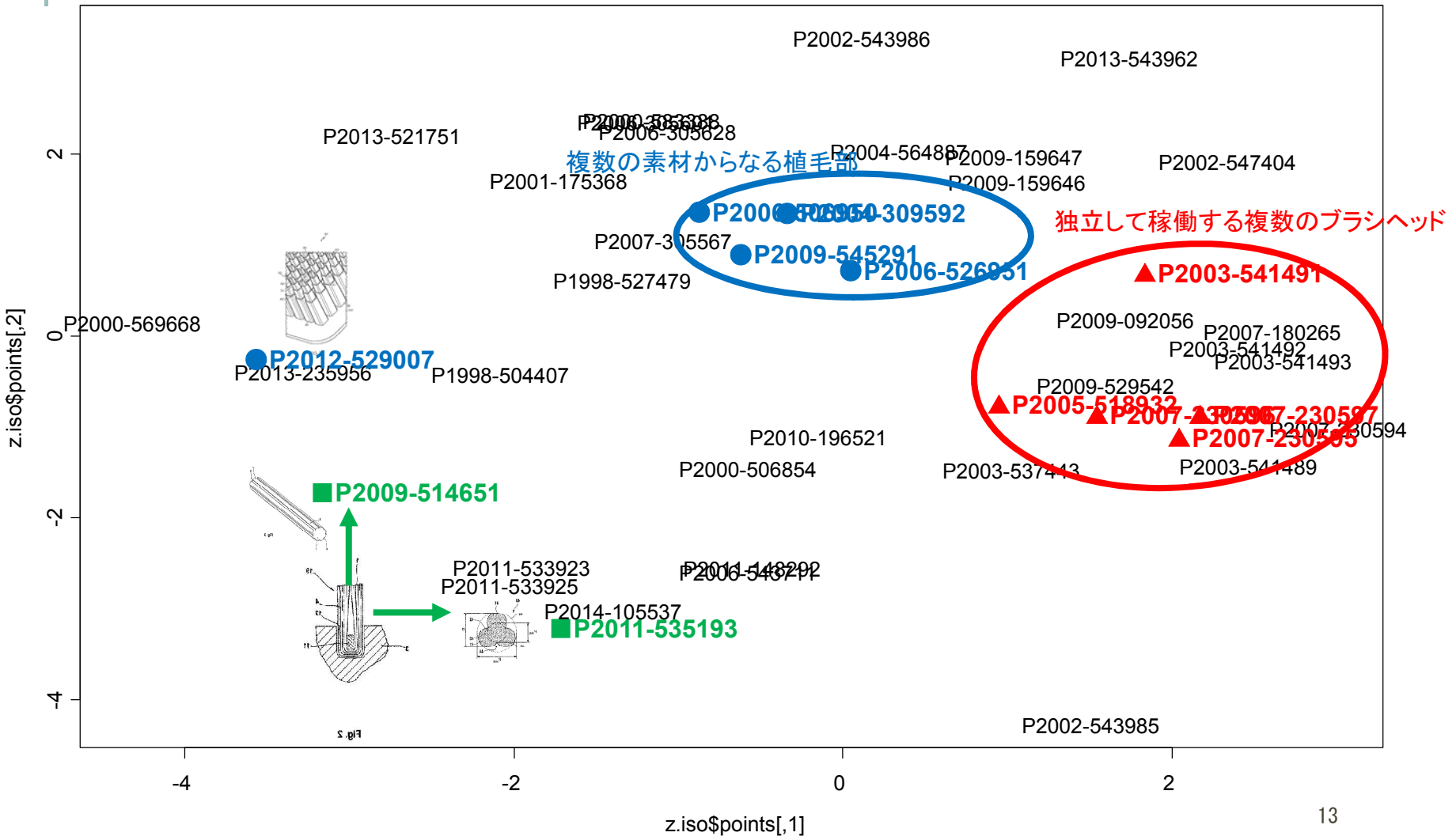
P2004-309592	TF
歯ブラシ	68
特徴	64
こと	64
歯清掃要素	35
前記歯清掃要素	32
:	
タフト	7
:	
回動	6
回動自在	5
:	
:	
:	

P2006-506950	TF
記載	93
歯ブラシ	86
前記ヘッド	43
前記歯クリーニング要素	36
歯クリーニング要素	33
:	
:	
:	
回転	24
:	
回転自在	4
:	
ブリストルタフト	4

P2009-545291	TF
立上り要素	13
歯ブラシ	13
:	
該歯ブラシ	3
:	
:	
:	
:	
旋回可能	1
:	
歯清掃要素	1
歯肉処置要素	1
:	

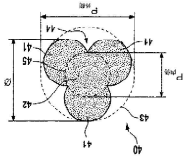
特許公報で多用される「前記」等を含む複合語も類似度計算結果に影響しているでは？

ノイズ除去後の専門用語

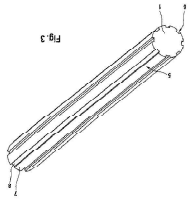


関連技術が非類似となった理由

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



P2011-535193
毛の断面形状を特定



P2009-514651
毛の側面形状を特定

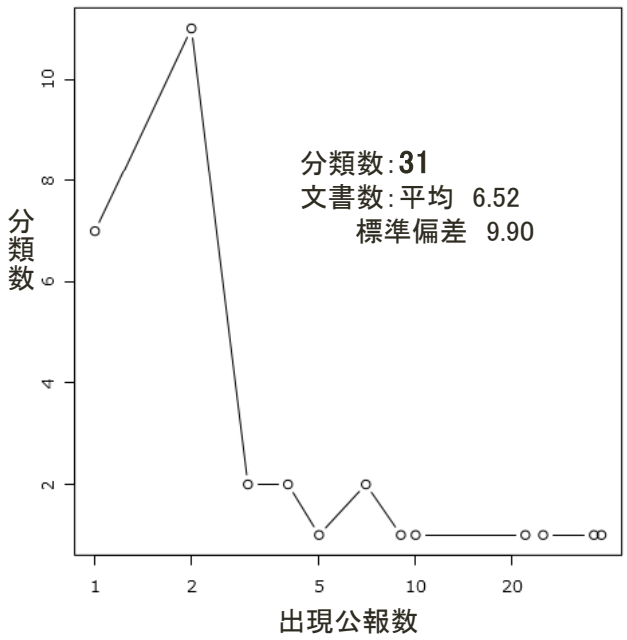
公報記載のテキスト情報からテキストマイニングして類似度を求める場合の限界？

発表内容

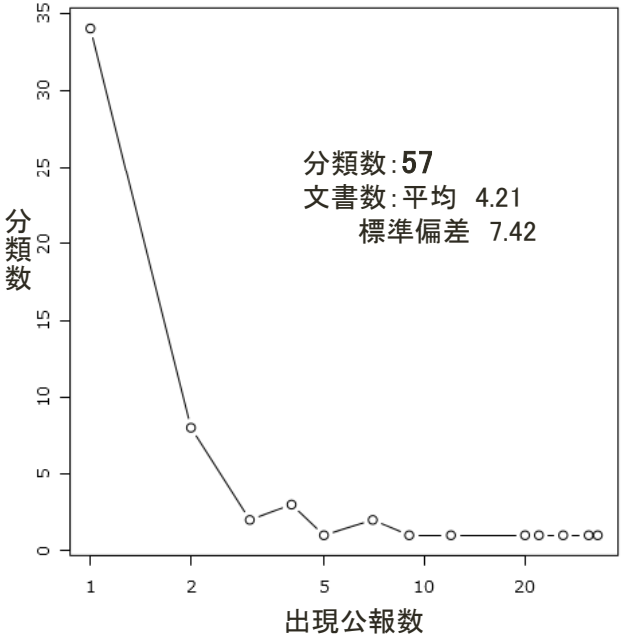
- 検討に至った背景
- 用いたテスト集合
- テキスト情報による公報間類似度
- **技術分類を併用した場合の類似度**
- まとめ
- 今後の課題

テスト集合に付与された分類

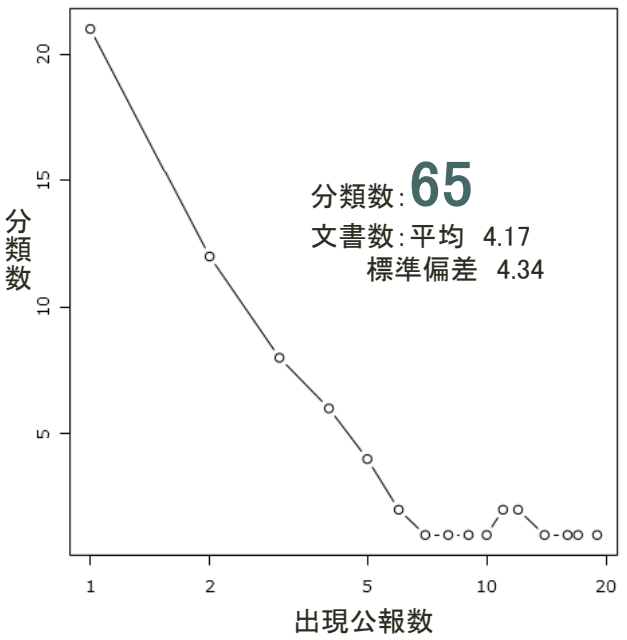
① 付与IPC分類の分布



② 付与FI分類の分布



③ 付与CPC分類の分布



IPC，FI，CPCいずれの技術分類も、付与されている分類は適度にバラついている

（KHCoder使用）

CPC分類を選んだもう一つの理由

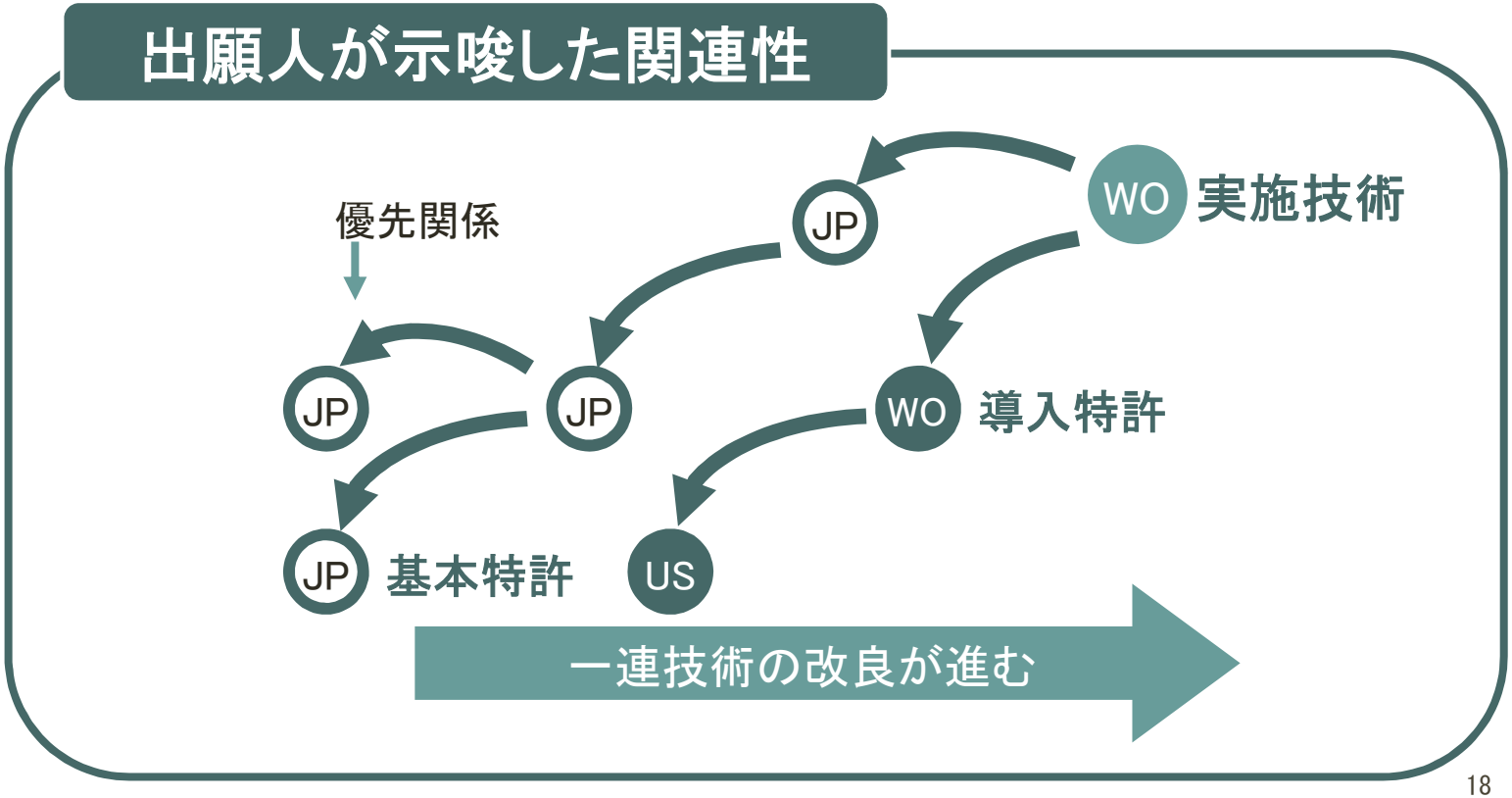
テスト集合を各技術分類に基づいてクラスタリング（KHCoder使用）し、どの分類を利用したときに同一ファミリーに属する公報が同じクラスターに集結してくるかを確認した。

Database 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
41657854	0.0	0.0			.0	0.0	0.0
42066018	0.0	0.0			.0	0.0	0.0
43352823	0.0	0.0			.0	0.0	0.0
44976057	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

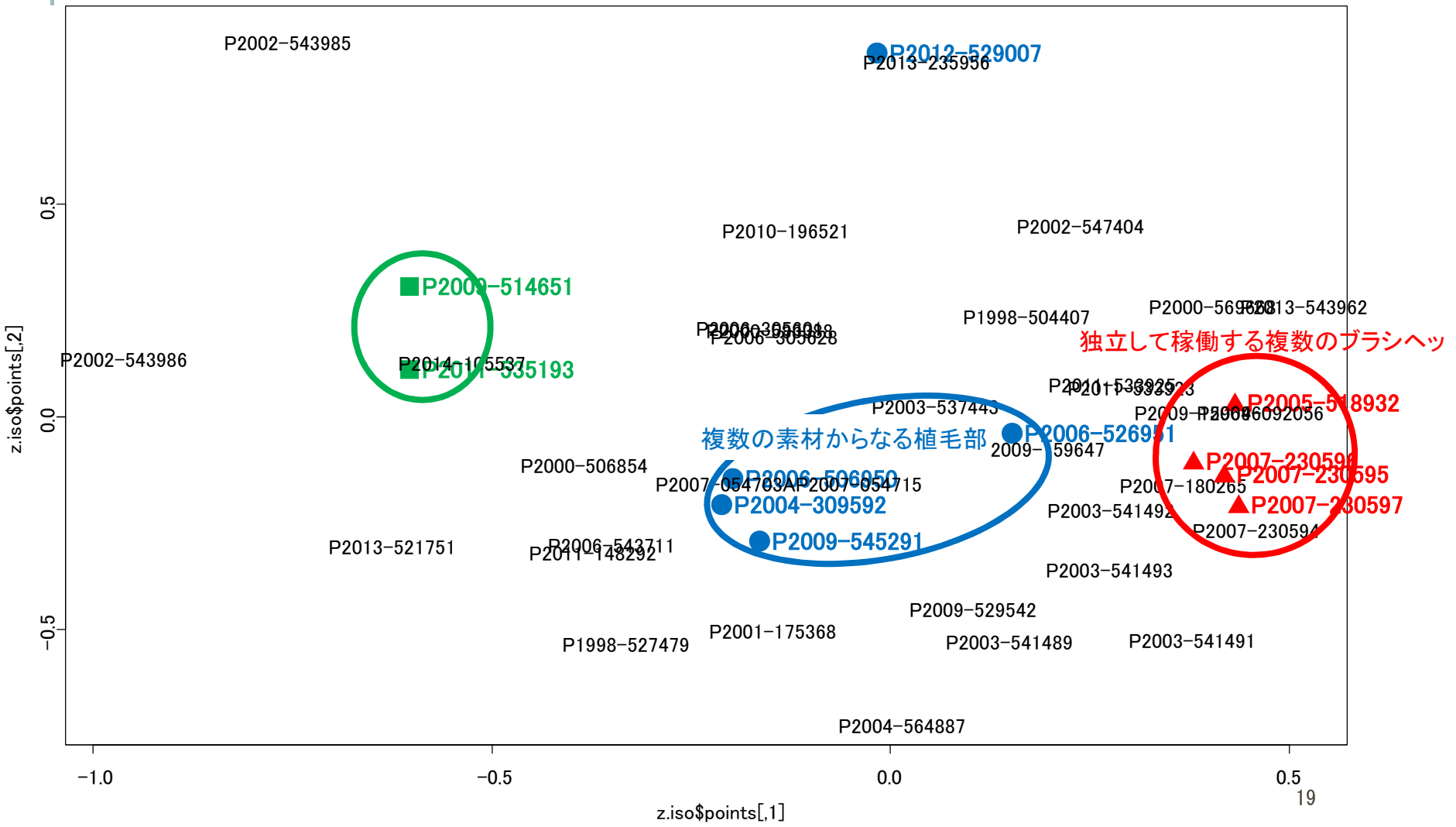
バラツキが無い＝0

PATBASEの拡張ファミリー

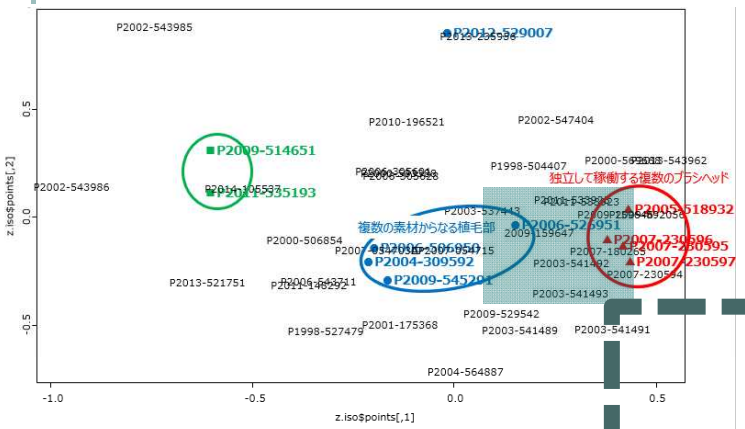
：**優先関係**に繋がりがある限り、同一ファミリー



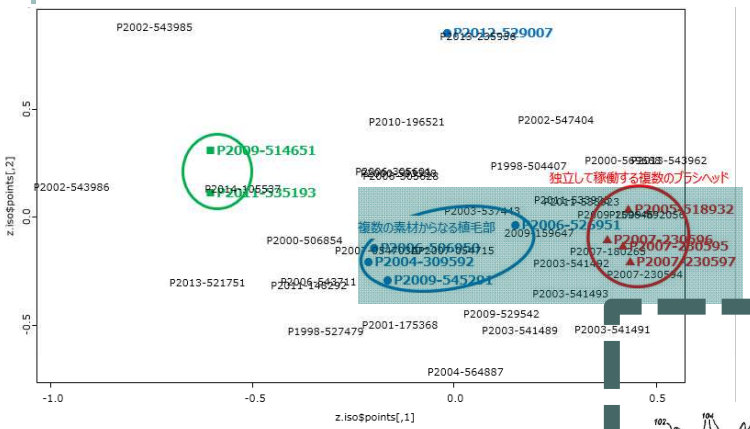
専門用語#ノイズ+CPC分類



専門用語#ノイズ+CPC分類



専門用語#ノイズ+CPC分類



中間エリアの公報文献：
独立して稼働する複数の素材からなる植毛部を有するブラシヘッド

複数の素材からなる植毛部

独立して稼働する複数のブラシヘッド

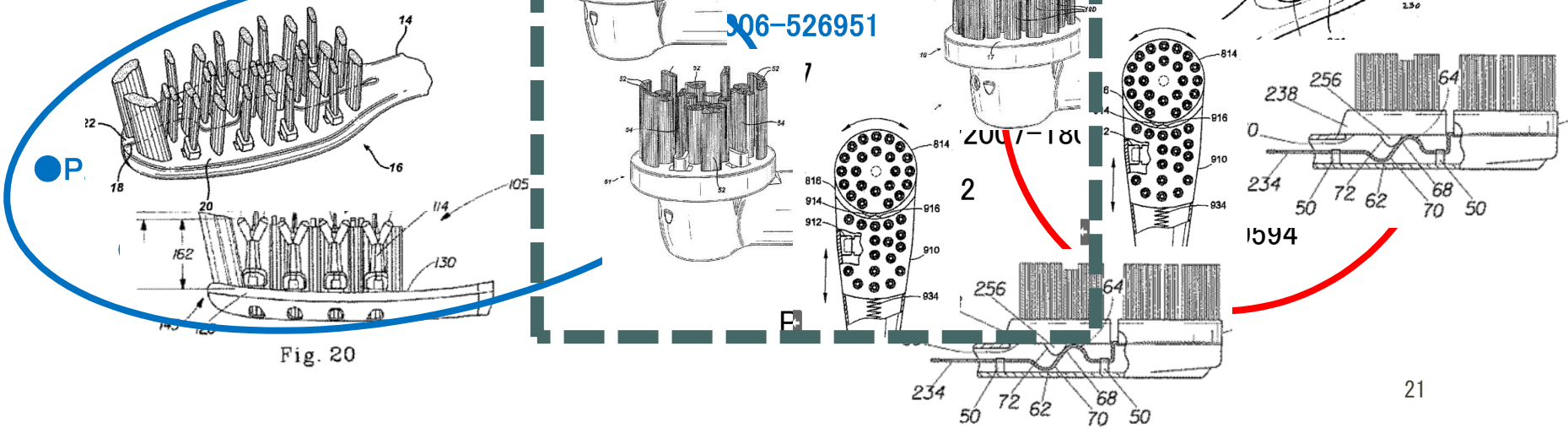


Fig. 20

発表内容

- 検討に至った背景
- 用いたテスト集合
- テキスト情報による公報間類似度
- 技術分類を併用した場合の類似度
- **まとめ**
- **今後の課題**

まとめ

テキストマイニングによる公報間類似度マップの作成において、

・公報に多用される**ノイズを強制排除**して**専門用語**を抽出すると、**目視判断に近い**結果を得ることができた。

しかし、関連技術であっても使用される用語や語数が異なると非類似と計算されてしまった。そこで、・・・

・テキスト情報に**技術分類情報**を**併用**することにより
ほぼ目視判断に近いレベルにまでマップを改善する事ができた

今後の課題

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。そのため、

・他の技術分野でも同様の結果が得られるのか
・スケールUPした場合でも同様の精度が得られるか

等についても、今後、確認していく必要がある。

また、**専門用語の重み付け**に関しても、
・文書単位だけでなく、技術分類単位＊等でも評価したい。

＊Konishi, K.; Kitauchi; A. and Takaki, T. Invalidity Patent Search System of NTT DATA. Proceedings of the 4th NTCIR Workshop. 2004.

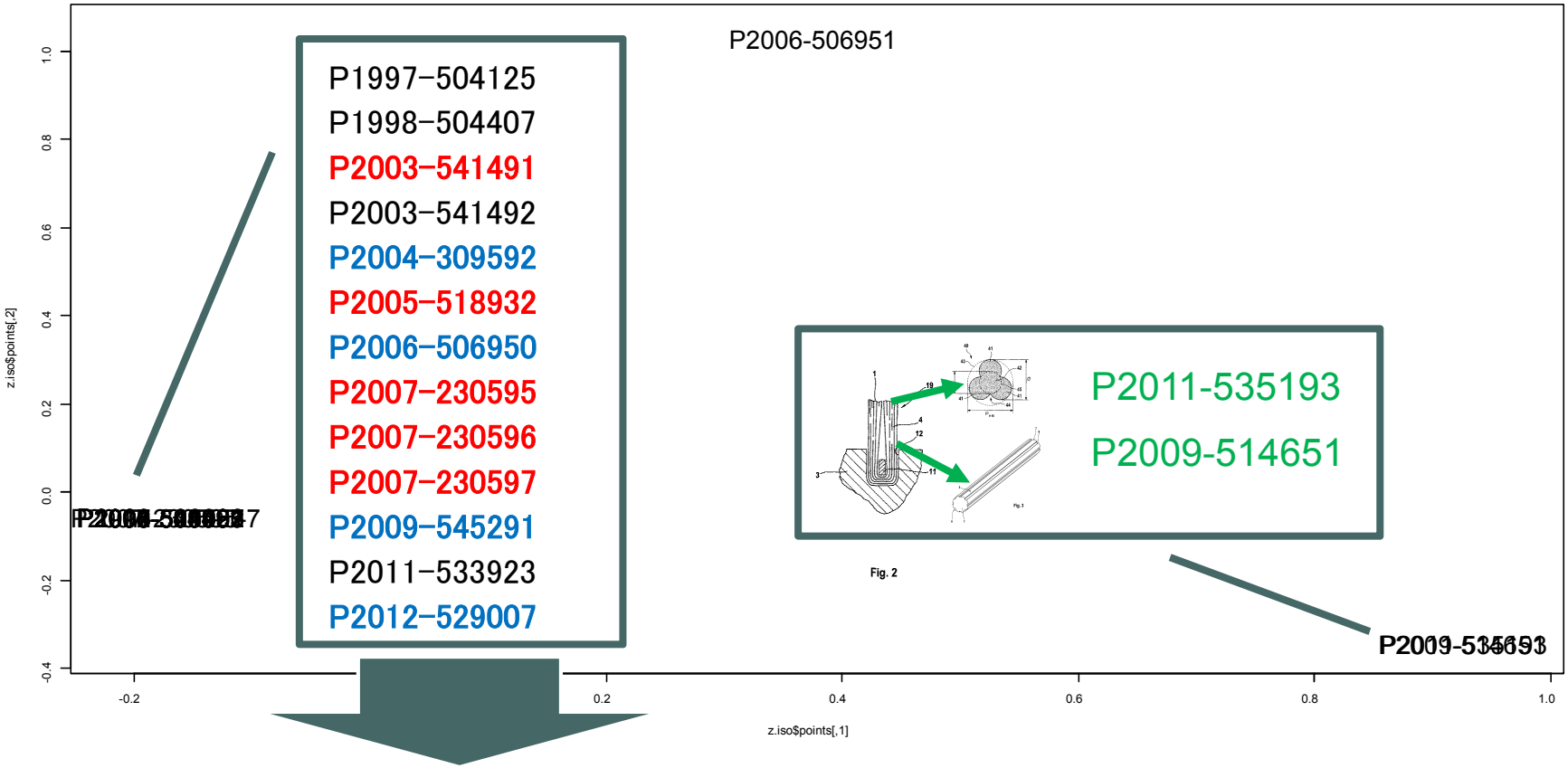
謝辞

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、研究会の皆様には多くの情報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。

参考資料

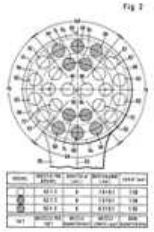
CPC分類説明文による公報間類似度

電動歯ブラシのヘッド部関連技術が一箇所に集まってしまった

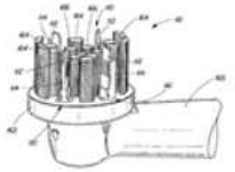


複雑な構成要素を有する駆動式ヘッド

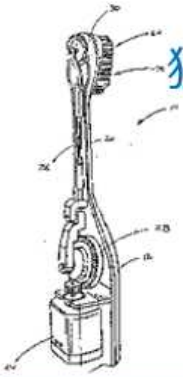
- P&G
- GLETTE
- BRAUN



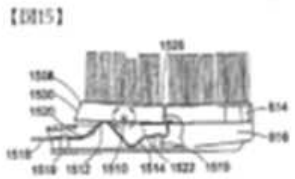
ブリストル部の
タフトに特徴



複数の素材からなる植毛部

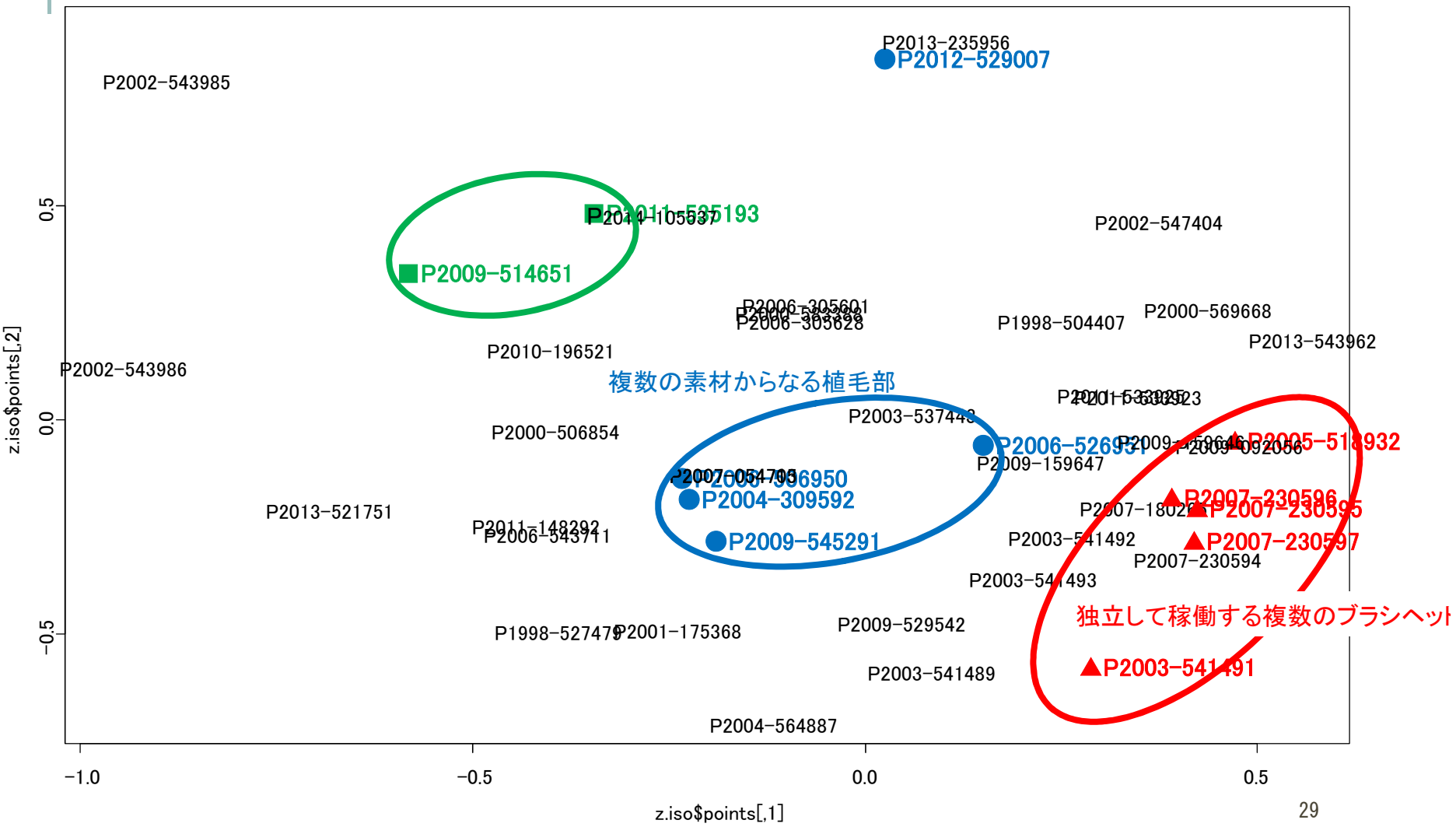


独立して稼働するブラシヘッド



テキスト情報
＋
技術分類併用

専門用語#ノイズ+CPC&IPC分類



＊DWPIファミリー、同引用・彼引用、同関連出願の各情報

専門用語#ノイズ+関連文献情報＊