A13

自動テキスト分類への挑戦

○花王株式会社　安藤俊幸
オリンパスメディカルシステムズ株式会社　中西昌弘
ユーエムジー・エービーエス株式会社　道中孝徳
富士フイルム株式会社　多田幸輔

発表内容

1．はじめに

2．重要な概念（専門用語）の抽出
・専門用語－文書行列作成

3．特許公報間の関係の解析・可視化
・多次元尺度法によるクラスタリング
・パテントインテグレーションによるクラスタリング検討

4．公報（文書）の自動分類
・カテゴリーマッチングによる自動分類
・自己組織化マップ

5．まとめ

はじめに

背景

- アジアの特許情報活用の重要性UP
- ビジネスのグローバル化→多言語対応
- 特許調査の効率化手法が求められている

目的

多言語の膨大な特許情報の中からユーザーにとって重要な情報を迅速に抽出、活用できる特許情報の分析・評価支援手法の開発

(1) 検索結果の大まかな把握（全体像の俯瞰、分類体系構築支援）
(2) 特許調査上の重要性に基づいた必要な観点に自動分類支援
(3) 自動分類結果の戦略的利用
（商用ＡＳＰ型特許ＤＢ、市販パテントマップソフトとの連携）

解析手法検討用データセット（母集団）の作成

対象：インクジェットカートリッジ

使用DB：Thomson Innovation
検索期間：1993.01.01-2010.12.31
検索日：2011.07.04

母集団の抽出条件

集合No.	Search Query	Collections	Results	備考
1	FTC=(2C056KC00 or 2C056KC01 or 2C056KC02 or 2C056KC04 or 2C056KC05 or 2C056KC06 or 2C056KC07 or 2C056KC09 or 2C056KC10 or 2C056KC11 or 2C056KC12 or 2C056KC13 or 2C056KC14 or 2C056KC15 or 2C056KC16 or 2C056KC17 or 2C056KC18 or 2C056KC20 or 2C056KC21 or 2C056KC22 or 2C056KC23 or 2C056KC24 or 2C056KC25 or 2C056KC26 or 2C056KC27 or 2C056KC30) AND DP>=(19930101) and DP<=(20101231);	JP App	8846	Fターム：インクタンク
2	1 and DWPI Family Members = US2 DWPI Family Members = CN		1202	Excelマクロ
3	2 and CN重複除去		768	Excelマクロ
4 4	CN1558829A CN1359334A			除去 除去
5	3 not 4		766	

JPのFターム（インクタンク）の集合でUS公開 and CNがある公報を抽出し
CNの重複を除去、ファミリー数が異常に多い2件を除去した

JP,US公開,CN各々766件を検討用母集団とした。

パテントマップEXZによる予備検討
・ランキング解析
・マトリックス解析
・時系列推移

テキストマイニングによる解析
・自動分類検討
・クラスタリング解析
・ネットワーク解析
・パテントインテグレーション検討

専門用語－文書行列作成

termmi を使用して各特許文書毎に専門用語を抽出する。
専門用語は各文書ごとに重要度付きで抽出されファイルに出力される。

termmiによる専門用語抽出例

1994-040044.txt

インク	28841.77
供給部インク室	16984.09
インク室	15414.73
記録ヘッド	12414.78

文書により同じ専門用語でも**重要度（重み）**が異なる。

1994-262774.txt

インク	129787.99
プリントヘッド	18409.37
吐出口	13881.00
吐出	13070.02

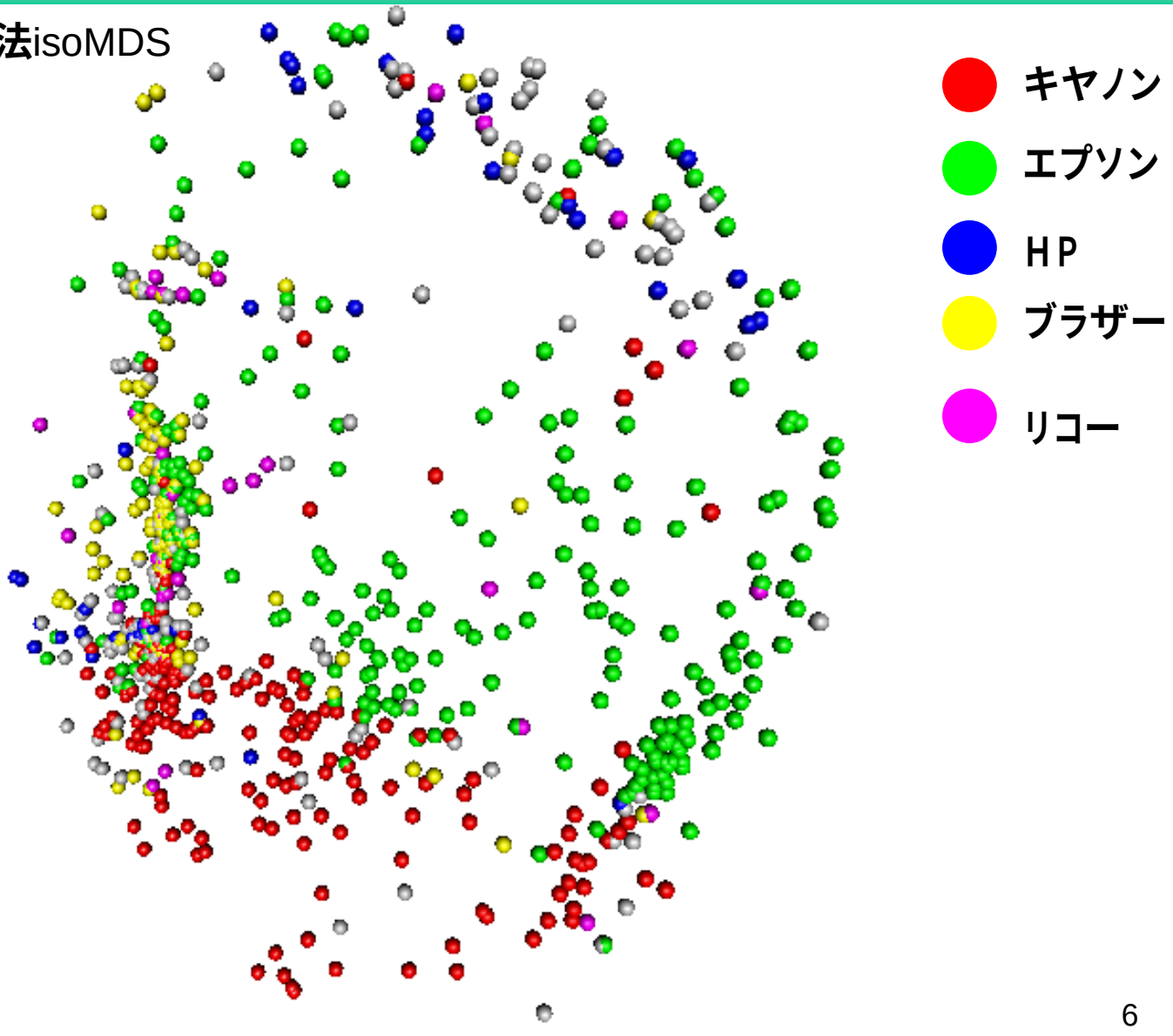
文書間の類似度を計算する

専門用語－文書の統計情報

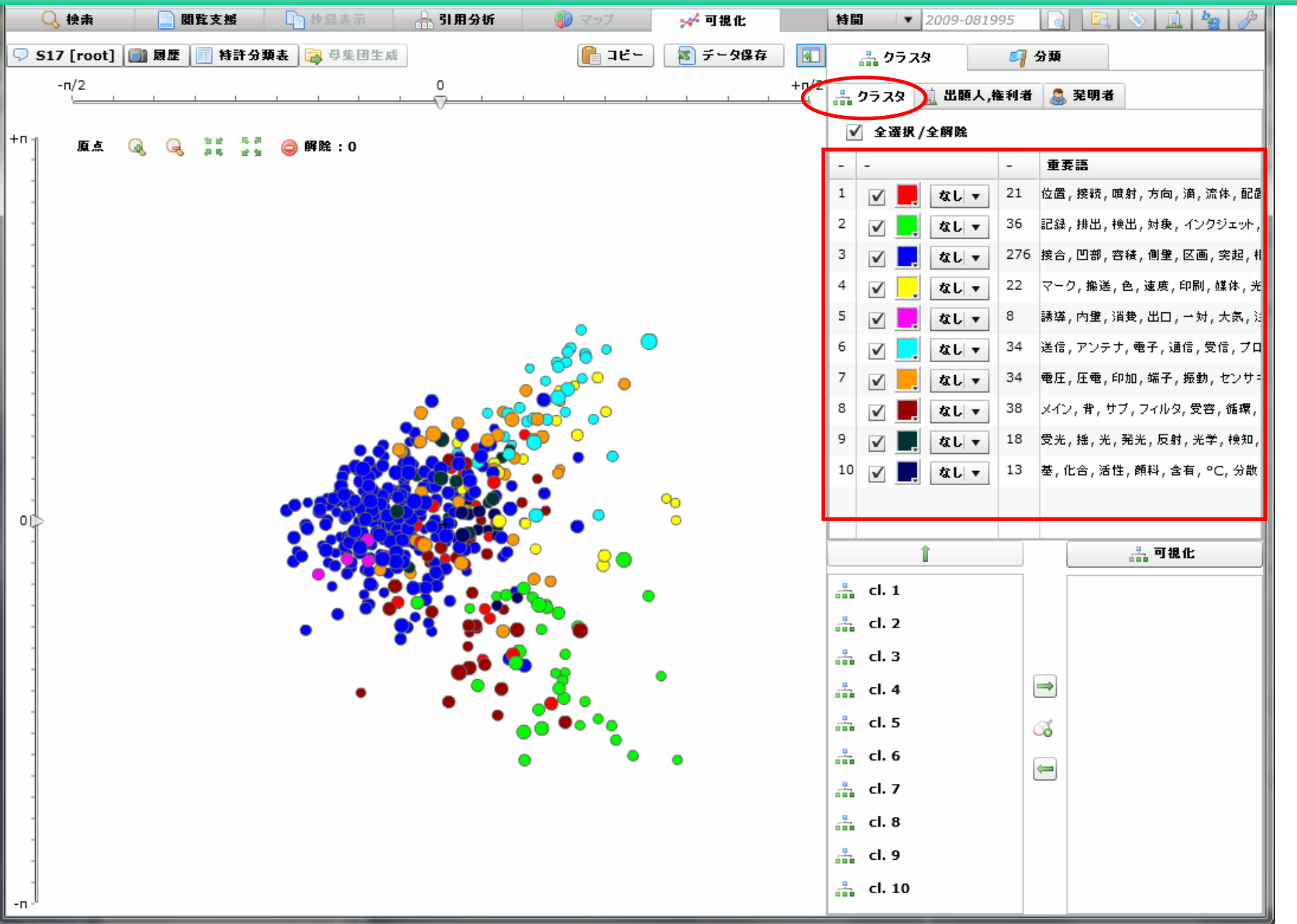
重要度の合計、平均、標準偏差、文書数(DF)、文書リスト

多次元尺度法によるクラスタリング（出願人）

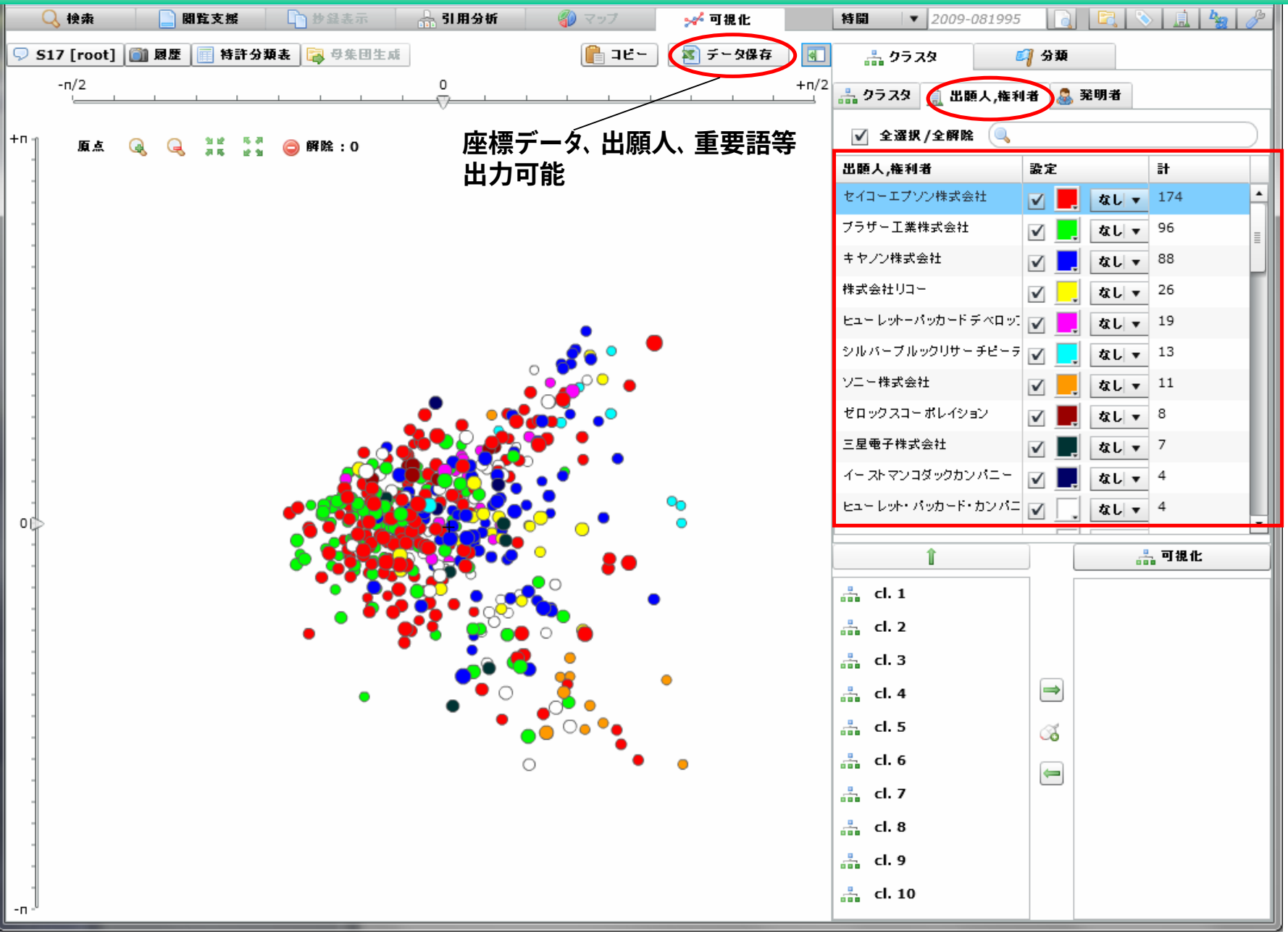
非計量多次元尺度法isoMDS



パテントインテグレーションによるクラスタリング検討（インクジェットカートリッジ）



パテントインテグレーションによるクラスタリング検討（インクジェットカートリッジ）



パテントインテグレーションの出力データ

文献番号	発明の名称	出願日	公開日	出願人,権利者	発明者	重要語	軸1	軸2
特表2008-513253	液滴付着装置における流体の取扱	2005/9/16	2008/5/1	フジフイルムテ	パリフカ, ロ	管,流体,バルブ,液,滴,付着,アクチュエータ	-0.06805	-0.14371
特開2010-076301	液体吐出装置	2008/9/26	2010/4/8	ブラザー工業株	平比呂志,遠	液体,排気,吐出,タンク,供給,ポンプ,排出,弁	-0.1339	-0.35188
特開2009-154328	液滴吐出ヘッド及びこれを備えた画	2007/12/25	2009/7/16	富士ゼロックス	浜崎聡信,平	ノズル,液体,防止,流体,循環,圧力,液,滴,イ	-0.18018	-0.30356
特開2006-137018	インクカートリッジ及び該カートリッ	2004/11/10	2006/6/1	株式会社リコー	勝山悟朗	部材,フィルム,縁,辺,口金,ケース,連結,パ	0.138671	0.078319
特表2005-517553	掃引プリンター付き携帯電話	2003/2/12	2005/6/16	シルバーブルッ	シルバーブ	印刷,インク,速度,吐出,キャップ,データ,光学	-0.38977	0.03564
特開2010-012602	パッケージ、インクカートリッジ	2008/6/30	2010/1/21	ブラザー工業株	中村宙健,服	部材,インク,包装,タンク,収容,アダプタ,圧力	0.052724	-0.10861
特開2010-012603	インクカートリッジ用のアダプタ	2008/6/30	2010/1/21	ブラザー工業株	中村宙健,服	カートリッジ,インク,アダプタ,収容,開口,部材	0.122294	0.02366
特開2005-225216	液体噴射装置	2004/3/30	2005/8/25	セイコーエプソ	岩崎充孝	液体,廃液,噴射,加,圧,空気,収容,部材,キャ	0.049407	-0.15729
特開2007-196674	液体収納容器	2006/12/22	2007/8/9	キヤノン株式会	小倉英幹,林	部材,収納,液体,ばね,面部,容器,位置,保護	0.067372	0.058427
特開2009-028972	液体吐出装置及び画像記録装置	2007/7/26	2009/2/12	ブラザー工業株	近本忠信	吐出,液体,ヘッド,位置,キャップ,インク,一方	-0.19247	-0.15043
特開2009-056679	インク容器、及びインク容器の製造	2007/8/31	2009/3/19	ブラザー工業株	服部信吾,林	インク,壁,位置,浮力,開口,フレーム,容器,対	0.211148	0.020547
特開2008-213465	流体収容容器、流体収容容器の再	2007/12/27	2008/9/18	セイコーエプソ	松山雅英,小	被覆,孔,フィルム,流体,収容,穴,容器,刃,部	0.231967	-0.00757
特開2005-145074	インクカートリッジ、インクカートリッ	2004/11/18	2005/6/9	3ティーサップ	ステイガー	オリフィス,インク,カートリッジ,シール,ハウ	0.115764	0.05432
特表2008-521659	プリントヘッドおよびプリントヘッドを	2005/12/2	2008/6/26	フジフイルムテ	モイニハン,	流体,射出,貯留,アセンブリ,導管,ノズル,シ	-0.07656	-0.15948
特開2009-292144	プリンタ	2009/3/4	2009/12/17	キヤノン株式会	岩田直宏,林	収容,カートリッジ,カバー,レバー,方向,弾性	0.081773	0.086939
特開2007-182055	液体収容容器	2006/8/12	2007/7/19	セイコーエプソ	勝村隆義	液体,内壁,収容,誘導,大気,消費,容器,連,供	0.127368	-0.06154
特開2005-212480	消耗物質用の容器用の交換式の	2005/1/28	2005/8/11	ヒューレット?パ	ブライアン・	デバイス,メモリ,物質,消耗,容器,交換,本体	-0.02186	0.119141
特開2009-285889	液体収容体	2008/5/27	2009/12/10	セイコーエプソ	鰐部晃久	液体,収容,分離,連,気泡,半径,通路,検出,設	0.061904	-0.00783
特開2009-241585	液体検出装置及びそれを用いた液	2009/2/9	2009/10/22	セイコーエプソ	鰐部晃久	流出,配置,センサ,液体,バッファ,部材,孔,下	0.109591	-0.01853
特開2008-012818	液体収納容器および記録装置	2006/7/6	2008/1/24	キヤノン株式会	瀧本雅文	液体,収納,データ,書換,アンテナ,領域,高周	-0.10694	0.186906
特開2008-037016	液体収容容器の製造方法	2006/8/8	2008/2/21	セイコーエプソ	岩室猛,木林	収容,体内,検出,導入,残,排出,流入,減圧,弁	0.027701	-0.09277
特開2007-245430	印刷装置、カートリッジおよび印刷	2006/3/14	2007/9/27	セイコーエプソ	小嶋輝人	手順,格納,情報,カートリッジ,収容,信号,前	-0.19517	0.224455
特表2005-515101	液滴付着装置	2003/1/16	2005/5/26	ザー・テクノロ	テムプル、	列,ノズル,プリント,ヘッド,方向,供給,チャン	-0.11893	-0.11358
特開2008-179804	再充填用インク及びインクカートリ	2007/12/26	2008/8/7	株式会社リコー	坂内昭子,後	混合,残留,充填,体積,径,保存,カートリッジ	-0.02361	-0.02566
特表2005-528237	インテリジェントなインク・カートリッ	2002/4/28	2005/9/22	プリント?ライト	ピーターチ	インク,カートリッジ,パーセンテージ,マイク	-0.11365	0.134045
特開2008-044190	液体注入方法及び液体収容容器	2006/8/11	2008/2/28	セイコーエプソ	品田聡,宮崎	液体,内壁,注入,連,大気,収容,誘導,消費,弁	0.164693	-0.059
特開2006-224394	液滴吐出装置の制御方法、液滴吐	2005/2/16	2006/8/31	セイコーエプソ	臼田秀範,林	描画,機能,滴,吐出,ヘッド,加,圧力,領域,タ	-0.28946	-0.36514
特開2005-342934	液体タンクおよび該液体タンクが搭	2004/6/1	2005/12/15	キヤノン株式会	小川将史,林	液体,圧接,収容,導入,吸収,連,供給,部材,設	0.009725	-0.07266
特開2008-044191	液体収容容器	2006/8/11	2008/2/28	セイコーエプソ	鰐部晃久	対向,外面,液体,接点,本体,容器,縁,接続,収	0.067876	0.117319
特開2009-096055	液体検出装置及びそれを用いた液	2007/10/16	2009/5/7	セイコーエプソ	鰐部晃久	液体,センサベース,本体,ケース,センサチ	0.076801	0.125627
特開2007-237715	半導体装置、インクカートリッジ及び	2006/3/13	2007/9/20	セイコーエプソ	橋元伸晃	インク,導,半導体,接点,電極,電気,収容,検	-0.17002	0.208716
特開2008-273114	液体収容容器、液体収容容器の再	2007/5/2	2008/11/13	セイコーエプソ	上原保直,服	フィルム,液体,穴,接合,収容,容器,封,止,孔	0.252858	0.031795
特開2006-188044	液体収納容器および該液体収納容	2005/11/9	2006/7/20	キヤノン株式会	小倉英幹,河	液体,収納,弾性,面,容器,係,圧縮,合,保持,	0.103703	0.132904

一部抜粋　　特徴：形態素レベル

インクジェットカートリッジ自動分類検討

1．検索モデルを応用した自動分類

①ブーリアンモデル
- キーワード検索　⇒　単語のマッチング（BOWモデル（Bag of Words））
- 複合検索モデル（特許分類＋キーワード）
- 近接演算モデル

②ベクトル空間モデル
- vector_space.plスクリプト（termmiに付属する文書間の類似度計算プログラム）
- **非計量多次元尺度法**を用いた**クラスタリング**

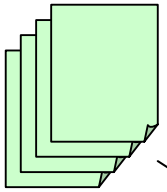
2．機械学習を応用した自動分類

- **自己組織化（SOM）マップ**（教師なし）
- **学習ベクトル量子化**
- 機械学習（教師あり）、例えばSVM（サポートベクターマシン）

テキストの自動分類とクラスタリング

自動分類

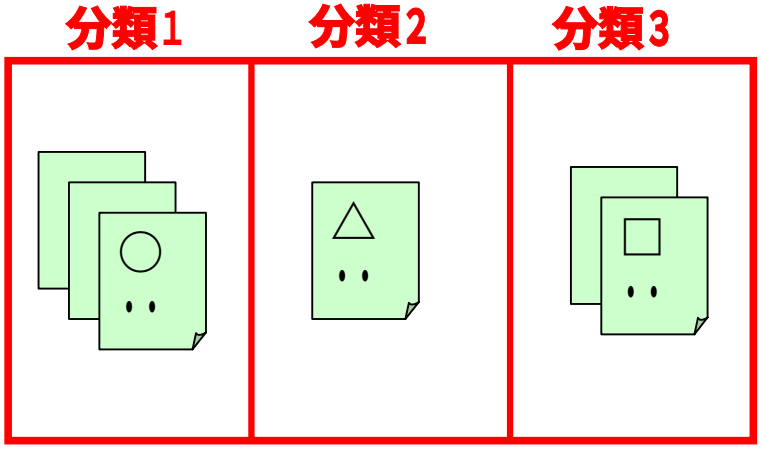
文書集合



カテゴリゼーション(注)

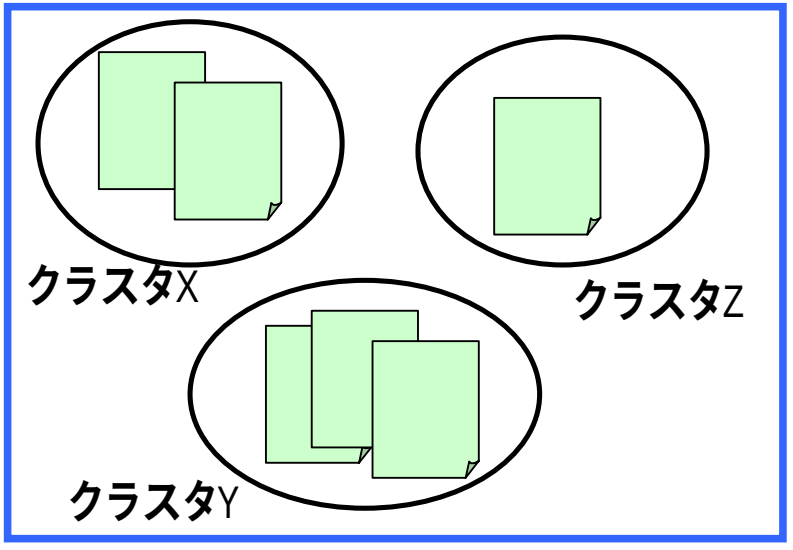
特徴語による分類表

分類1	分類2	分類3
○○○	△△△	□□□
○‥	△‥	□‥
○‥	△‥	□‥
○‥	△‥	□‥



あらかじめ決めたカテゴリに振り分ける

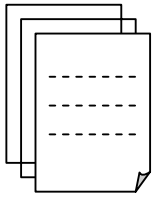
クラスタリング



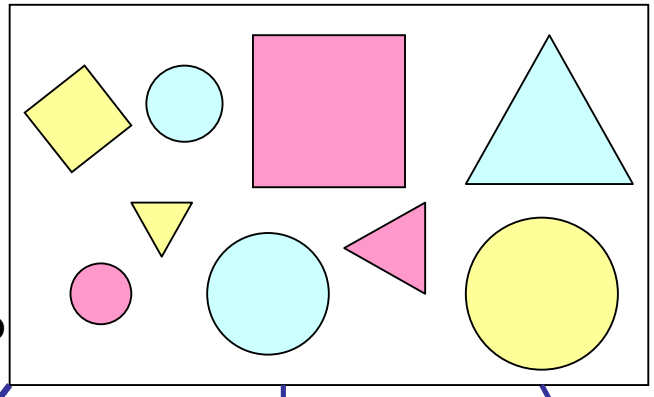
何らかの類似度(観点の)で似た文書をまとめる

(注) クラシフィケーション

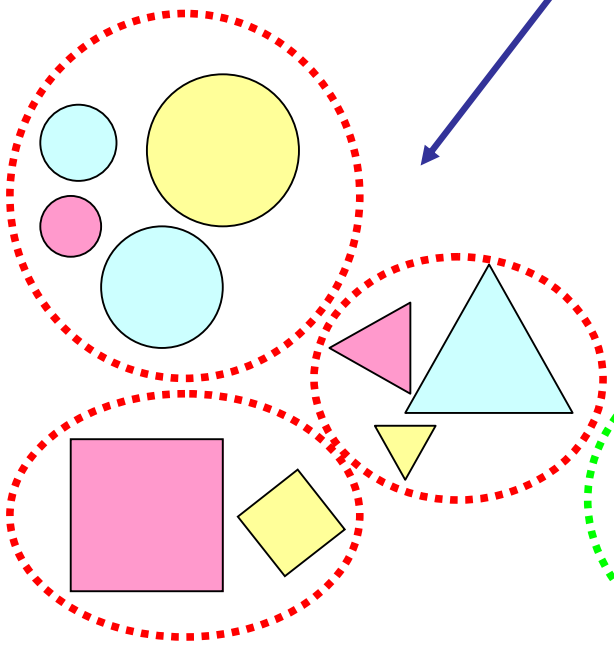
クラスタリングとは



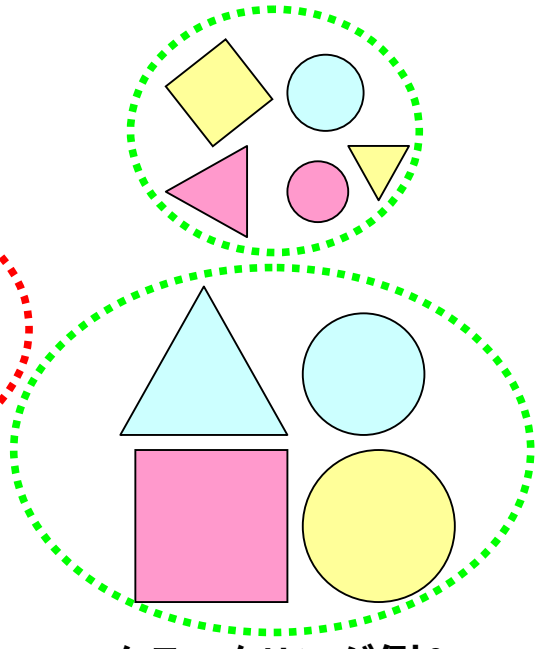
特許文書集合を文書間の
何らかの**類似度**に従って、
いくつかのグループに分ける



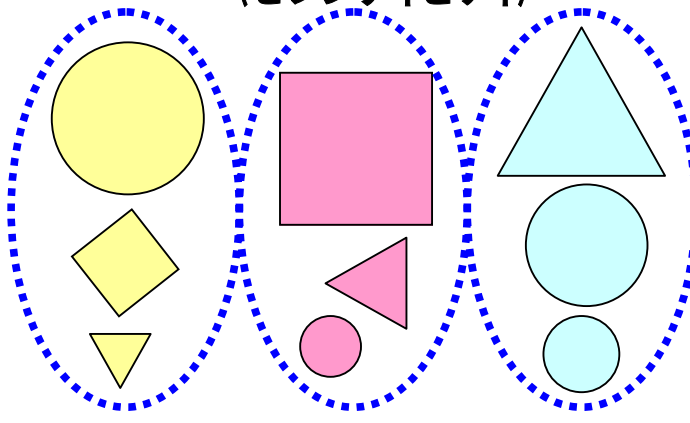
- **観点**によりクラスタリング結果が異なる（デッタッチメント）
- **類似度**の設定方法が多様（数値化方法が様々）
- 文書データをn次元ベクトルで表現
- クラスタリングには厳密な正解はない
- 人が行うデータ分析**支援**（気付きのためのツール）（セレンディピティ）



クラスタリング例1
観点：**形状**



クラスタリング例2
観点：**サイズ**



クラスタリング例3
観点：**カラー**

カテゴリーマッチングによる自動分類

考え方

カテゴリー：請求項の最後のターム

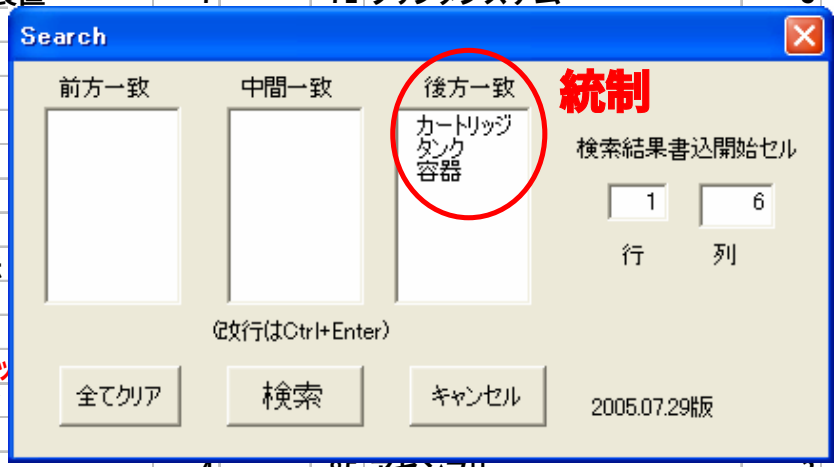
①請求の範囲（全請求項）を各請求項に分割
②各請求項を文に分割（「。」で分割）
③最初の文を末尾からスキャンして漢字、カタカナ部分を抽出する
⑤抽出したカテゴリーのランキングを作成
⑥カテゴリー自動分類用LUT（参照テーブル）編集
⑦分類表読込（「K03_カテゴリー分類読込」マクロ）
⑧自動分類実行（「K03_カテゴリー分類」マクロ）
①～③は「K03_カテゴリー抽出全請求項」マクロを使用

例1 画像形成装置本体に着脱自在に装着されるインクカートリッジであって、インクを収容する内袋部と、前記内袋部を内部に収容し、気体が導入されることで前記内袋部を加圧して前記内袋部内のインクをインクカートリッジ外に供給する外袋部と、前記外袋部を収容するカートリッジケースと、を備え、前記外袋部内に気体を導入したときに前記外袋部が前記カートリッジケースの内壁面に当接することを特徴とするインクカートリッジ。【請求項2】前記外袋部

例2 吐出口からインクを吐出するための熱エネルギーを発生する発熱部が設けられたサーマルインクジェットヘッドを具備し、且つ、インクを収容するインクカートリッジであって、該発熱部が該インクと接する面に、珪素の、酸化物、窒化物及び炭化物からなる群より選ばれる少なくとも一つを含有する保護層を有し、該インクが、該保護層を溶解する物質と下記一般式（1）で表される化合物とを含有し、且つ、該下記一般式（1）で表される化合物のインク中の含有量X（質量%）が1≦X≦30を満足することを特徴とするインクカートリッジ。（Aは、窒素原子とカルボニル基とR3と共に環を形成する、炭素数が1～4のアルキレン基又はアルケニレン基を表す。R1とR4は、それぞれ独立に、水素原子、水酸基、置換若しくは非置換のアルキル基、置換若しくは非置換のアルケニル基、置換若しくは非置換のアシル基、カルバモイル基、置換若しくは非置換のカルボキシル基、及び、置換若しくは非置換のスルホニル基のいずれかを表す。R2は、Aの任意の炭素原子に結合する基であり、水素原子、水酸基、置換若しくは非置換のアルキル基、置換若しくは非置換のアルケニル基、置換若しくは非置換のアシル基、カルバモイル基、置換若しくは非置換のカルボキシル基、及び、置換若しくは非置換のスルホニル基のいずれかを表し、nは0～4の整数である。R3は、炭素又は窒素原子を表す。）【請求項2】

カテゴリー（請求項の最後のターム）ランキング

No.	カテゴリー	件数	No.	カテゴリー	件数	No.	カテゴリー	件数
1	インクカートリッジ	137	31	液体吐出方法	9	61	パッケージ	4
2	インクジェット記録装置	84	32	インク供給システム	9	62	画像記録方法	4
3	製造方法	63	33	液体充填方法	8	63	記録用インク	4
4	インクタンク	57	34	流体噴射装置	8	64	インクジェット装置	4
5	記録装置	54	35	インクジェットカートリッジ	7	65	記録媒体	4
6	方法	50	36	画像記録装置	7	66	半導体装置	4
7	液体収納容器	46	37	記録ユニット	7	67	液体消費装置	4
8	液体収容容器	45	38	印刷方法	7	68	装着方法	3
9	液体噴射装置	40	39	システム	7	69	ユニット	3
10	インクジェットプリンタ	39	40	電子機器	7	70	使用	3
11	液体吐出装置	30	41	液体吐出記録装置	7	71	プリンタシステム	3
12	液体容器	29	42	記録液				
13	画像形成装置	28	43	液滴吐出装置				
14	印刷装置	23	44	回路基板				
15	液体収容体	21	45	インク				
16	インクジェット記録方法	18	46	再生方法				
17	液体供給システム	17	47	液体検出装置				
18	液体カートリッジ	16	48	液体注入方法				
19	インク供給装置	15	49	インク供給方法				
20	制御方法	15	50	容器				
21	液体吐出ヘッド	14	51	電気光学装置				
22	液体供給装置	14	52	プリントカートリッ				
23	装置	14	53	インクスティック				
24	インクジェット記録ヘッド	13	54	シール方法				
25	インクジェット式記録装置	13	55	印刷流体容器	4	85	アセンブリ	3
26	カートリッジ	11	56	インク充填方法	4	86	インクジェットヘッドカートリッジ	3
27	インク容器	11	57	プリントヘッド	4	87	基板	3
28	ヘッドカートリッジ	10	58	インクジェットプリント装置	4	88	記憶装置	3
29	プリンタ	9	59	液体吐出カートリッジ	4	89	インクジェットプリントヘッド	3
30	記録ヘッド	9	60	流体収容容器	4	90	インクセット	3



一部抜粋

・物の発明のカテゴリーはかなり良く抽出できている
・方法、製造方法は更に工夫が必要

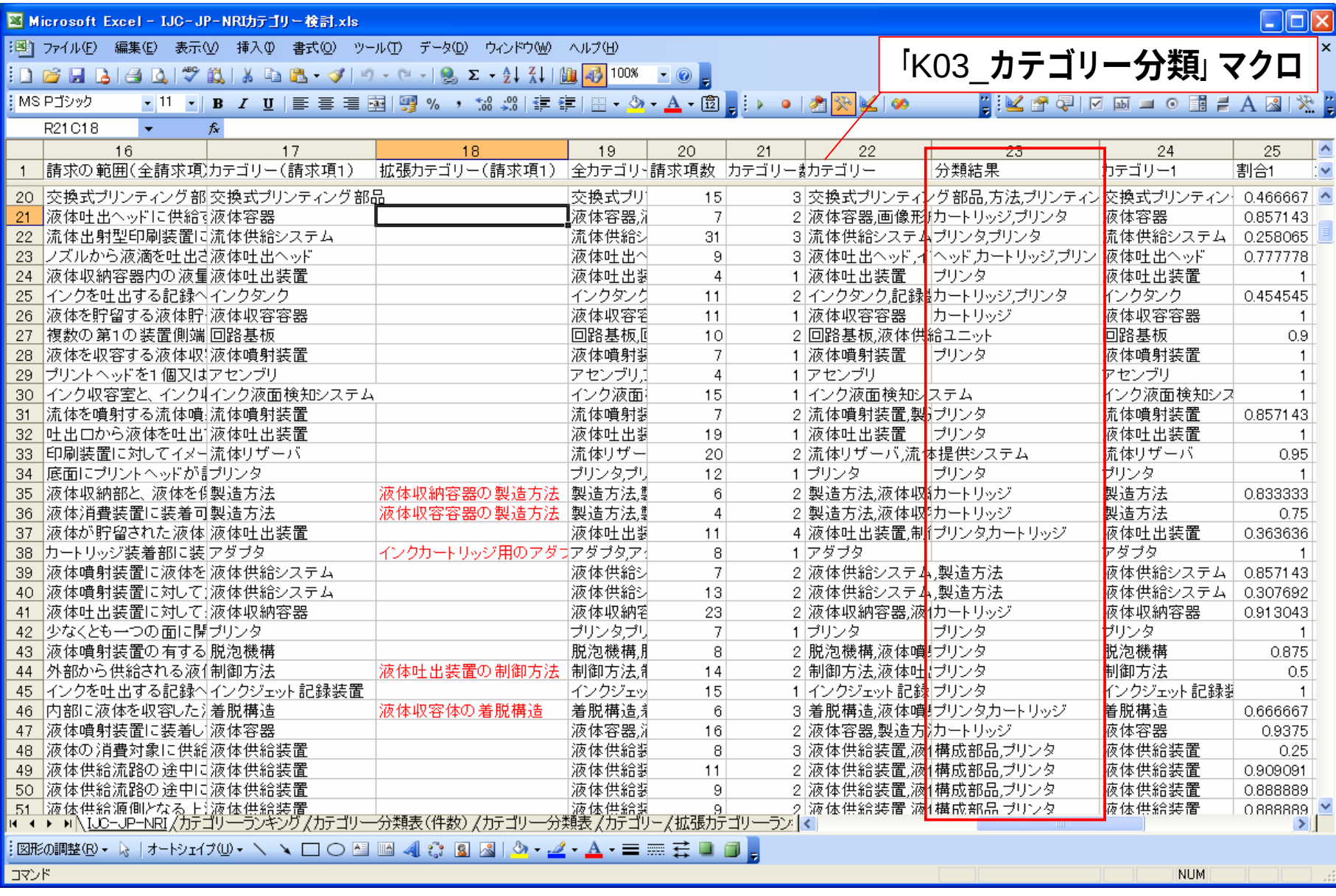
カテゴリー自動分類用LUT（参照テーブル）

カートリッジ	プリンタ	構成部品	ヘッド	記録方法
インクカートリッジ	インクジェット記録装置	インク供給装置	液体吐出ヘッド	インクジェット記録方法
インクタンク	記録装置	液体供給装置	インクジェット記録ヘッド	画像記録方法
液体収納容器	液体噴射装置	装置	記録ヘッド	印刷方法
液体収容容器	インクジェットプリンタ	液体検出装置	プリントヘッド	デジタル画像印刷方法
液体容器	液体吐出装置	電気光学装置	液体吐出記録ヘッド	インクジェットプリント方法
液体カートリッジ	画像形成装置	半導体装置	インクジェットプリントヘッド	
カートリッジ	印刷装置	液体消費装置	液体噴射ヘッド	
インク容器	インクジェット式記録装置	記憶装置	液滴吐出ヘッド	
ヘッドカートリッジ	プリンタ	液体使用装置	液体記録吐出ヘッド	
インクジェットカートリッジ	流体噴射装置	インク残量検出装置	インクジェットヘッド	
容器	画像記録装置	弁装置	インク噴射式プリントヘッド	
プリントカートリッジ	液体吐出記録装置	移動通信装置	流体噴射器ヘッド	
印刷流体容器	液滴吐出装置	機能液供給装置	ページ幅インクジェット印字ヘッド	
液体吐出カートリッジ	インクジェットプリント装置	液体充填装置	ヘッドキット	
流体収容容器	インクジェット装置	カートリッジ収容装置	液体吐出ヘッドキット	
液体収納カートリッジ	流体吐出装置	流体供給装置	ヘッドモジュール	
交換式インク容器	マーキング装置	インク充填装置	印字ヘッド組立体	
液体タンク	印字装置	液滴付着装置	ページ幅インクジェット印字ヘッド組立体	
インクジェットヘッドカートリッジ	インクジェット画像記録装置	画像処理装置	インクジェットヘッドユニット	
記録液カートリッジ	液滴噴射装置	製造装置		
消耗品容器	吐出装置	インクエンド検出装置		
流体噴射カートリッジ	プリンタ装置	判定装置		
流体滴噴射カートリッジ	プリンター	泡－液体／気体分離装置		
リフィルユニット用インクカートリッジ	インクジェトプリンタ	記録ヘッド装置		
記録装置用インクカートリッジ	プリンタ制御装置	制御装置		
消耗物質用容器	インクジェット画像形成装置	吸入装置		
インクジェット記録カートリッジ	流体射出型印刷装置	制御用装置		
インク・カートリッジ	流体出射型印刷装置	開放装置		
プリンタ用インクカートリッジ	写真アルバムページ作成装置	レバー形状保持装置		
インクジェット記録装置用インクカートリッジ	デジタル画像印刷装置	廃液回収装置		
インクジェット記録ヘッド一体型インクタンク	描画装置	インク循環装置		
インクジェット記録用インクタンク	インクジェット式プリンタ	液体カートリッジ着脱装置		
再充填流体収容容器	インクジェットプリンタ装置	カートリッジ着脱装置		
液体供給容器	プリンタシステム	再生装置		
印刷カートリッジ	プリンタ本体	インク貯留装置		
インクタンクカートリッジ	プリンタユニット	印刷制御装置		
インクジェット記録装置用インクタンクカートリ	インクジェットプリンタユニット	排出装置		
インク供給容器		液体収容装置		
インク保持容器		インクカートリッジ判定装置		
インクジェットプリンタ用インクカートリッジ		インクカートリッジ保持装置		

「K03_カテゴリー分類**読込**」マクロ

一部抜粋

カテゴリーマッチングによる自動分類結果



	16	17	18	19	20	21	22	23	24	25
1	請求の範囲(全請求項)	カテゴリー(請求項1)	拡張カテゴリー(請求項1)	全カテゴリー	請求項数	カテゴリー数	カテゴリー	分類結果	カテゴリー1	割合1
20	交換式プリンティング部	交換式プリンティング部品		交換式プリ	15	3	交換式プリンティング部品,方法,プリンティン		交換式プリンティン	0.466667
21	液体吐出ヘッドに供給	液体容器		液体容器,液	7	2	液体容器,画像形	カートリッジ,プリンタ	液体容器	0.857143
22	流体出射型印刷装置に	流体供給システム		流体供給シ	31	3	流体供給システム	プリンタ,プリンタ	流体供給システム	0.258065
23	ノズルから液滴を吐出さ	液体吐出ヘッド		液体吐出ヘ	9	3	液体吐出ヘッド,イ	ヘッド,カートリッジ,プリン	液体吐出ヘッド	0.777778
24	液体収納容器内の液量	液体吐出装置		液体吐出装	4	1	液体吐出装置	プリンタ	液体吐出装置	1
25	インクを吐出する記録ヘ	インクタンク		インクタンク	11	2	インクタンク,記録装	カートリッジ,プリンタ	インクタンク	0.454545
26	液体を貯留する液体貯	液体収容容器		液体収容容	11	1	液体収容容器	カートリッジ	液体収容容器	1
27	複数の第1の装置側端	回路基板		回路基板,回	10	2	回路基板,液体供	給ユニット	回路基板	0.9
28	液体を収容する液体収	液体噴射装置		液体噴射装	7	1	液体噴射装置	プリンタ	液体噴射装置	1
29	プリントヘッドを1個又は	アセンブリ		アセンブリ,ア	4	1	アセンブリ		アセンブリ	1
30	インク収容室と、インク	インク液面検知システム		インク液面	15	1	インク液面検知シ	ステム	インク液面検知シス	1
31	流体を噴射する流体噴	流体噴射装置		流体噴射装	7	2	流体噴射装置,製	プリンタ	流体噴射装置	0.857143
32	吐出口から液体を吐出	液体吐出装置		液体吐出装	19	1	液体吐出装置	プリンタ	液体吐出装置	1
33	印刷装置に対してイメー	流体リザーバ		流体リザー	20	2	流体リザーバ,流体	提供システム	流体リザーバ	0.95
34	底面にプリントヘッドが	プリンタ		プリンタ,プリ	12	1	プリンタ	プリンタ	プリンタ	1
35	液体収納部と、液体を	製造方法	液体収納容器の製造方法	製造方法,製	6	2	製造方法,液体収	カートリッジ	製造方法	0.833333
36	液体消費装置に装着可	製造方法	液体収容容器の製造方法	製造方法,製	4	2	製造方法,液体収	カートリッジ	製造方法	0.75
37	液体が貯留された液体	液体吐出装置		液体吐出装	11	4	液体吐出装置,制	プリンタ,カートリッジ	液体吐出装置	0.363636
38	カートリッジ装着部に装	アダプタ	インクカートリッジ用のアダプ	アダプタ,ア	8	1	アダプタ		アダプタ	1
39	液体噴射装置に液体を	液体供給システム		液体供給シ	7	2	液体供給システム	,製造方法	液体供給システム	0.857143
40	液体噴射装置に対して	液体供給システム		液体供給シ	13	2	液体供給システム	,製造方法	液体供給システム	0.307692
41	液体吐出装置に対して	液体収納容器		液体収納容	23	2	液体収納容器,液	カートリッジ	液体収納容器	0.913043
42	少なくとも一つの面に開	プリンタ		プリンタ,プリ	7	1	プリンタ	プリンタ	プリンタ	1
43	液体噴射装置の有する	脱泡機構		脱泡機構,脱	8	2	脱泡機構,液体噴	プリンタ	脱泡機構	0.875
44	外部から供給される液	制御方法	液体吐出装置の制御方法	制御方法,制	14	2	制御方法,液体吐	プリンタ	制御方法	0.5
45	インクを吐出する記録ヘ	インクジェット記録装置		インクジェッ	15	1	インクジェット記録	プリンタ	インクジェット記録装	1
46	内部に液体を収容した	着脱構造	液体収容体の着脱構造	着脱構造,着	6	3	着脱構造,液体噴	プリンタ,カートリッジ	着脱構造	0.666667
47	液体噴射装置に装着し	液体容器		液体容器,液	16	2	液体容器,製造方	カートリッジ	液体容器	0.9375
48	液体の消費対象に供給	液体供給装置		液体供給装	8	3	液体供給装置,液	構成部品,プリンタ	液体供給装置	0.25
49	液体供給流路の途中に	液体供給装置		液体供給装	11	2	液体供給装置,液	構成部品,プリンタ	液体供給装置	0.909091
50	液体供給流路の途中に	液体供給装置		液体供給装	9	2	液体供給装置,液	構成部品,プリンタ	液体供給装置	0.888889
51	液体供給源側となる上	液体供給装置		液体供給装	9	2	液体供給装置,液	構成部品,プリンタ	液体供給装置	0.888889

カテゴリーマッチングによる自動分類（請求項1）

分類可能例

No.	カテゴリー（請求項1）	分類結果	件数
1	インクカートリッジ	カートリッジ	89
2	インクタンク	カートリッジ	41
3	液体収納容器	カートリッジ	39
4	液体収容容器	カートリッジ	30
5	インクジェット記録装置	プリンタ	21
6	液体容器	カートリッジ	19
9	液体収容体	カートリッジ	17
8	インクジェットプリンタ	プリンタ	17
10	液体噴射装置	プリンタ	15
11	画像形成装置	プリンタ	14
12	記録装置	プリンタ	13
14	液体吐出装置	プリンタ	12
15	液体吐出ヘッド	ヘッド	11
16	インク供給装置	構成部品	10
17	液体カートリッジ	カートリッジ	9
19	印刷装置	プリンタ	9
18	液体供給装置	構成部品	9
20	インク容器	カートリッジ	7
21	インクジェット記録ヘッド	ヘッド	7
23	インクジェット式記録装置	プリンタ	6
22	液体供給システム	構成部品	6
27	液体注入方法	インク充填方	5
25	液体検出装置	構成部品	5
26	装置	構成部品	5
28	インク供給システム	構成部品	5

カテゴリー件数（請求項1）	766	
分類可能件数	702	**92%**
方法	47	**6%**
製造方法	17	**2%**
定義で対応可能	10	**1%**
その他（分類不能）	11	**1%**

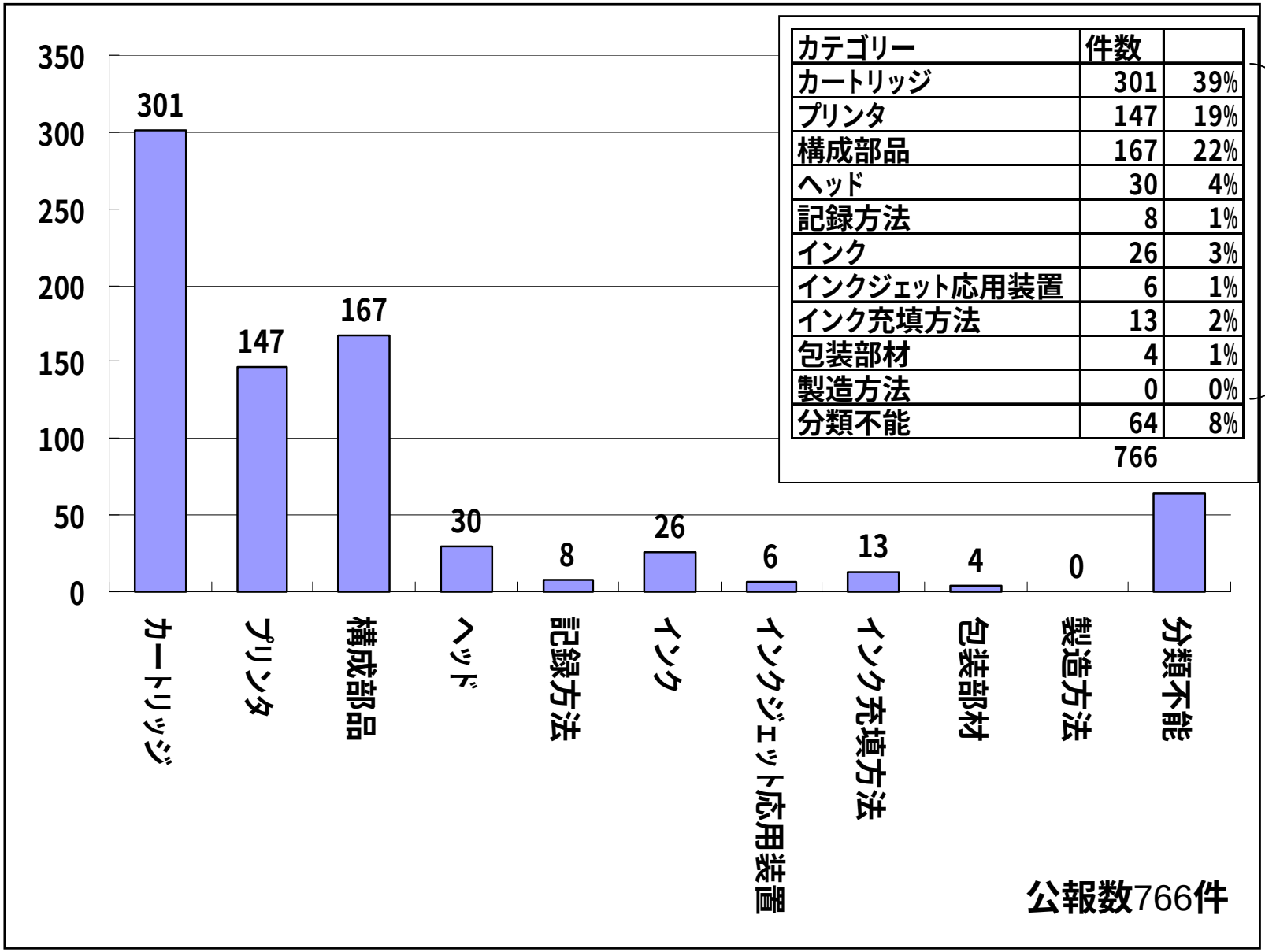
分類不能例（全数）

No.	カテゴリー（請求項1）	分類結果	件数	方法	定義
7	製造方法		17	17	
13	方法		12	12	
24	制御方法		5	5	
49	シール方法		3	3	
61	立体形半導体素子		2		2
63	システム		2		
65	リフィルユニット		2		2
173	無線通信機器		1		
174	保存方法		1	1	
175	保管形態		1		
176	インクジェットプリント用布帛		1		1
177	記録データ生成方法		1	1	1
178	検査方法		1	1	
179	バッファ		1		
180	部材		1		
181	構造		1		
182	ツール		1		
183	封止方法		1	1	
184	モジュール		1		
185	部品		1		
186	再利用方法		1	1	
187	可視剤量表示制御システム		1		1
196	インクエンド検出方法		1	1	1
197	液体消費状態検知方法		1	1	1
198	液体貯留手段		1		
199	通信方法		1		
200	着色材用カートリッジの真偽判定方法		1	1	1
201	装着方法		1	1	
			64	46	10

17

分類表定義で対応可能

カテゴリーマッチングによる自動分類結果（請求項1）



カテゴリー	件数	
カートリッジ	301	39%
プリンタ	147	19%
構成部品	167	22%
ヘッド	30	4%
記録方法	8	1%
インク	26	3%
インクジェット応用装置	6	1%
インク充填方法	13	2%
包装部材	4	1%
製造方法	0	0%
分類不能	64	8%
	766	

92% 分類可能

公報数766件

カテゴリーマッチングによる自動分類（全請求項）

分類可能例

No.	カテゴリー（全請求項）	分類結果	件数
1	インクカートリッジ	カートリッジ	137
2	インクジェット記録装置	プリンタ	84
4	インクタンク	カートリッジ	57
5	記録装置	プリンタ	54
7	液体収納容器	カートリッジ	46
8	液体収容容器	カートリッジ	45
9	液体噴射装置	プリンタ	40
10	インクジェットプリンタ	プリンタ	39
11	液体吐出装置	プリンタ	30
12	液体容器	カートリッジ	29
13	画像形成装置	プリンタ	28
14	印刷装置	プリンタ	23
15	液体収容体	カートリッジ	21
16	インクジェット記録方法	記録方法	18
17	液体供給システム	構成部品	17
18	液体カートリッジ	カートリッジ	16
20	インク供給装置	構成部品	15
21	装置	構成部品	14
22	液体吐出ヘッド	ヘッド	14
23	液体供給装置	構成部品	14
24	インクジェット式記録装置	プリンタ	13
25	インクジェット記録ヘッド	ヘッド	13
26	カートリッジ	カートリッジ	11
27	インク容器	カートリッジ	11
28	ヘッドカートリッジ	カートリッジ	10

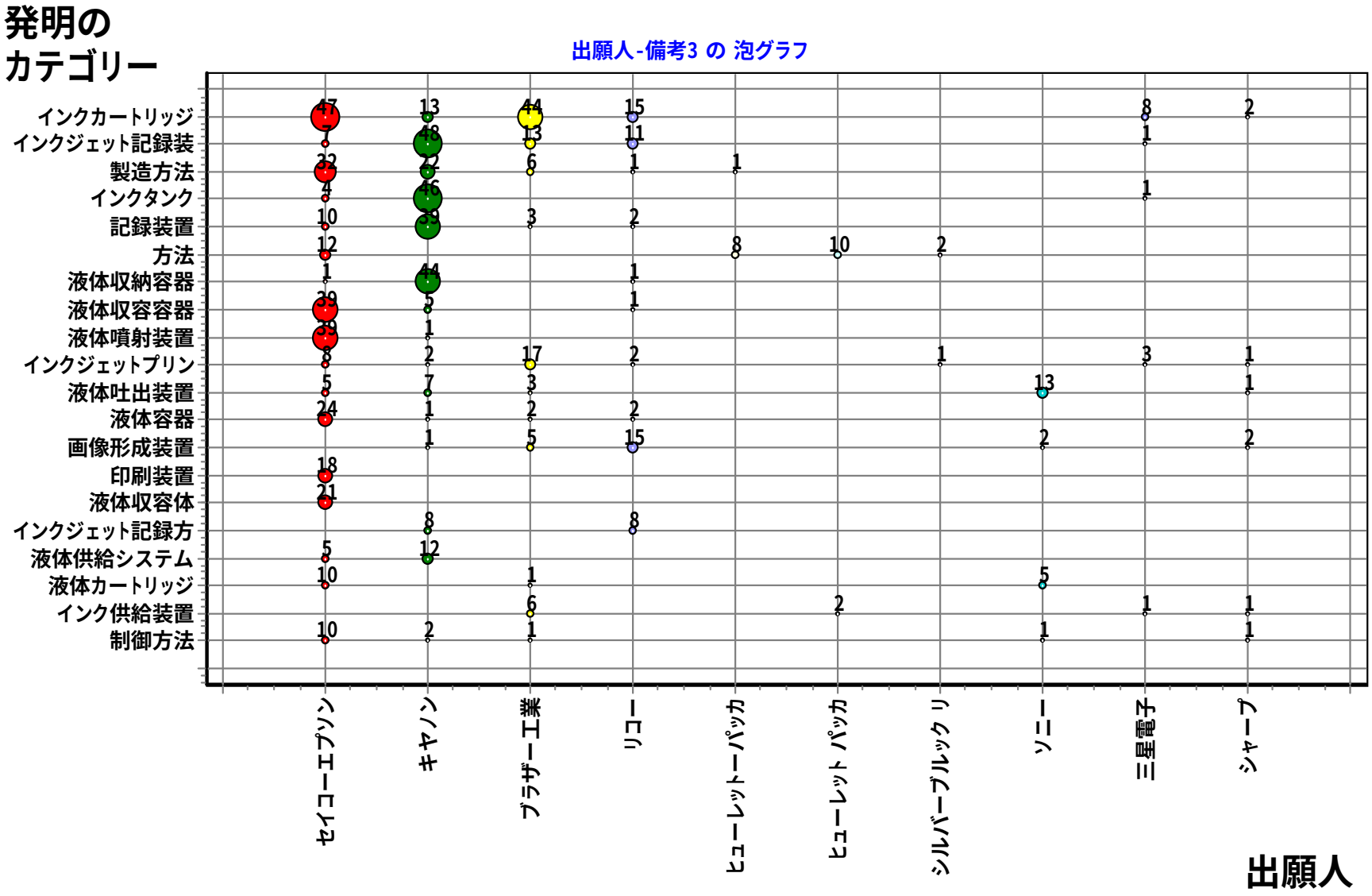
全カテゴリー数	560	
分類可能カテゴリー数	320	57%
カテゴリー件数（全請求項）	1763	
分類可能件数	1320	75%
全請求項数	11077	
全公報数	766	
一部分類可能公報数	750	98%

分類不能例（一部抜粋）

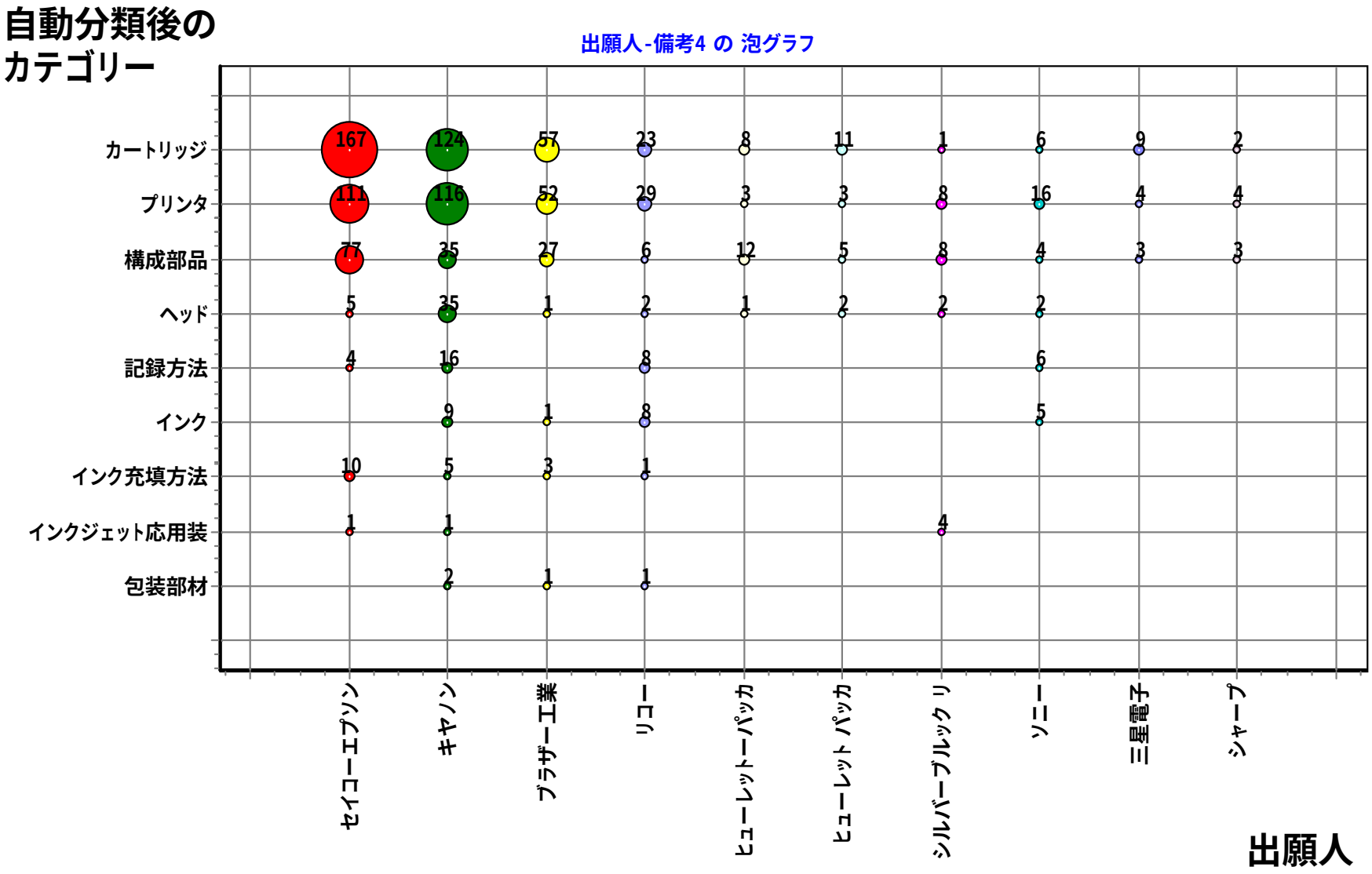
No.	カテゴリー（全請求項）	分類結果	件数
3	製造方法		63
6	方法		50
19	制御方法		15
35	電子機器		7
40	システム		7
47	再生方法		5
49	インク供給方法		5
57	記録媒体		4
65	シール方法		4
68	開封方法		3
69	インク注入方法		3
70	再充填方法		3
71	ユニット		3
73	コンピュータプログラム		3
74	使用		3
79	インク記録物		3
80	特徴		3
81	基板		3
85	装着方法		3
87	記録システム		3
90	プログラム		3
93	インクセット		3
94	リフィルユニット用ユニット本体		2
95	セット		2
97	インクカートリッジ製造方法		2

分類表定義で対応可能

出願人－カテゴリー抽出結果（全請求項：未統制）



出願人－カテゴリー自動分類結果（全請求項）



カテゴリー自動分類の課題

○○の製造法――――　「カテゴリー拡張抽出」オプションで対応
○○方法
○○する方法
○○を特徴とする□ □

液体収容体の製造方法　⇒　請求項1：成功
○○を特徴とする請求項1記載の製造方法
下位クレーム：失敗

上記「○○」の抽出　「係り受け」検討

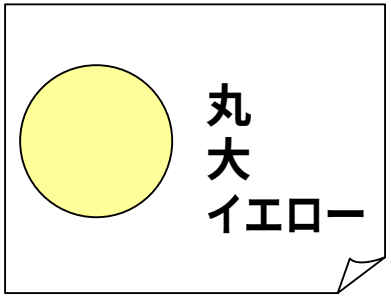
「課題、解決手段」の抽出への応用

分類表（分類体系）を作るのが大変　――→　（絶対）必要！

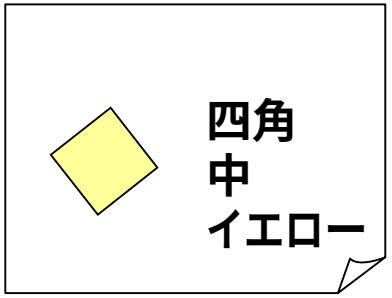
しかし、楽にできないか？

文書のクラスタリング実験 (多次元尺度法)

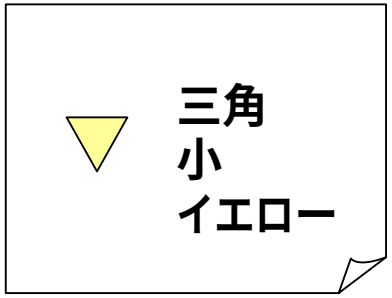
文書1　　ターム重み



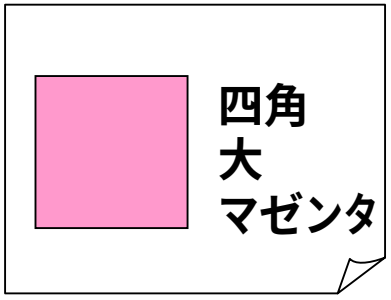
文書2　　ターム重み



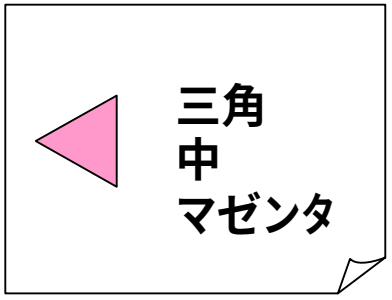
文書3　　ターム重み



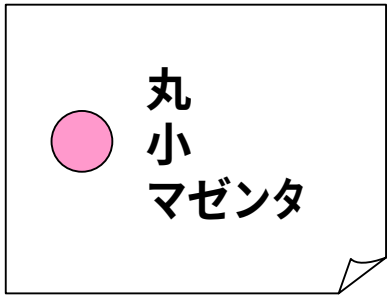
文書4



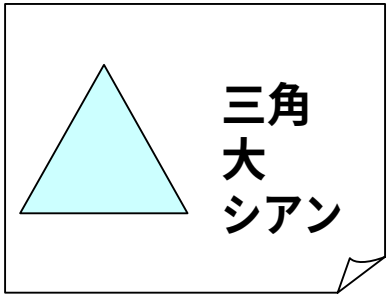
文書5



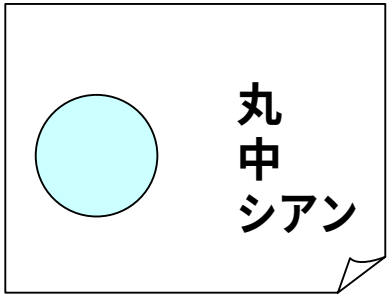
文書6



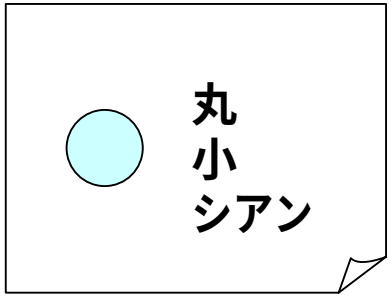
文書7



文書8



文書9



文書間相互の類似度と距離マトリックス

類似度マトリックス（⑦重み付きCosine係数）

（多次元尺度法）

文書番号	doc1	doc2	doc3	doc4	doc5	doc6	doc7	doc8	doc9
doc1	0	0	0	0	0	0	0	0	0
doc2	0.67	0	0	0	0	0	0	0	0
doc3	0.67	0.67	0	0	0	0	0	0	0
doc4	0.17	0.17	0	0	0	0	0	0	0
doc5	0	0.17	0.17	0.67	0	0	0	0	0
doc6	0.17	0	0.17	0.67	0.67	0	0	0	0
doc7	0.17	0	0.17	0.17	0.17	0	0	0	0
doc8	0.17	0.17	0	0	0.17	0.17	0.67	0	0
doc9	0.17	0	0.17	0	0	0.33	0.67	0.83	0

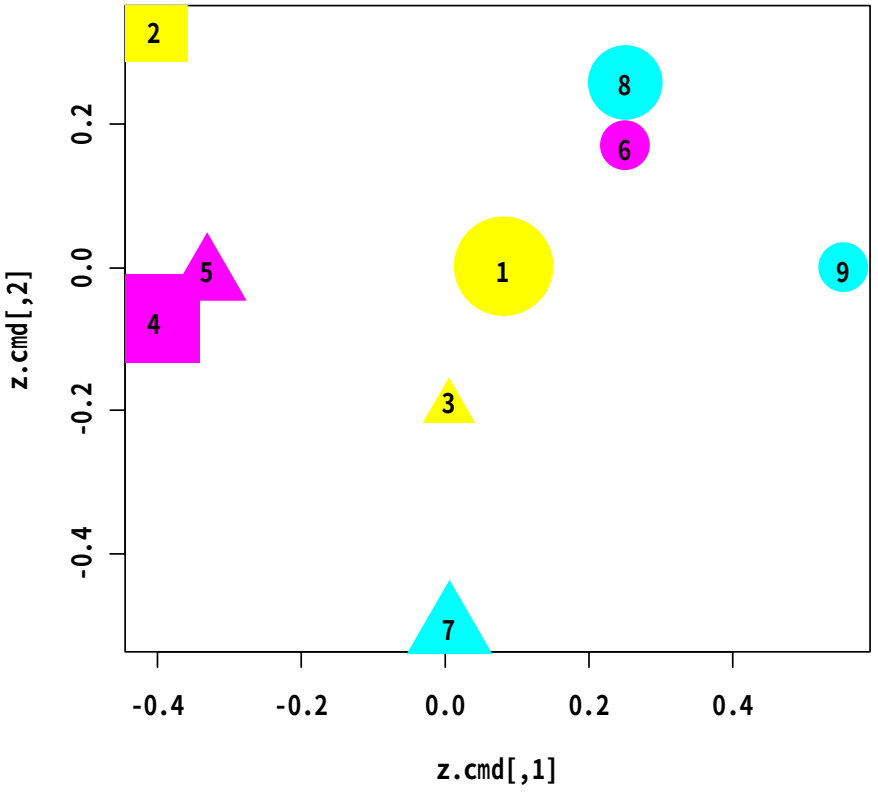
距離マトリックス（⑦重み付きCosine係数）

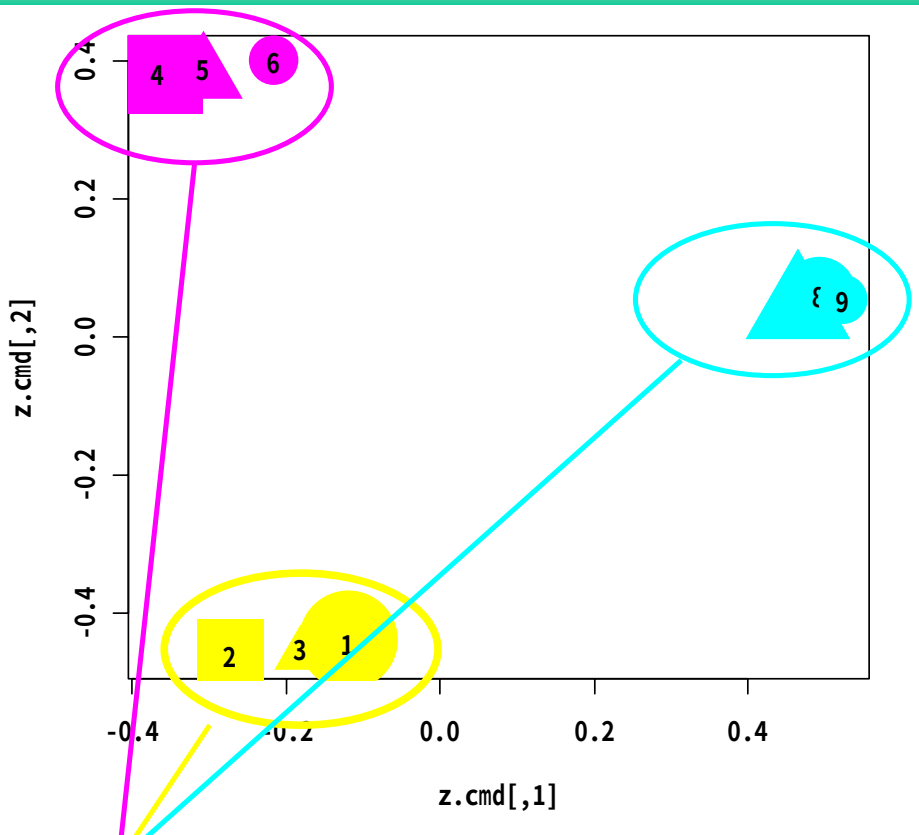
文書番号	doc1	doc2	doc3	doc4	doc5	doc6	doc7	doc8	doc9
doc1	0	0	0	0	0	0	0	0	0
doc2	0.33	0	0	0	0	0	0	0	0
doc3	0.33	0.33	0	0	0	0	0	0	0
doc4	0.83	0.83	1	0	0	0	0	0	0
doc5	1	0.83	0.83	0.33	0	0	0	0	0
doc6	0.83	1	0.83	0.33	0.33	0	0	0	0
doc7	0.83	1	0.83	0.83	0.83	1	0	0	0
doc8	0.83	0.83	1	1	0.83	0.83	0.33	0	0
doc9	0.83	1	0.83	1	1	0.67	0.33	0.17	0

距離＝１－類似度

対称行列のため
必要な下三角行列
のみ算出

文書のクラスタリング結果（多次元尺度法）





ターム重み
形　状：1
サイズ：1
カラー：1

特定の観点に重み付けすることで
その観点でクラスタリングできる

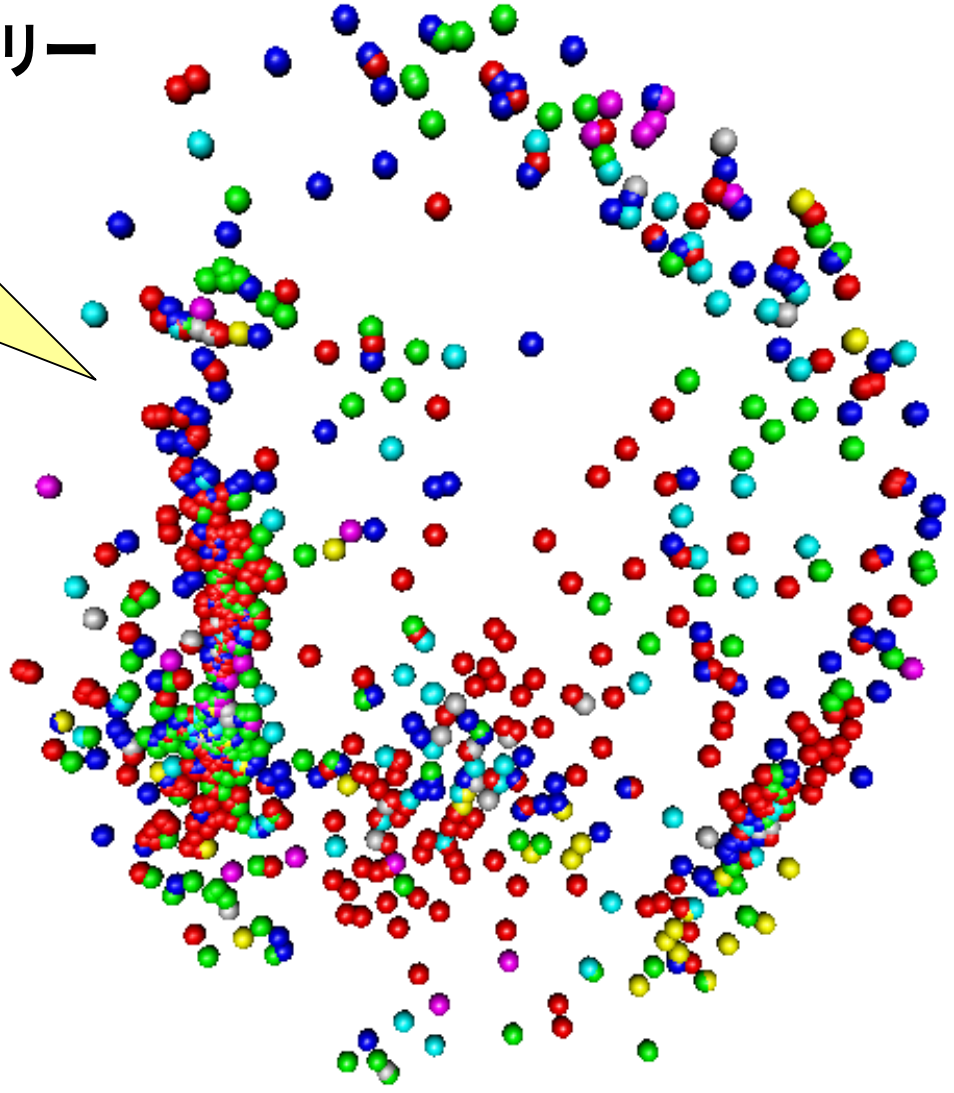
ターム重み
形　状：1
サイズ：1
カラー：2

但し、形状、サイズでは
クラスタリングしていない

多次元尺度法によるクラスタリング（カートリッジ）

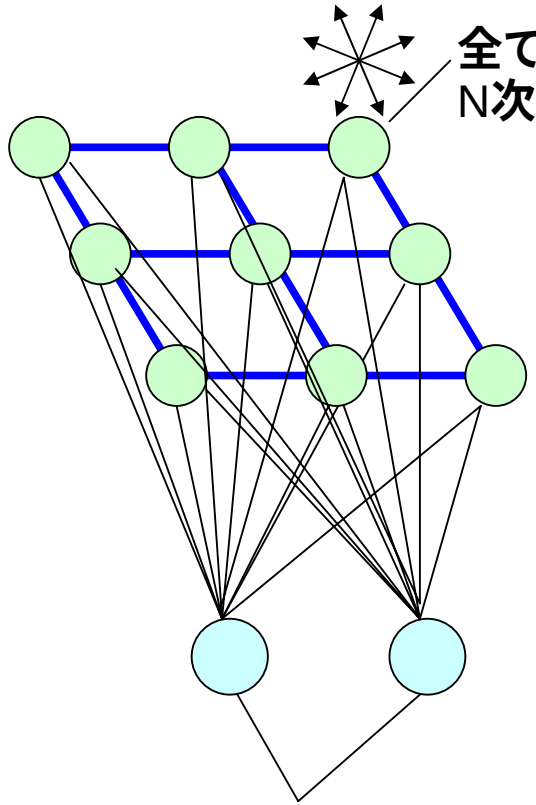
請求項1のカテゴリー

カテゴリー毎にクラスタリングできないか？



- カートリッジ
- プリンタ
- 構成部品
- ヘッド
- インク
- 分類不能

自己組織化マップの基本構造



全てのユニット(ノード)に
N次元の重みベクトルが存在

ノードの数($X_m \times Y_n$)は
任意に設定できる

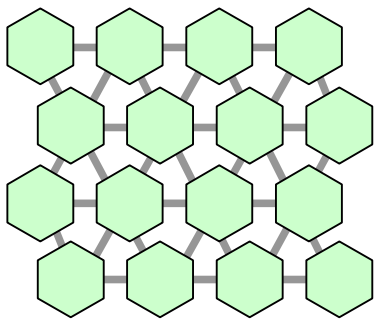
出力層
第2層

入力層
第1層

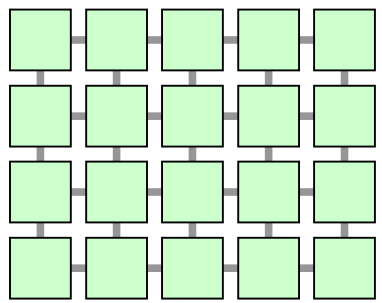
入力データ
文書ベクトル
(n次元ベクトルデータ)

第1層に入力されたデータは第2層の全てのユニットに伝えられる。そこで入力データと接続重みとがお互いにどれだけ似ているかを第2層のユニット間で競争する。競争の結果一番似ていたユニットは勝者と呼ばれる。勝者のユニットは重みが調節され、入力情報にさらに近づけられる。この過程を競合学習と呼ぶ。

第2層ユニットは一つひとつ明確な位置座標を持ち、空間的な拡がりを表現している。



(a) **蜂の巣状**hexagonal

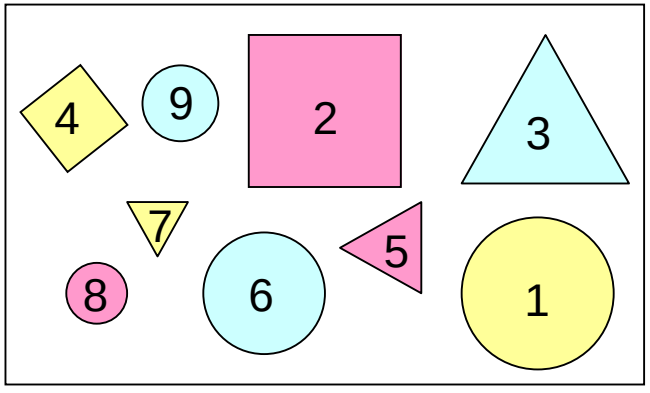


(b) **格子状**rectangular

豊田秀樹編著. “データマイニング入門：Rで学ぶ最新データ解析”. 東京図書,,2008.p114-115

文書のクラスタリング実験（自己組織化マップ）

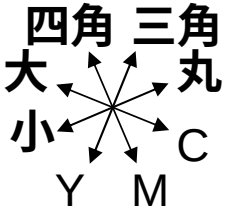
分類対象



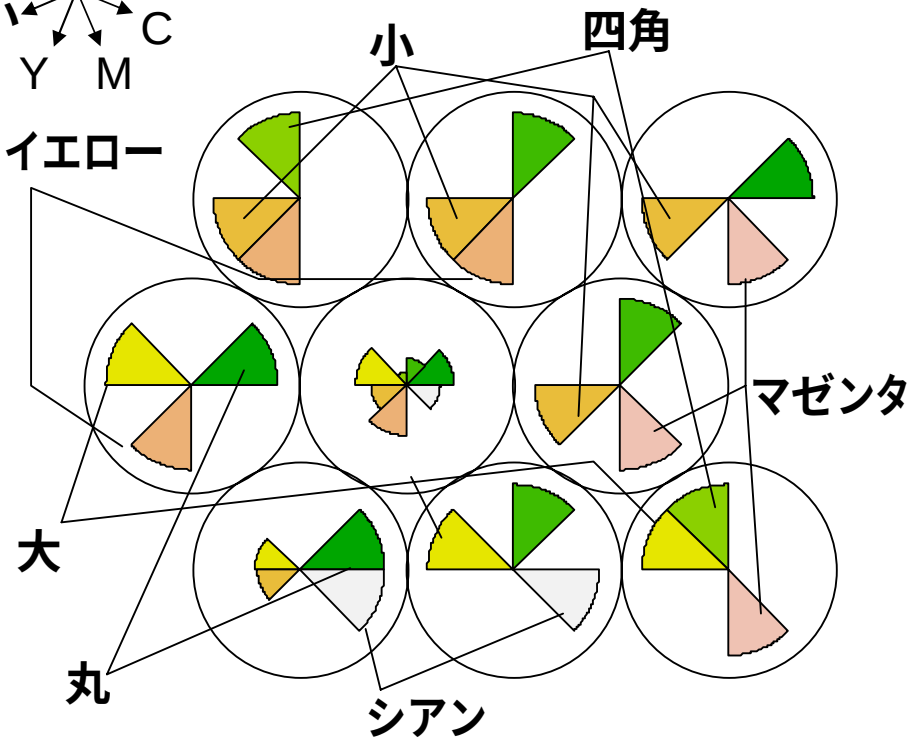
	形状			サイズ		カラー			
	図形番号	丸	三角	四角	大	小	イエロー	マゼンタ	シアン
	図1	1	0	0	1	0	1	0	0
	図2	0	0	1	1	0	0	1	0
	図3	0	1	0	1	0	0	0	1
	図4	0	0	1	0	1	1	0	0
9文書	図5	0	1	0	0	1	0	1	0
	図6	1	0	0	1	0	0	0	1
	図7	0	1	0	0	1	1	0	0
	図8	1	0	0	0	1	0	1	0
	図9	1	0	0	0	1	0	0	1

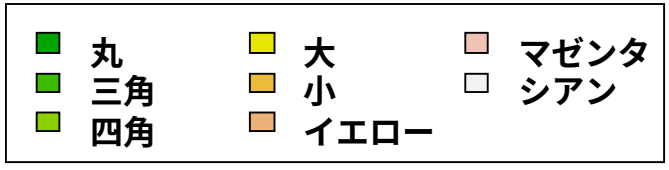
8次元ベクトル
（文書ベクトル）

文書のクラスタリング実験（自己組織化マップ）



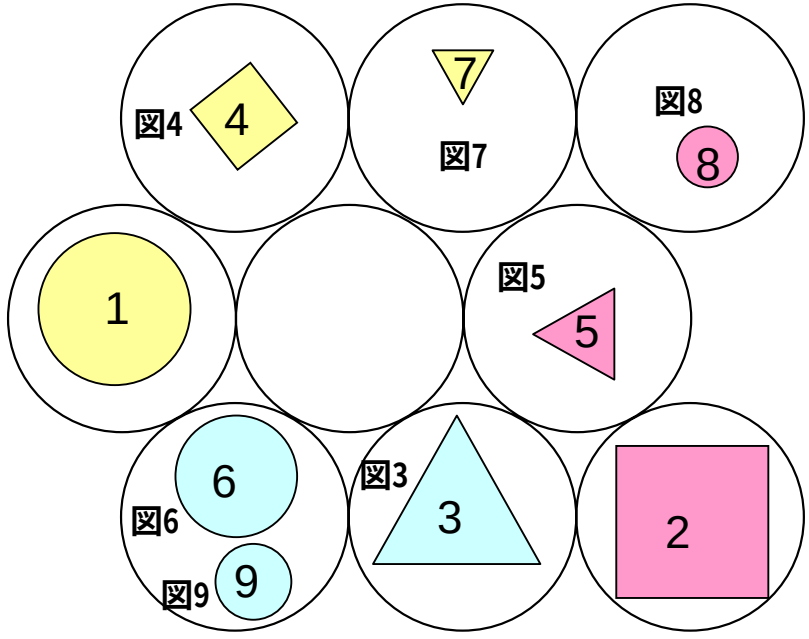
学習後のコード情報





8次元ベクトル

自己組織化マップによる分類

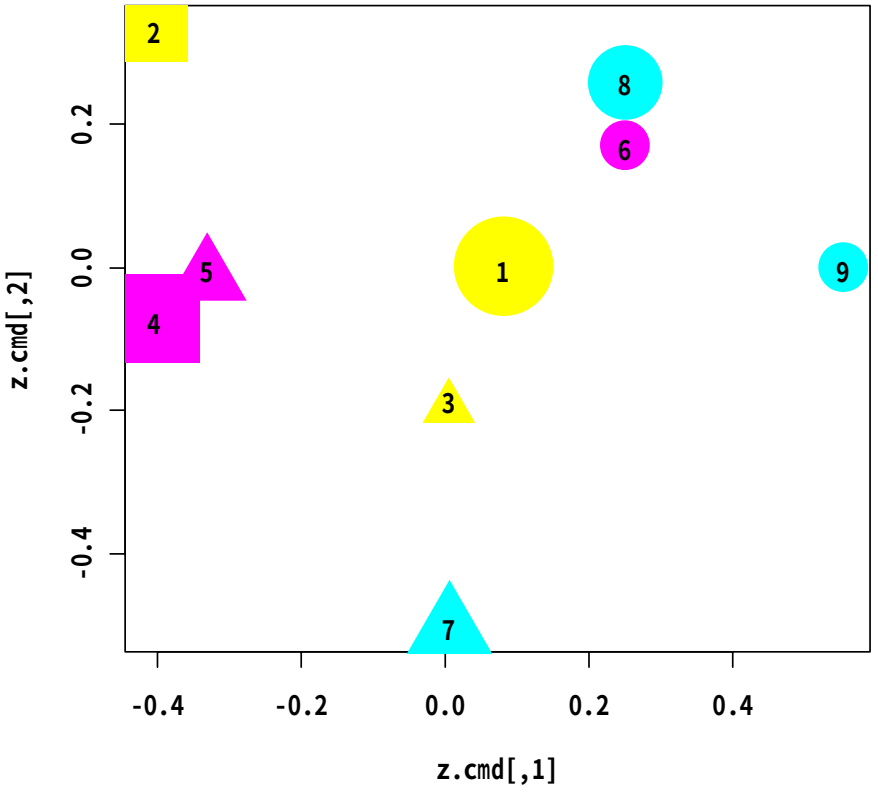


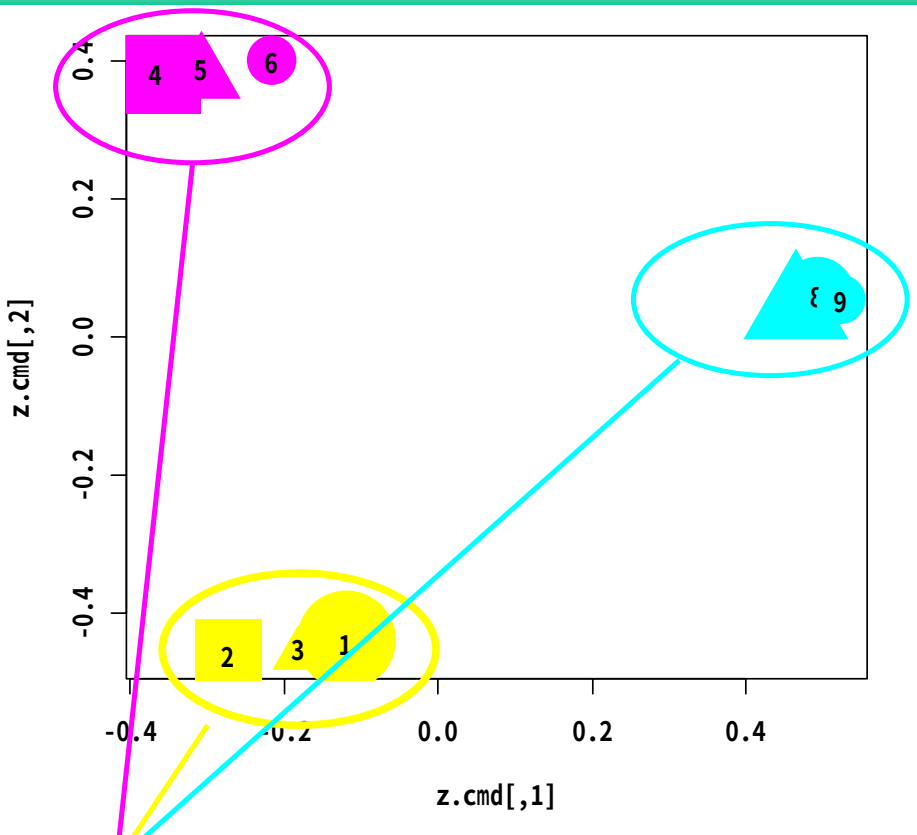
形状、サイズ、カラーで分類

各属性の似ているもの同士が
近くなるようにポジションが決まる

library(kohonen)
som**関数**

文書のクラスタリング結果（多次元尺度法）





ターム重み
形　状：１
サイズ：１
カラー：１

特定の観点に**重み付け**することで
その観点で**クラスタリング**できる

但し、形状、サイズでは
クラスタリングしていない

ターム重み
形　状：１
サイズ：１
カラー：２

自己組織化マップの入力データ（専門用語）

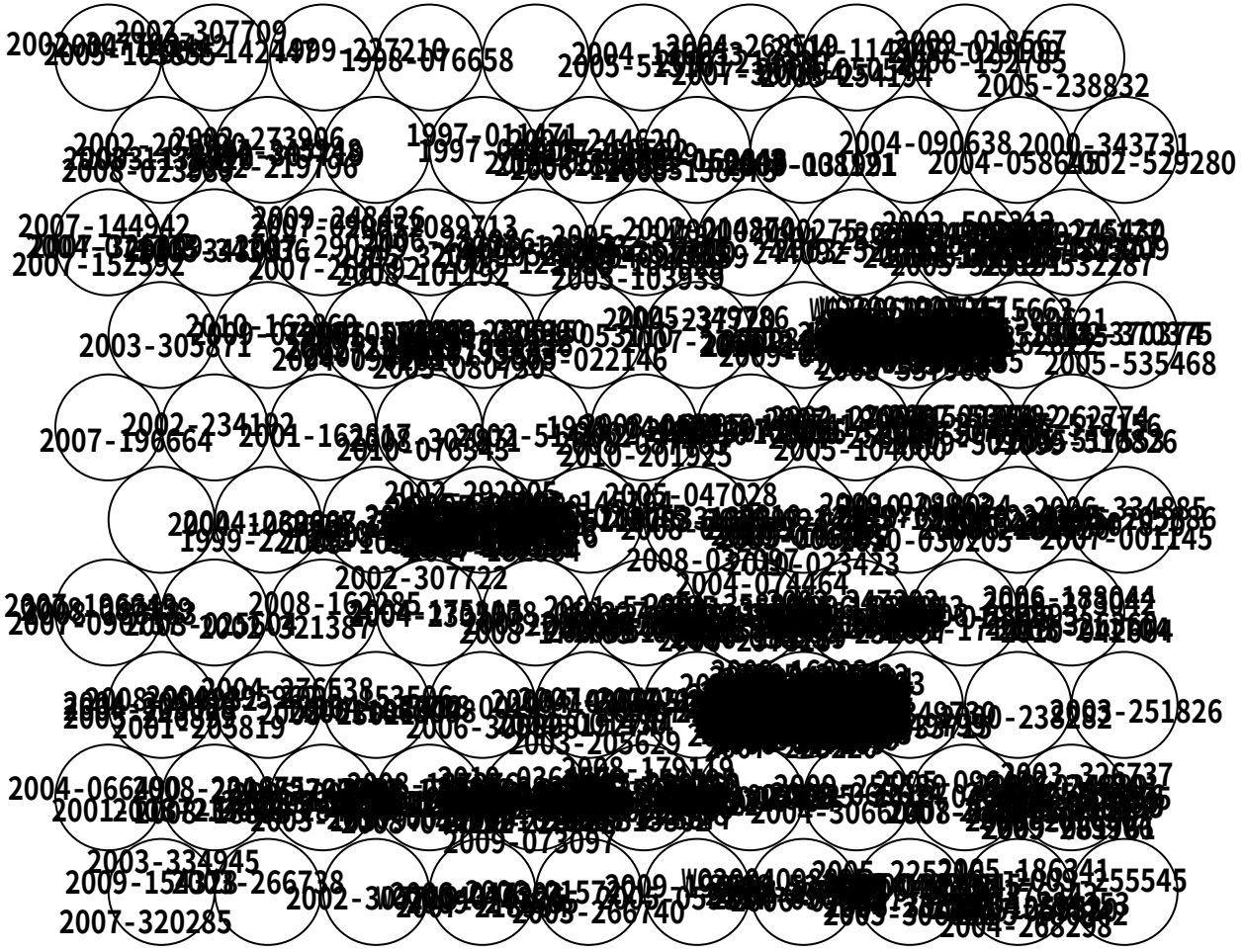
→ 専門用語：ノイズ除去の上位190語

公報番号	インク	インクカートリッジ	液体	インクタンク	液体収納容器	記録ヘッド	液体収容容	記録装置	液体収容部	カートリッジ
1994-040044	28841.77	0	0	6506.54	0	12414.78	0	0	0	0
1994-262774	129788	0	0	0	0	2482.96	0	0	0	0
1996-020115	28841.77	6654.28	0	97598.11	0	0	0	0	0	2302.88
1996-058107	115367.1	0	0	69403.1	0	39727.31	0	0	0	0
1996-174860	24034.81	0	0	0	0	12414.78	0	0	0	0
1996-183179	4806.96	0	0	0	0	0	0	0	0	0
1996-218279	24034.81	0	0	0	0	0	0	0	0	0
1996-290578	19227.85	0	0	34701.55	0	12414.78	0	0	0	0
1997-011471	0	0	132048.3	0	0	0	0	0	0	0
1997-048127	33648.74	0	95275.34	0	0	0	0	0	0	0
1998-024588	0	0	18386.47	0	0	0	0	0	0	0
1998-044475	312452.6	23289.99	0	0	0	0	0	0	0	0
1998-076658	0	0	342656.9	0	0	0	0	0	0	0
1998-193635	14420.89	0	0	0	0	0	0	0	0	0
1998-230615	9613.92	0	0	26026.16	0	0	0	0	0	0
1999-208097	4806.96	0	0	0	0	0	0	0	0	0
1999-227210	76911.4	0	45130.43	8675.39	0	0	0	0	0	0
1999-227222	105753.2	0	0	0	0	0	0	2832.14	0	0
1999-314377	0	0	0	0	0	0	0	0	0	0
1999-348308	4806.96	9981.43	0	0	0	0	0	0	0	0

↓ 766公報

一部抜粋

自己組織化マップによる特許のポジショニングマップ



library(kohonen)
特許SOM <- som(scale(特許データ1[,2:191]), grid =somgrid(10, 10,"hexagonal"), rlen=500)

自己組織化マップの入力データ（カテゴリー分類）

文書ベクトル
11カテゴリー⇒11次元

カテゴリー分類結果（全請求項：2値）

公報番号	カートリッジ	プリンタ	構成部品	ヘッド	記録方法	インク	インクジェット応用装置	インク充填方法	包装部材	製造方法	分類不能
2004-244620	1	1	0	0	1	1	0	0	0	0	0
2004-130792	0	1	0	0	1	1	0	0	0	0	0
2002-187918	1	1	0	0	0	1	0	0	0	0	0
2008-179804	1	0	0	0	0	1	0	0	0	0	0
2009-190403	0	0	0	0	0	1	0	0	0	0	0
2008-101192	1	1	0	0	1	1	0	0	0	0	0
2008-069327	1	1	0	0	1	1	0	0	0	0	0
2008-221846	0	0	0	0	0	1	0	0	0	0	0
2008-221843	0	0	0	0	0	1	0	0	0	0	0
2008-114600	0	0	0	0	0	1	0	0	0	0	0
2006-182800	1	1	0	0	1	1	0	0	0	0	0
2008-094095	0	0	0	0	0	1	0	0	0	0	0
2008-018720	0	0	0	0	0	1	0	0	0	0	0
2008-174703	0	1	0	0	0	1	0	0	0	0	0
2006-117883	1	1	0	0	1	1	0	0	0	0	0
2006-282986	1	1	0	0	1	1	0	0	0	0	0
2005-320509	1	1	0	0	1	1	0	0	0	0	0
2006-045537	1	1	0	0	1	1	0	0	0	0	0
2005-089713	1	1	0	0	1	1	0	0	0	0	0
2005-200552	1	1	0	0	1	1	0	0	0	0	0

766公報

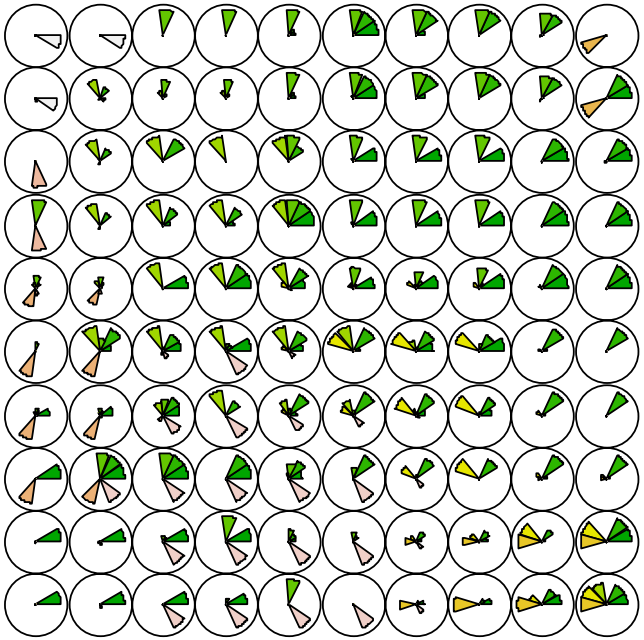
一部抜粋

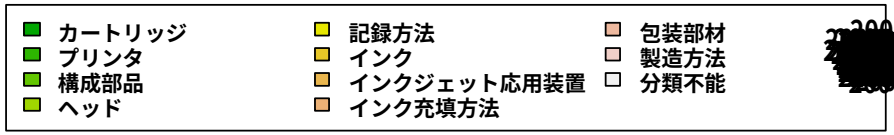
課題：重み付き分類結果（カテゴリー出現頻度、寄与率）

自己組織化マップによる特許のポジショニングマップ

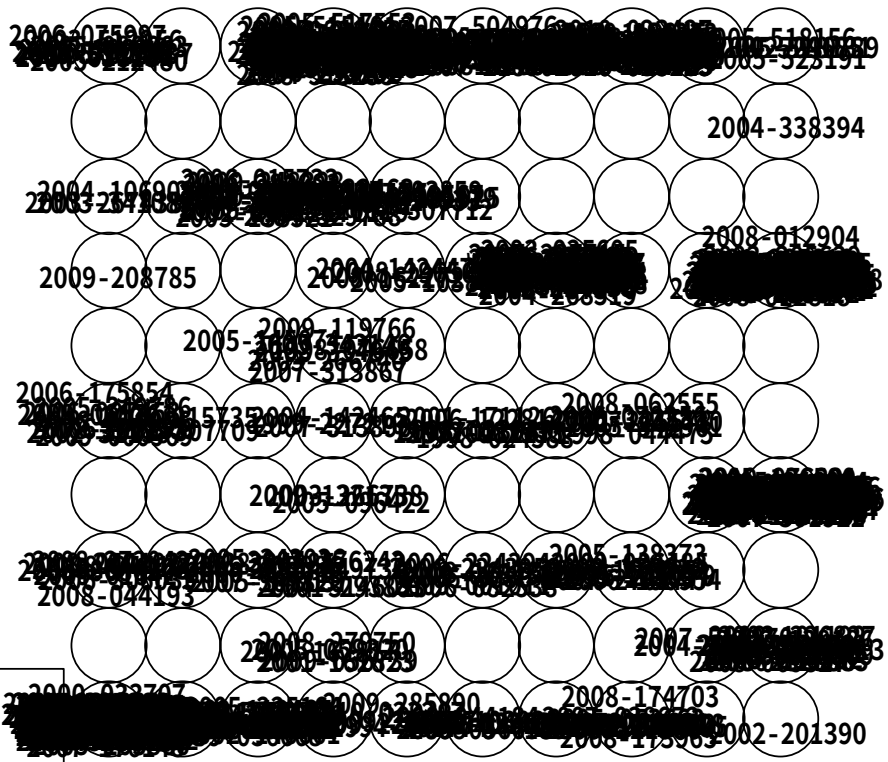
入力データ：カテゴリー分類結果

コード情報





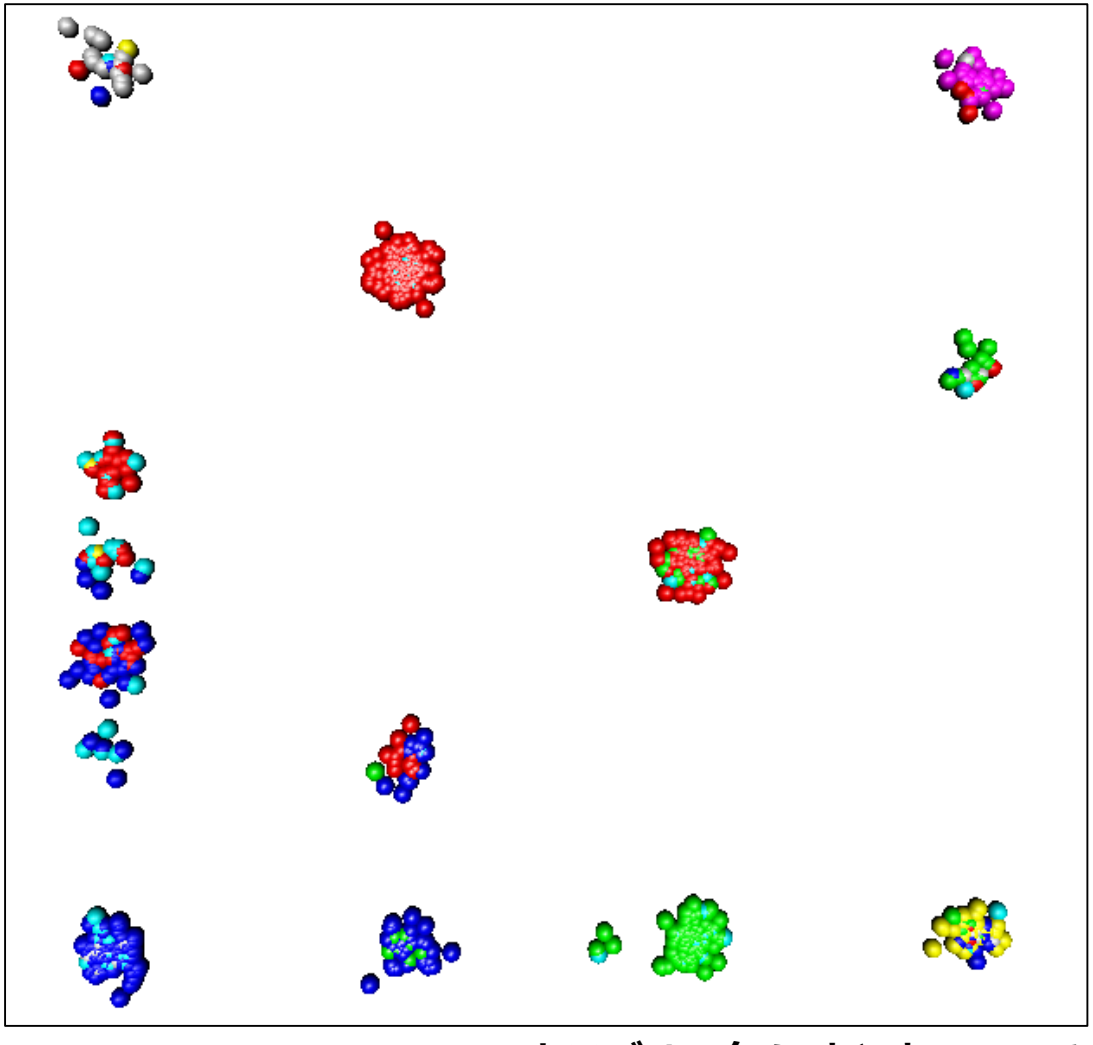
特許のポジショニングマップ



カテゴリー毎にまとまっている

```
library(kohonen)
som(scale(特許データ1[,2:12]), grid =somgrid(10, 10,"rect"), rlen=500)
```

自己組織化マップによる特許のポジショニングマップ



入力データ:
カテゴリー分類結果
全カテゴリー: **2値**

請求項1

- カートリッジ
- プリンタ
- 構成部品
- ヘッド
- インク
- 分類不能

カテゴリー毎にまとまっている

```
library(som)
som(標準化特許データ, xdim=10, ydim=10, topol="rect")
```

自己組織化マップの入力データ（カテゴリー分類）

文書ベクトル
11カテゴリー⇒11次元

→ カテゴリー分類結果（全請求項：頻度）

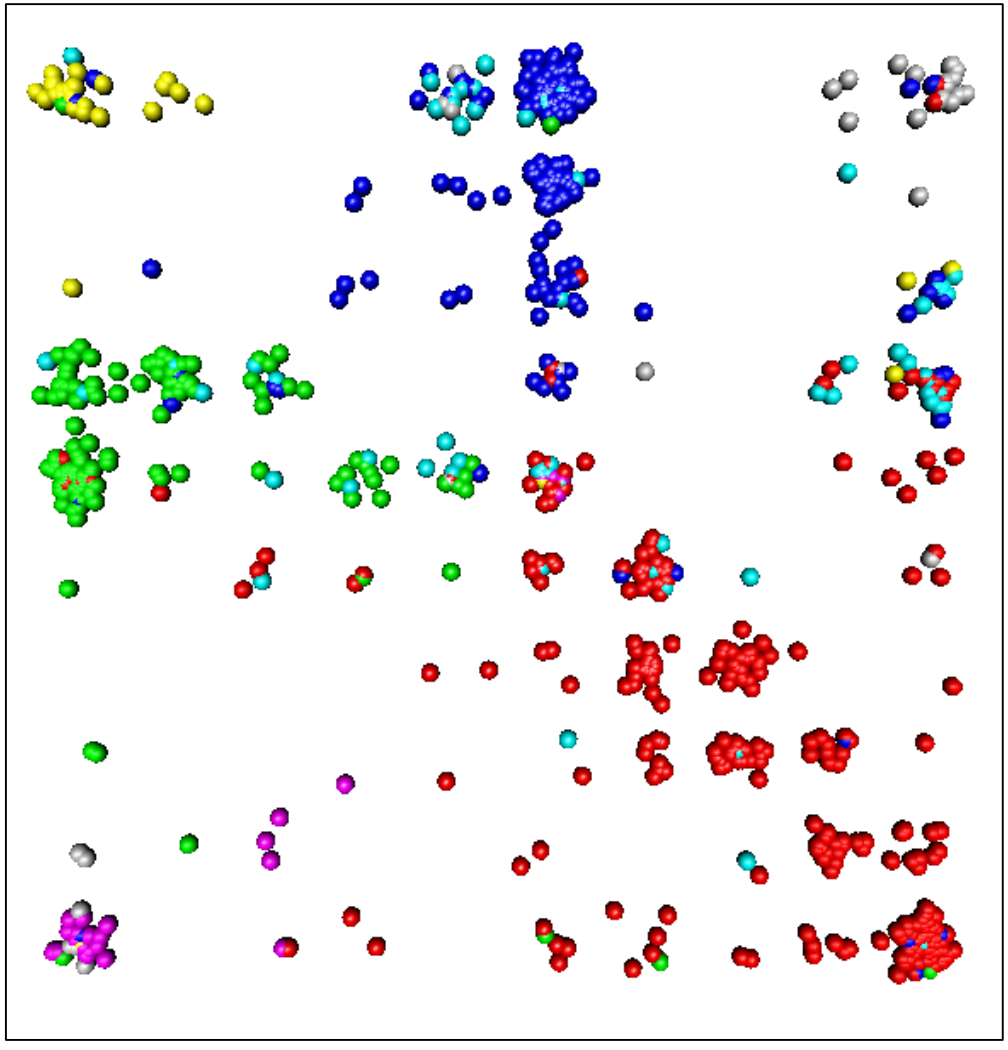
公報番号	カートリッジ	プリンタ	構成部品	ヘッド	記録方法	インク	インクジェット応用装置	インク充填方法	包装部材	製造方法	分類不能
2004-244620	8	5	0	0	3	6	0	0	0	0	0
2004-130792	0	7	0	0	7	3	0	0	0	0	0
2002-187918	1	2	0	0	0	9	0	0	0	0	0
2008-179804	2	0	0	0	0	4	0	0	0	0	0
2009-190403	0	0	0	0	0	4	0	0	0	0	0
2008-101192	1	2	0	0	9	11	0	0	0	0	0
2008-069327	1	2	0	0	3	4	0	0	0	0	0
2008-221846	0	0	0	0	0	1	0	0	0	0	0
2008-221843	0	0	0	0	0	3	0	0	0	0	0
2008-114600	0	0	0	0	0	2	0	0	0	0	0
2006-182800	6	7	0	0	7	6	0	0	0	0	0
2008-094095	0	0	0	0	0	2	0	0	0	0	0
2008-018720	0	0	0	0	0	2	0	0	0	0	0
2008-174703	0	2	0	0	0	8	0	0	0	0	0
2006-117883	2	2	0	0	2	2	0	0	0	0	0
2006-282986	1	5	0	0	2	15	0	0	0	0	0
2005-320509	2	7	0	0	3	15	0	0	0	0	0
2006-045537	1	2	0	0	2	7	0	0	0	0	0
2005-089713	1	5	0	0	1	17	0	0	0	0	0
2005-200552	5	3	0	0	2	2	0	0	0	0	0

766公報

一部抜粋

重み付き分類結果（カテゴリー出現頻度：寄与率）

自己組織化マップによる特許のポジショニングマップ



入力データ：
カテゴリー分類結果
全カテゴリー：頻度（寄与率）

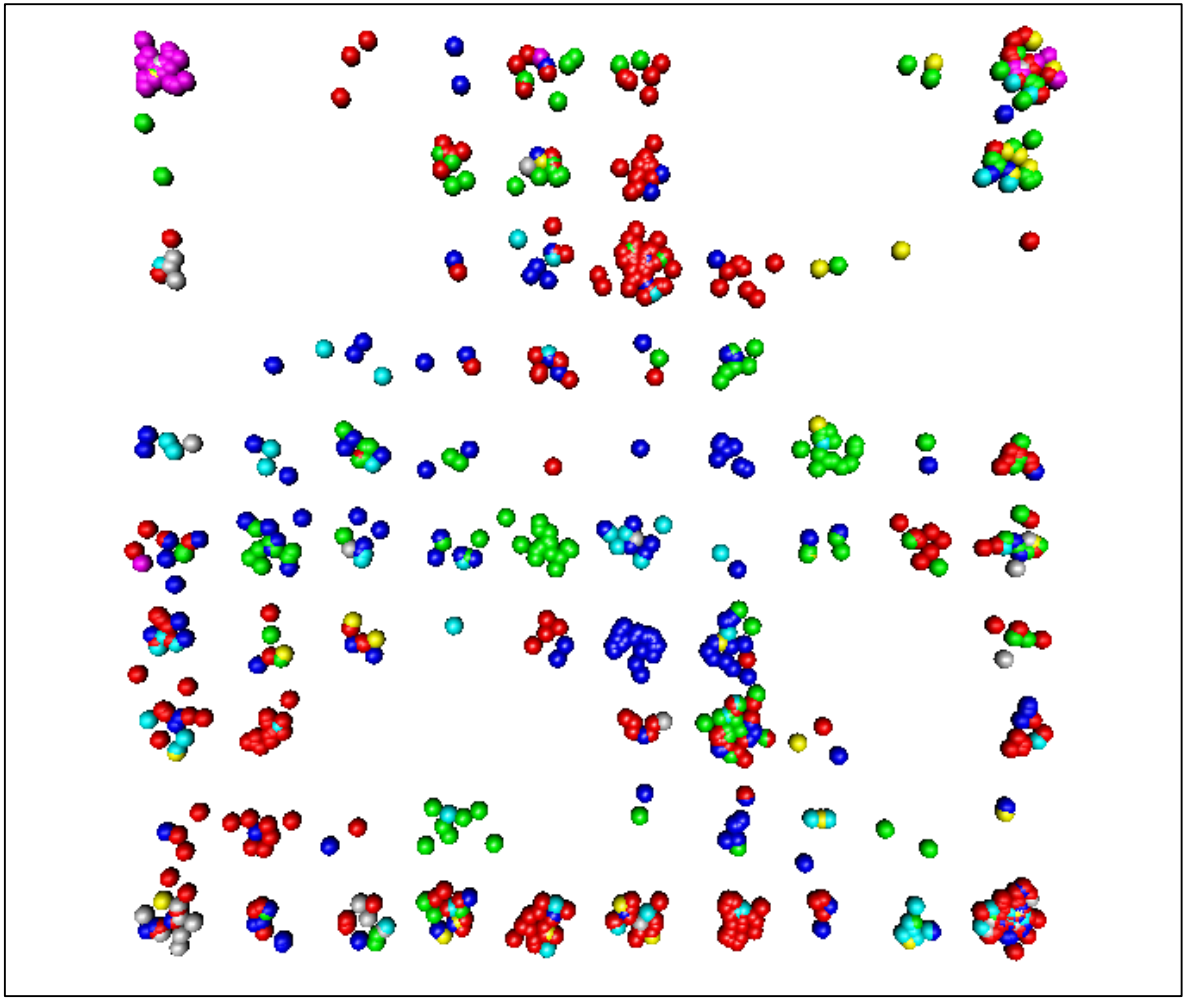
請求項1

- カートリッジ
- プリンタ
- 構成部品
- ヘッド
- インク
- 分類不能

カテゴリー毎にバランス良くまとまっている

```
library(som)
som(標準化特許データ, xdim=10, ydim=10, topol="rect")
```

自己組織化マップによる特許のポジショニングマップ



入力データ:
560カテゴリー分類無し
全カテゴリー: 2値

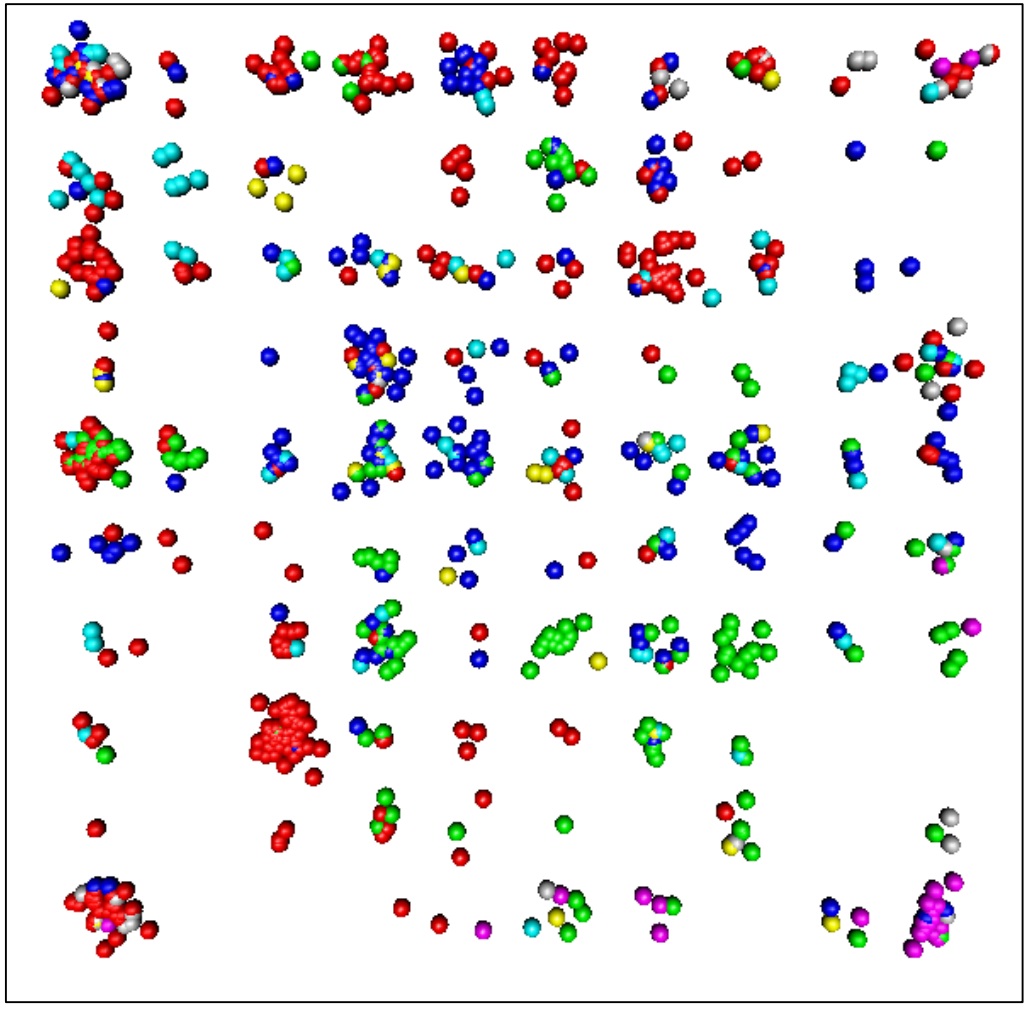
請求項1

- カートリッジ
- プリンタ
- 構成部品
- ヘッド
- インク
- 分類不能

一つのユニットには同じカテゴリーが集まる傾向

```
library(som)
som(標準化特許データ, xdim=10, ydim=10, topol="rect")
```

自己組織化マップによる特許のポジショニングマップ



入力データ:
560カテゴリー分類無し
全カテゴリー:頻度

請求項1
- カートリッジ
- プリンタ
- 構成部品
- ヘッド
- インク
- 分類不能

一つのユニットには同じカテゴリーが集まる傾向

```
library(som)
som(標準化特許データ, xdim=10, ydim=10, topol="rect")
```

まとめ

テキストマイニングによる解析から分類体系構築、特許の自動分類まで
一連の流れを検討した。

・多次元尺度法による非類似度（距離）を用いた文書間のクラスタリングは
全体の俯瞰に適している。

・LUT（参照テーブル）使用のテキスト自動分類はマッチングが成立して
分類可能な場合、物の発明のカテゴリー分類では良い結果が得られる。

・自己組織化マップによる自動分類では複数の観点をそれぞれ
考慮した分類が可能である。

・カテゴリー分類と自己組織化マップによる自動分類を組み合わせると良い

・専門用語－文書の統計情報、専門用語間の関係を解析、可視化することで
調査・解析用ユーザー辞書（日本語、英語、中国語）作成支援を行うことができる。

・上記ユーザー辞書は網羅的あるいは高精度の検索にも有用である

ご清聴ありがとうございました

アジア特許情報研究会

テキストマイニング検討チーム一同

予備資料

ハッシュテーブルを用いたターム（専門用語）統計出力

ハッシュテーブル

・キー（ユニーク）と値のセット
・キーを与えると値を返す

キー1：値1
キー2：値2
キー3：値3
・
・
・

キーとしてタームを使用して
値に配列の番号（添え字）を格納する

ハッシュテーブル

ターム1：配列（1）
ターム2：配列（2）
ターム3：配列（3）
・
・
・

統計情報を格納する
数値配列

No.	Total重要度	平均	標準偏差
1			
2			
3			
・			
・			

文書番号を格納する
文字列配列

No.	文書番号	・
1		
2		
3		
・		
・		

インバーテッド（転置）ファイル

ターム（専門用語）統計出力

Term	Total重要度	平均重要度	標準偏差	Min.重要度	Max.重要度	文書数	文書		
インク	60855295.43	96442.62	126737.9	11682.72	1693994.6	631	1990-501818	1992-211470	1993
液体	25770803.51	88864.84	131284.8	8273.13	1472617.36	290	1994-079880	1996-310004	1997
インクカートリッジ	17360646.99	53090.66	56751.38	5777.25	381298.74	327	1990-501818	1992-211470	1994
インクタンク	6406614.54	34444.16	40270.55	3889.87	287850.31	186	1994-040044	1995-195703	1995
液体吐出ヘッド	4288402.65	107210.07	146133	4292.7	725465.5	40	1994-079880	1997-011471	1997
液体収納容器	4057128.66	69950.49	50402.78	3739.29	205660.9	58	1997-267483	1997-286117	1998
液体収容容器	3173056.46	68979.49	38969.89	9065.88	231179.83	46	2002-234178	2004-175417	2004
液体容器	2870065.03	47050.25	53737.49	4335.45	268797.62	61	1997-011471	1997-048127	1997
インクジェット記録装置	2411262.46	16629.4	13840.82	2679.18	61621.16	145	1992-211470	1993-301340	1994
インク容器	2405955.98	48119.12	48729.95	5151.94	231837.3	50	1995-195705	1995-232436	1997
液体吐出装置	2354878.11	50103.79	44024.09	4418.16	251834.99	47	1994-079880	1997-011471	1997
記録ヘッド	2179325.27	12819.56	14434.47	3415.87	88812.64	170	1992-211470	1993-301340	1994
記録装置	2097227.17	20764.63	22006.14	3566.71	167635.51	101	1995-195705	1995-232461	1995
カートリッジ	1737010.64	22558.58	37240.02	2856.93	225697.08	77	1996-020115	1996-025754	1997
装置	1681928.44	16819.28	39805.02	2529.22	326268.77	100	1990-501818	1994-199031	1994
液体収容体	1527857.94	43653.08	28947.37	3410.4	109132.71	35	2004-090623	2004-237731	2004
記録	1493850.09	14092.93	14597.7	5029.8	85506.56	106	1992-211470	1993-301340	1994
インク供給口	1420258.23	15780.65	17568.11	4383.51	109587.83	90	1995-195703	1996-132635	1996
液体噴射装置	1197365.29	16864.3	17486.54	2878.28	69078.76	71	2002-307713	2004-090623	2004
液体収容部	1141228.38	18114.74	19740.08	4755.12	142653.55	63	1997-267483	1997-286117	1998
液体カートリッジ	1045256.45	45445.93	32490.14	4861.66	111818.13	23	2004-090623	2004-237718	2004
インク供給装置	1022227.19	51111.36	38554.72	5324.1	165047.09	20	1997-240020	1999-508506	2001
印刷装置	1017980.27	26788.95	25575.63	2242.25	109870.12	38	1998-044475	1998-319797	1999
液流路	996339.69	49816.98	50472	3629.65	206890.21	20	1997-011471	1997-048127	1997
容器	963304.06	12350.05	16227.33	2271.94	77246.08	78	1995-314709	1996-174860	1998
インク収容部	933592.47	17614.95	12199.85	5334.81	58682.96	53	1992-211470	1994-025575	1994
吐出口	928649.51	13656.61	18241.81	2412.08	127840.06	68	1994-079880	1994-199031	1994

一部抜粋

左側のTermを含む文書リスト

インバーテッド（転置）ファイル

カテゴリー（請求項の最後のターム）統計情報

No.	カテゴリー	頻度	平均	標準偏差	Min.	Max.	文書数	文書		
1	インクカートリッジ	1136	8.29	6.57	1	44	137	1998-044475	2000-033707	2
2	液体収納容器	496	10.78	9.13	1	47	46	2000-033715	2001-063098	2
3	インクタンク	403	7.07	4.37	1	21	57	1996-020115	1996-058107	1
4	インクジェット記録装置	392	4.67	4.24	1	16	84	1996-058107	1998-230615	1
5	液体収容容器	351	7.8	5.44	1	27	45	2002-234178	2004-175417	2
6	液体容器	301	10.38	10.58	1	58	29	2001-146019	2001-146023	2
7	製造方法	290	4.6	3.95	1	18	63	1994-262774	1998-193635	1
8	液体吐出ヘッド	287	20.5	23.42	3	82	14	1997-011471	1997-048127	1
9	装置	273	19.5	24.78	1	89	14	2002-207807	2002-529280	2
10	方法	271	5.42	5.04	1	19	50	2001-253086	2002-014870	2
11	記録装置	208	3.85	3.29	1	14	54	2000-334977	2001-113723	2
12	インクジェットプリンタ	205	5.26	5.1	1	22	39	2002-273911	2002-507507	2
13	液体吐出装置	199	6.63	4.73	1	19	30	1997-011471	1997-048127	1
14	液体噴射装置	177	4.43	5.07	1	21	40	2002-307713	2004-142128	2
15	画像形成装置	170	6.07	4.65	1	15	28	2002-347303	2003-053947	2
16	印刷装置	162	7.04	6.24	1	23	23	1998-044475	2000-343731	2
17	液体収容体	150	7.14	3.66	1	13	21	2004-090623	2004-284353	2
18	インク供給装置	121	8.07	6.79	1	27	15	2001-514985	2004-345084	2
19	カートリッジ	105	9.55	9.2	1	34	11	2001-187457	2002-529280	2
20	インクジェット記録ヘッド	99	7.62	5.69	1	23	13	2002-001933	2002-103637	2
計	560	11076								

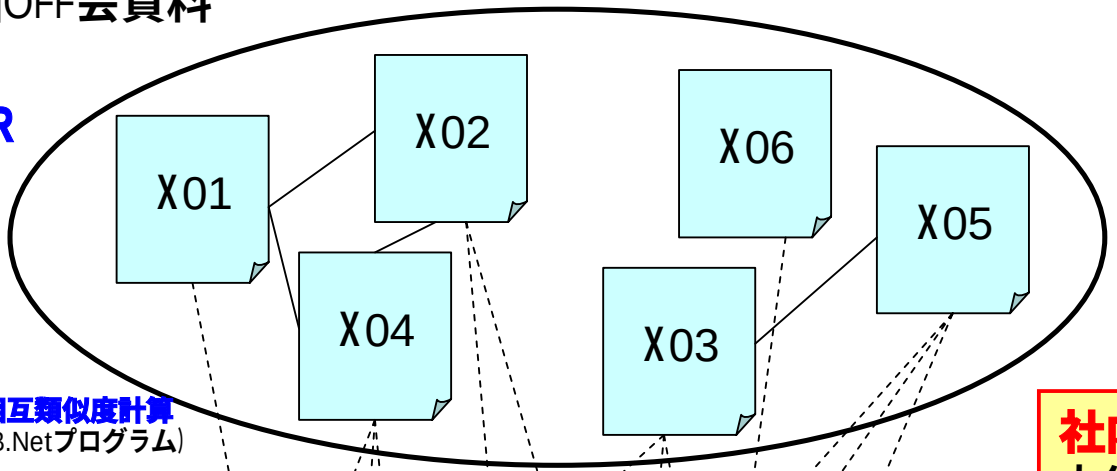
一部抜粋

特許公報、専門用語、単語のマルチレベルネットワーク

2011.06第2回OFF会資料

統計解析言語R

特許公報

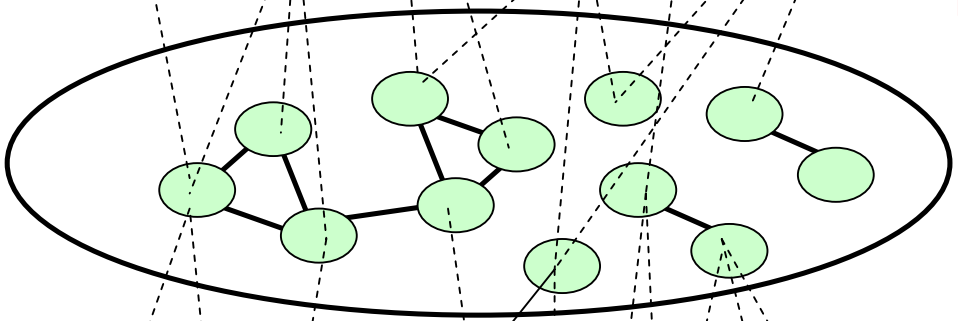


特許分類
IPC
FI
Fターム

社内(独自)分類
大分類、中分類、小分類

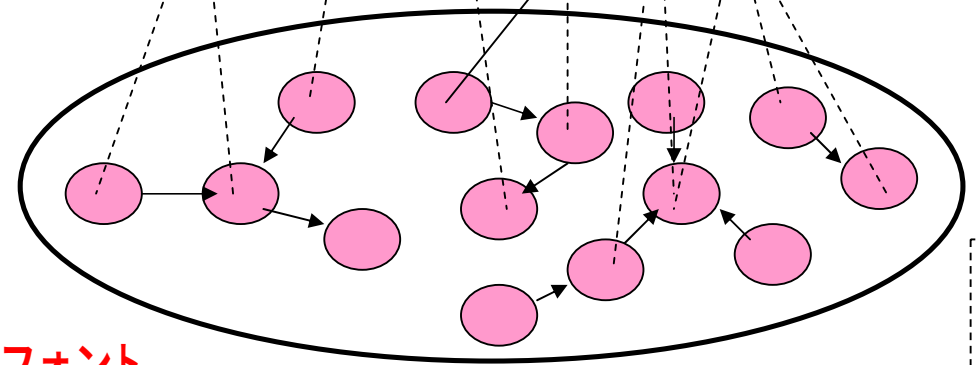
文書相互類似度計算
(自作VB.Netプログラム)

概念:concept



インクジェット記録用シート、記録媒体、インク受容層
インクジェット記録媒体、インク

ターム:専門用語
(特徴語)
「termmi」



インクジェット、記録、用、シート、媒体、インク、受容、層、

JP:漢字、ひらがな、カタカナ
CN:漢字、簡体字、繁体字
KR:ハングル

単語
(形態素:名詞)
和布蕪

文字、文字コード、フォント

検索モデルを応用した自動分類

自動分類の手法	検討項目
①ブーリアンモデル ・単語（形態素）のマッチング BOWモデル（Bag of Words） 例：インク、容器、カートリッジ、プリンタ・・・ ・専門用語（ターム）によるマッチング 例：インクカートリッジ、インクタンク、液体収納容器 ・カテゴリー（請求項の最後のターム）のマッチング	・動詞、形容詞の利用 ・名詞バイグラム（インクー容器） ・ネットワーク分析による重要語抽出 タームバイグラム 例：液体収納容器の製造方法 ⇒液体収納容器ー製造方法 カテゴリーの寄与率
②ベクトル空間モデル ・単語（形態素）使用のクラスタリング パテントインテグレーションのクラスタリング ・非計量多次元尺度法を用いたクラスタリング	既存分類の典型的なクラスター（教師データ）との類似度から分類する ・分類追加クラスタリング ・形態素追加（ターム+ワード） ・重要度（重み）補正

情報検索とテキストの自動分類

情報検索

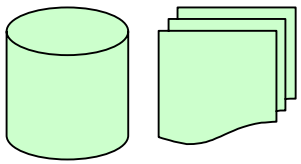
検索モデル
・ブーリアンモデル
・ベクトル空間モデル

検索質問（クエリ） → **検索システム** → **文書集合から検索質問に一致する文書を抽出する**

検索式

内部表現で
比較・照合

文書集合

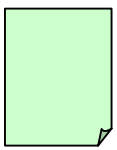


全文＝インクジェット用紙
FI=B41M5/00　B

自動分類

分類モデル
・規則に基づくモデル
・ベクトル空間モデル

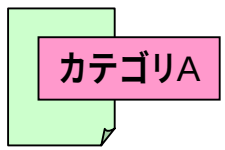
分類対象文書 → **分類システム** → **文書にもっとも類似したカテゴリを付与する**



分類対象文書と
各カテゴリ間の
類似度を計算する

例
記録用紙
インク
記録装置

あらかじめ
決めておいた
カテゴリ



注）テキストの自動分類は
情報検索と基本的には同じ
情報検索の技術を利用可能

自己組織化マップとは

自己組織化マップ(SOM: Self-Organizing Map)は、コホネン(T. Kohonen)により提案された教師なしのニューラルネットワークアルゴリズムで、高次元データを2次元平面上へ非線形写像するデータ解析方法である。

自己組織化マップは、**入力層**と**出力層**により構成された**2層のニューラルネットワーク**である。出力層は競合層とも呼ばれている。 [1]

入力層には***n*次元の入力データ**をそのまま与え、競合層では*m*次元上に配置されたノード（ユニット）がその入力データを学習する。**ノードには対応した*n*次元の重みベクトル**が存在し、**学習はこの重みベクトルの更新**によって行われる。
入力層および競合層のノード配置の次元は自由に設定できる。そのため、高次元データの視覚化によく用いられる。自己組織化写像は高次元のデータ間に存在する非線形な関係を簡単に幾何学的関係を持つ像に変換することができる。
通常、競合層は2次元、もしくは3次元に設定される。競合層を2次元に設定すると、平面上に配置されたノードが入力データ間の関係を表現することになる。

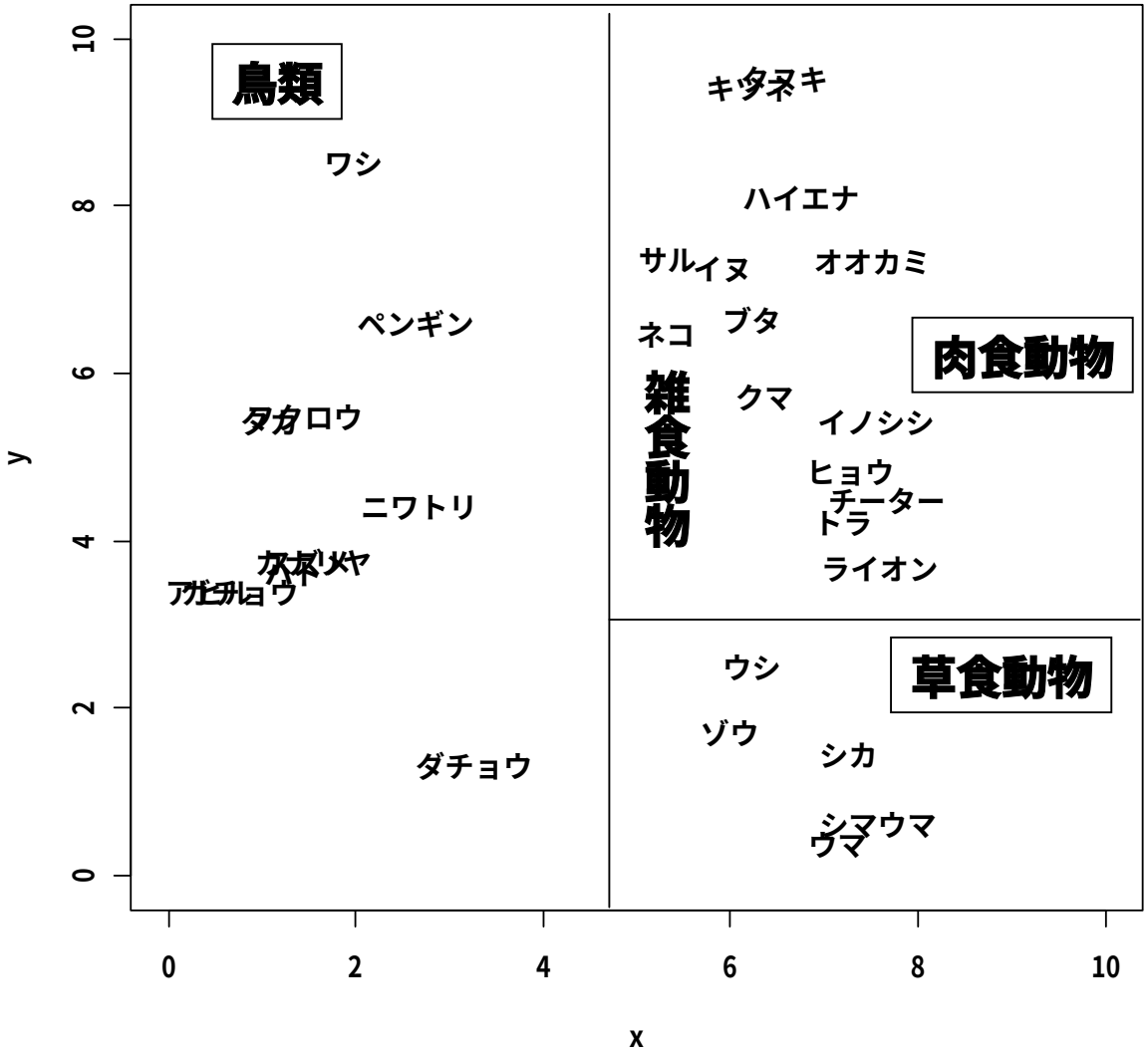
[1]金明哲　“Rと自己組織化マップ” http://www1.doshisha.ac.jp/~mjin/R/30/30.html

自己組織化マップとは

30種類の動物とその13属性

		1	2	3	4	5	6	7	8	9	10	11	12	13
	動物名	小	中	大	2本足	4本足	毛	蹄	たてがみ	羽	狩猟	走る	飛ぶ	泳ぐ
1	アヒル	1	0	0	1	0	0	0	0	1	0	0	1	1
2	イヌ	0	1	0	0	1	1	0	0	0	0	1	0	0
3	イノシシ	0	1	0	0	1	1	1	0	0	0	1	0	0
4	ウシ	0	0	1	0	1	1	1	0	0	0	0	0	0
5	ウマ	0	0	1	0	1	1	1	1	0	0	1	0	0
6	オオカミ	0	1	0	0	1	1	0	1	0	1	1	0	0
7	カナリヤ	1	0	0	1	0	0	0	0	1	0	0	1	0
8	ガチョウ	1	0	0	1	0	0	0	0	1	0	0	1	1
9	キツネ	0	1	0	0	1	1	0	0	0	1	0	0	0
10	クマ	0	0	1	0	1	1	0	0	0	1	0	0	0
11	サル	0	1	0	1	1	1	0	0	0	0	0	0	0
12	シカ	0	0	1	0	1	1	1	0	0	0	1	0	0
13	シマウマ	0	0	1	0	1	1	1	1	0	0	1	0	0
14	スズメ	1	0	0	1	0	0	0	0	1	0	0	1	0
15	ゾウ	0	0	1	0	1	0	1	0	0	0	0	0	0
16	タカ	1	0	0	1	0	0	0	0	1	1	0	1	0
17	タヌキ	0	1	0	0	1	1	0	0	0	1	0	0	0
18	ダチョウ	0	0	1	1	0	0	0	0	1	0	1	0	0
19	チーター	0	0	1	0	1	1	0	0	0	1	1	0	0
20	トラ	0	0	1	0	1	1	0	0	0	1	1	0	0
21	ニワトリ	1	0	0	1	0	0	0	0	1	0	0	0	0
22	ネコ	1	0	0	0	1	1	0	0	0	1	0	0	0
23	ハイエナ	0	1	0	0	1	1	0	0	0	1	1	0	0
24	ハト	1	0	0	1	0	0	0	0	1	0	0	1	0
25	ヒョウ	0	0	1	0	1	1	0	0	0	1	1	0	0
26	フクロウ	1	0	0	1	0	0	0	0	1	1	0	1	0
27	ブタ	0	1	0	0	1	1	1	0	0	0	0	0	0
28	ペンギン	0	1	0	1	0	0	0	0	1	0	0	0	1
29	ライオン	0	0	1	0	1	1	0	1	0	1	1	0	0
30	ワシ	0	1	0	1	0	0	0	0	1	1	0	1	0

自己組織化マップとは


鳥類
ワシ
ペンギン
タカ
ロウ
ニワトリ
アヒル
ダチョウ
キツネ
ハイエナ
サル
イヌ
オオカミ
ネコ
ブタ
肉食動物
雑食動物
クマ
イノシシ
ヒョウ
チーター
トラ
ライオン
草食動物
ウシ
ゾウ
シカ
ウマ
0
2
4
6
8
10
x
y

自己組織化マップの学習回数と類似度の変化