テキストマイニングと機械学習による効率的な特許調査

アジア特許情報研究会 1)
〇安藤俊幸　花王株式会社

目次

INFOPRO2016発表

①技術動向調査　対象：人工知能
②先行技術調査　対象：即席麺

YEARBOOK2017

先行技術調査への機械学習適用の基礎検討
言語処理における分散表現学習の基礎検討

1）アジア特許情報研究会

http://www.geocities.jp/patentsearch2006/asia-research.html

Japio YEARBOOK2016寄稿論文とINFOPRO2016発表内容の全体像

Japio YEARBOOK2016 寄稿論文 機械学習を用いた効率的な特許調査方法 http://www.japio.or.jp/00yearbook/	INFOPRO2016発表 機械学習を利用した効率的な特許調査方法 動向調査と先行技術への機械学習の応用
①技術動向調査 対象：人工知能 （G06N）/IPC/CPC AND PD=2006-01-01:2016-06-30 22457 ファミリー（出願数ベース57778件） 言語：英語、日本語 教師データなしの機械学習を利用したクラスタリング	①技術動向調査 対象：人工知能 (G06N)/IPC/CPC AND (US AND JP AND CN)/PN AND PD=2006-01-01:2016-06-30 1449ファミリー（出願数ベース12867件） 言語：日本語、英語、中国語（可能） 教師データなしの機械学習を利用したクラスタリング
②先行技術調査 対象：即席麺の直近10 年 イントロ 教師データありの機械学習	②先行技術調査 対象：即席麺の直近10 年 評価 教師データありの機械学習を応用

↑商用ツールを用いた解析

http://www.japio.or.jp/00yearbook/files/2016book/16_2_10.pdf

↑自分で試して結果の解析/検証に軸足

https://www.jstage.jst.go.jp/article/infopro/2016/0/2016_139/_article/-char/ja/

Japio YEARBOOK2017寄稿論文とINFOPRO2017発表内容の全体像

Japio YEARBOOK2017 寄稿論文
機械学習を用いた効率的な特許調査方法
ニューラルネットワークの特許調査への適用に関する基礎検討

（基礎編）　特許情報フェア11/8-10配布予定

先行技術調査への機械学習適用の基礎検討
- 先行技術調査の流れ
- データセット作成（特許検索競技大会2016の事例）
- 分かち書きと重み付けの再現率への影響
- 形態素解析（MeCab）による分かち書き
- 専門用語による分かち書き
- 評価関数とフィルターの影響

言語処理における分散表現学習の基礎検討
- Doc2vecによる文書のベクトル化処理の概要
- 文書の分散表現ベクトルの学習モデルと再現率
- 分散表現ベクトルの次元数（Size）の影響
- 非計量多次元尺度法による公報群の可視化
- doc2vecの類似度による公報群の可視化
- word2vecによる類似語抽出

- Visual Mining Studio（VMS）の自己組織化マップ
- BayoLinkによるベイジアンネットワーク紹介

↑テキストマイニング/機械学習の基礎検討

INFOPRO2017発表予定（ 11/30 ～12/1 ）
機械学習を利用した効率的な特許調査方法
ニューラルネットワークの特許調査への応用

（応用編）

1．単語のOne hotベクトル表現による検討
①分かち書きの影響
- 形態素/専門用語/Nグラム（文字単位）

②重み付けの影響
- TF（Term Frequency、単語の出現頻度）
- TF-IDF（Inverse Document Frequency、逆文書頻度）

③新規性を考慮した評価関数
- Fタームと類似度による評価関数
- Fタームによるフィルター

2．単語/文書の分散表現ベクトルによる検討
①Doc2Vecによる文書の分散表現学習
- PV-DM(Paragraph Vector with Distributed Memory) モデル
- PV-DBOW(Paragraph Vector with Distributed Bag of Words) モデル

②Word2Vecによる単語の分散表現学習

3．可視化検討
①次元圧縮
- PCA:Principal Component Analysis主成分分析
- t-SNE:t-Stochastic Neighbor Embedding
- MDS:Multi-Dimensional Scaling多次元尺度法
- nMDS:Non metric Multi-Dimensional Scaling非計量多次元尺度法

↑自分で試して結果の解析/検証→応用検討

http://www.japio.or.jp/00yearbook/　12/上　Web公開予定

使用データベース／解析ツール

使用特許データベース
日本特許
・日立　Shareresearch
・発明通信社　HYPAT-i2
・NRIサイバーパテントデスク2
外国特許
・Questel 社Orbit.com

解析ツール
①テキストマイニング：Text Mining Studio(TMS)
②データマイニング：Visual Mining Studio(VMS)
③特許情報分析ツール：Patent Mining eXpress（PMX）
①～③はNTTデータ数理システム
④Questel 社Orbit.comのAnalysis module
⑤自作解析ツール
・PatAnalyzer 中国語/日本語解析ツール（C#2008)
・SimCalc1 類似度計算プログラム（VB.NET2008）
⑥R言語：統計解析、可視化
⑦Cytoscape：ネットワーク分析
⑧Excel , Excel VBA
⑨Python
⑩doc2vec,word2vec

テキストの自動分類とクラスタリング

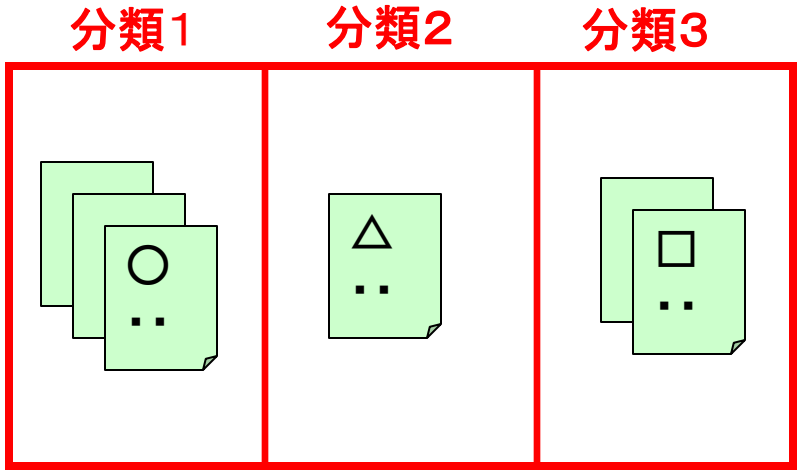
自動分類　　文書集合

軸　追加

クラス分類（注）

カテゴリによる分類表

分類1	分類2	分類3
○○○	△△△	□□□
○･･	△･･	□･･
○･･	△･･	□･･
○･･	△･･	□･･

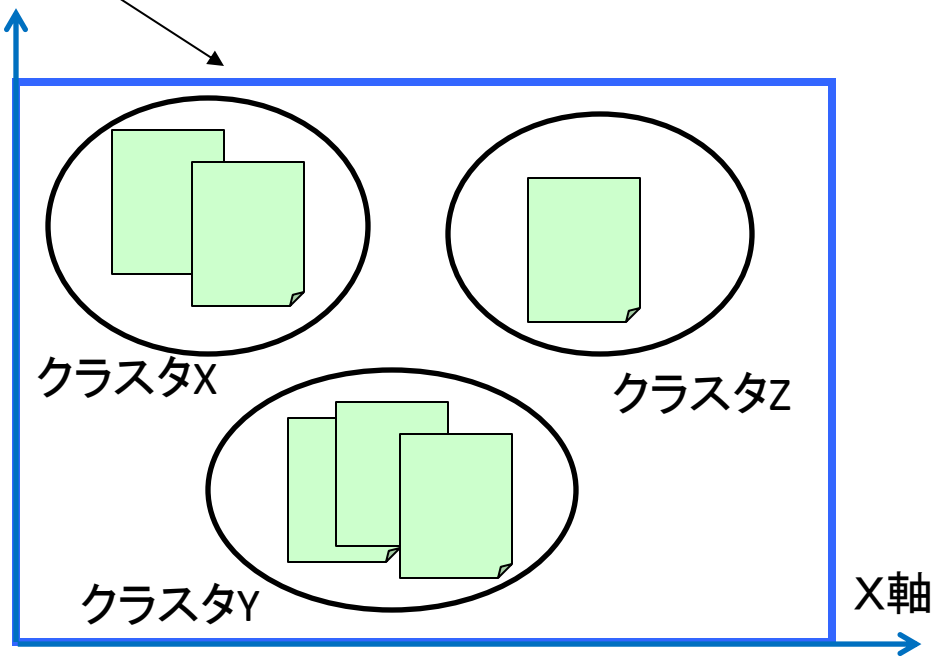


あらかじめ**決めたカテゴリ**に振り分ける

カテゴリ：IPC、特徴語

（注）クラシフィケーション、カテゴリゼーション

クラスタリング

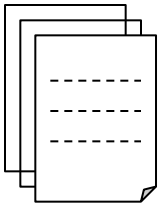


何らかの**類似度**で似た文書をまとめる

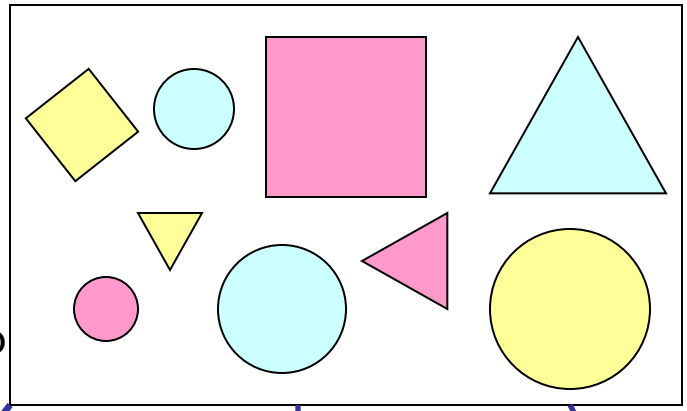
（観点の）

特徴

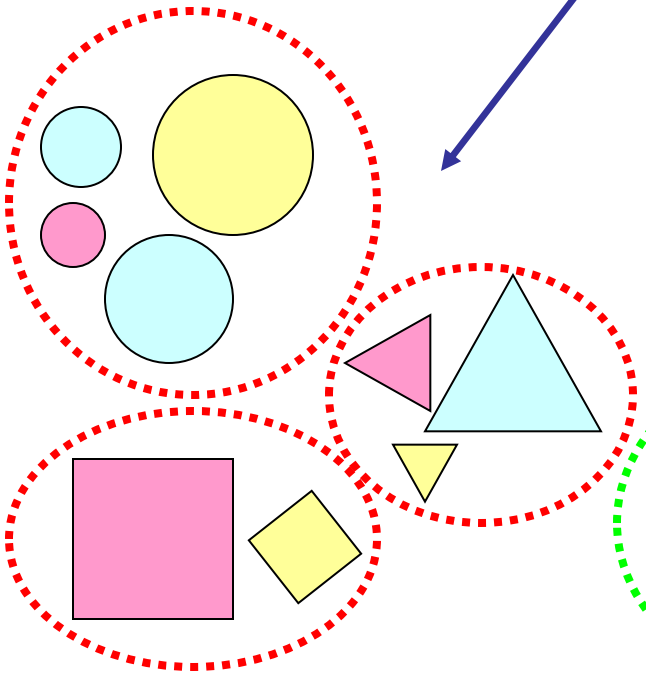
観点によるクラスタリングとは



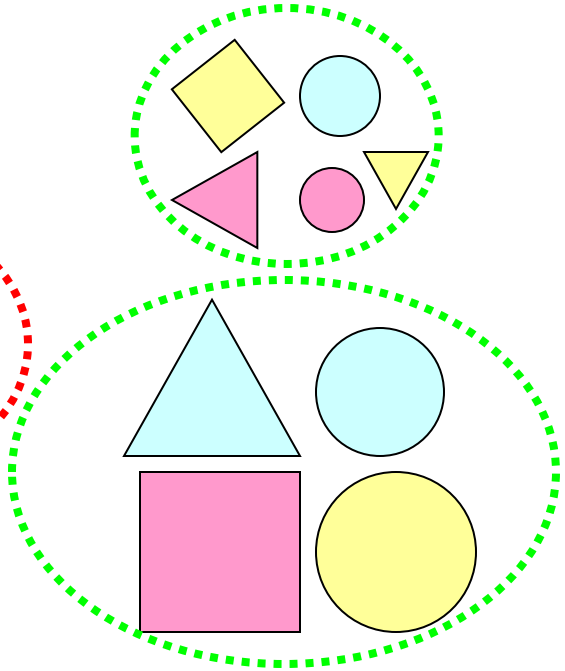
特許文書集合を文書間の
何らかの**類似度**に従って、
いくつかのグループに分ける



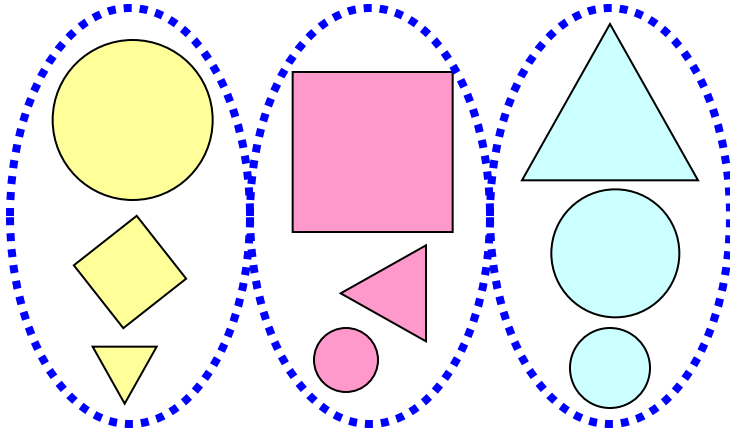
- **観点**によりクラスタリング結果が異なる（デッタッチメント）
- **類似度**の設定方法が多様（数値化方法が様々）
- 文書データをn次元ベクトルで表現
- クラスタリングには厳密な正解はない
- 人が行うデータ分析**支援**（気付きのためのツール）（セレンディピティ）



クラスタリング例1
観点：**形状**



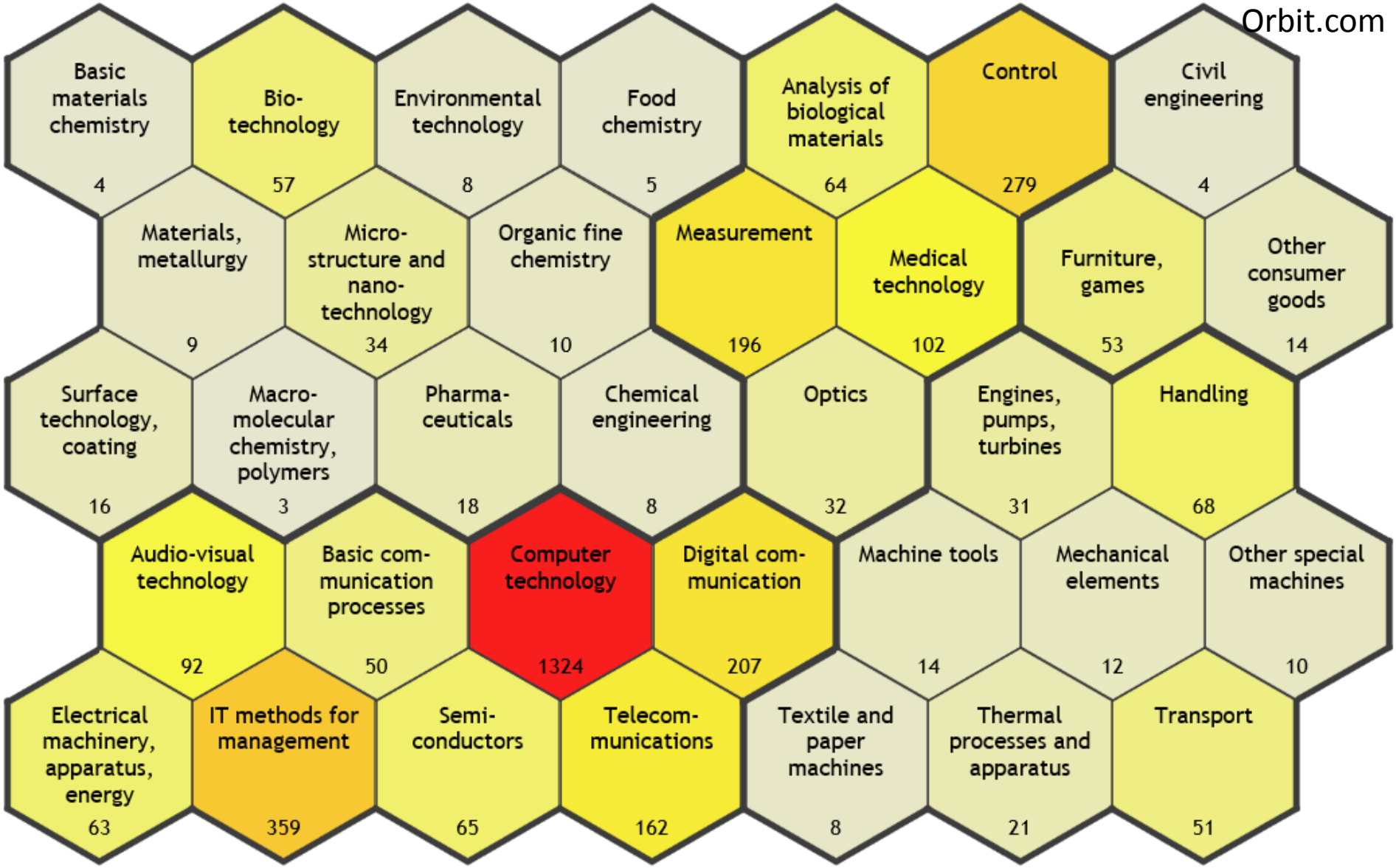
クラスタリング例2
観点：**サイズ**



クラスタリング例3
観点：**カラー**

IPCによるTechnology domainのヘキサゴンチャート

Orbit.com



・予め定められたIPC に基づいて公報をクラス分類
・技術領域としてComputer technology に集中している
・応用特許が幅広い分野に出願されている

・各Technology domain（ヘキサゴン：六角形）の位置は予め決まっており変わることはない
・ヘキサゴンの下部の数字はそこに属するファミリー数

コンセプトのタグクラウド（Orbit.com） INFOPRO2016

Computer (464) | Learning (292) | Probability (401) | Neural network (232) | Algorithm (431) | Database (438) | Behavior (318) | Classification (290) | Rule (339) | Search (297) | Vector (327) | Decision (292) | Subset (271) | Training data (153) | Computer readable medium (229) | Knowledge (231) | Recognition (231) | Category (247) | Computer readable storage (210) | Computation (210) | Learning data (90) | Inference (119) | Learning algorithm (115) | Attribute (216) | Update (218) | Estimation (223) | Computer program product (196) | Classifier (125) | Genetic algorithm (105) | Candidate (186) | Computer readable media (169) | Machine learning (96) | Neuron (124) | Language (176) | Architecture (209) | Likelihood (175) | Similarity (158) | Client (217) | Optical disc (173) | Storage media (143) | History (148) | Learning unit (63) | Computer readable instruction (122) | Map (194) | Correlation (185) | Sigma (177) | User interface (175) | Training (173) | File (216) | Confidence (116) | Information processing (108) | Probability distribution (91) | Management (250) | Cluster (144) | Bayesian network (66) | Score (126) | Computer system (152) | Variance (147) | Processor (157) | Computer executable instruction (109) | Decision tree (76) | Computing environment (124) | Readable storage (137) | Readable instruction (110) | Query (148) | Iteration (125) | Storage unit (152) | Storage medium (151) | Computing device (131) | Remote computer (113) | Threshold value (149) | Conditional probability (63) | Non transitory computer (108) | Scenario (143) | Recommendation (100) | Metadata (110) | Data structure (87) | Keyword (100) | Hidden layer (55) | Tree (128) | Medium (132) | Networking environment (84) | Resource (211) | Computer storage media (82) | Qubit (38) | Executable instruction (121) | Synaptic weight (37) | Artificial intelligence (61) | Optimization (144) | Hierarchy (96) | Qubits (35) | Web page (123) | Learning technique (53) | Microphone (143) | Outcome (95) | Artificial neural network (56) | Population (136) | Readable media (103) | Synapse (44) | Predicted (89) |

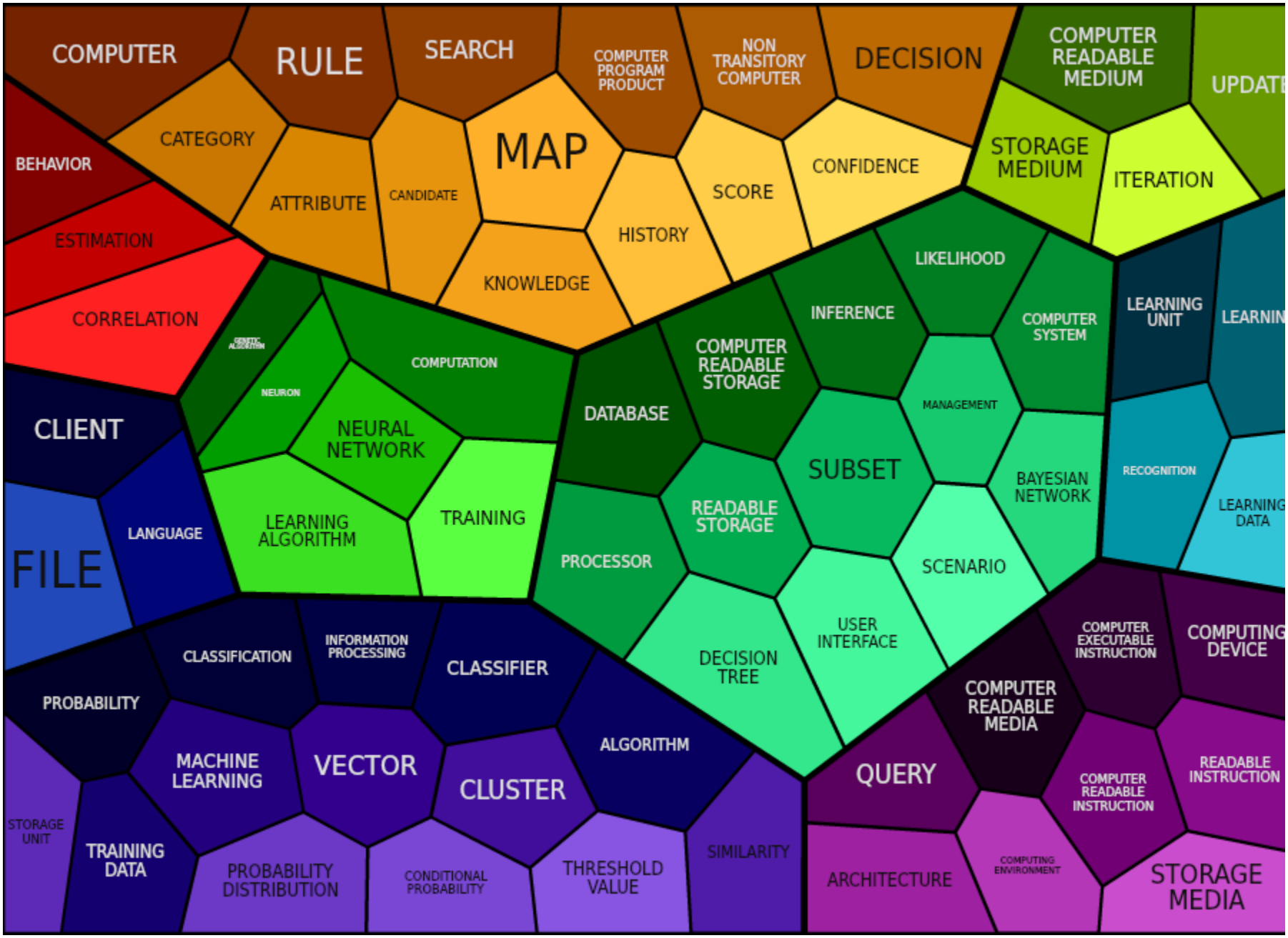
- コンセプトとはテキストマイニング的手法で公報より抽出されたテクニカルワード
- 対象集合全体あるいは個々の公報単位で表示可能
- テクニカルワードの頻度に比例して文字サイズを規定
- カッコ内の数字はコンセプトの該当公報数

コンセプトのドーナツチャート

Orbit.com

コンセプトのFoam Tree Chart

Orbit.com



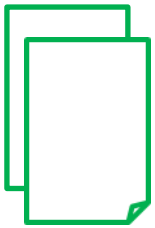
コンセプトによるLandscape map

Orbit.com

公報間の**類似度（距離）**による**クラスタリング**

専門用語による公報間相互類似度計算／Map作成フロー

分析対象公報

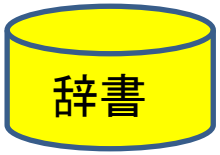


日本語検索
・NRI2
中国語検索
・日本版CNIPR
・Orbit（中国語）

抽出処理

PatAnalyzer（C#）

・形態素解析
・文字列抽出
・パターン抽出



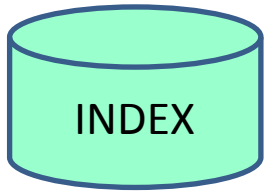
抽出パターン辞書
KW抽出辞書
ノイズ除去辞書

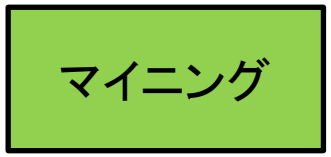
文書毎の抽出データ

KW1　頻度1
KW2　頻度2
・
・

解析ツール
・PatAnalyzer 中国語/日本語解析ツール（自作）
・MeCab：日本語形態素解析器2)
・saezuri lite（自然言語処理支援ライブラリ
・IKAnalyzerNet：中国語分詞ライブラリ
・SimCalc1 類似度計算プログラム（自作）
・R言語：統計解析5)
・Cytoscape：ネットワーク分析6)
・KH Coder　テキストマイニング

類似度計算プログラムSimCalc1（VB.NET）





・全文書間の非類似度
・抽出KW/文書番号
（インバーテッドファイル）

KW1　文書1,文書2
KW2　文書3,文書5,・・
・

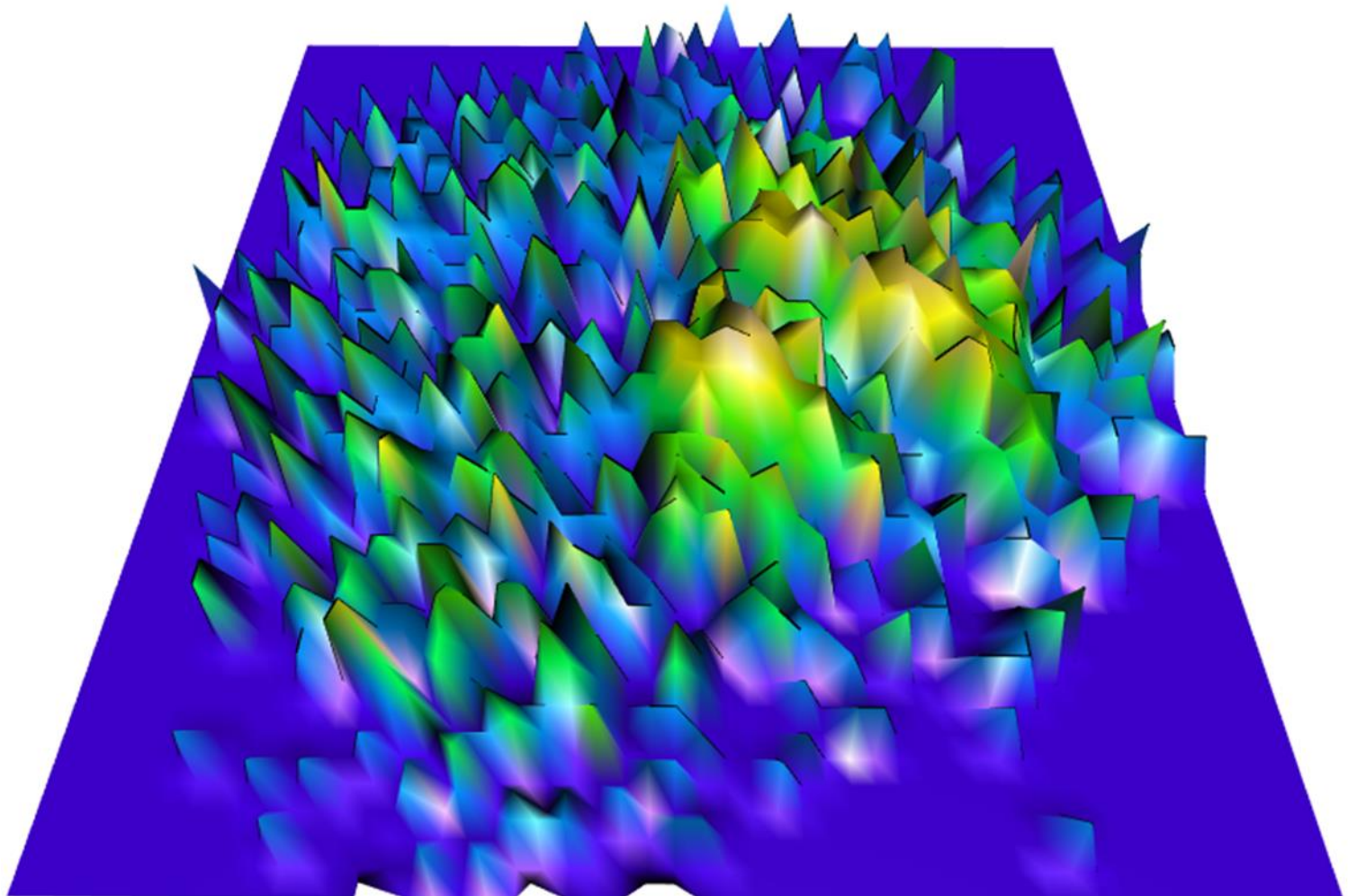
〇KW相互間の関係
〇文書相互間の関係

可視化/解析ツール
・ネットワーク分析
・R（多次元尺度法等）
・ Cytoscape

日本語の専門用語による公報間相互類似度計算Map

INFOPRO2016発表資料

各公報より専門用語抽出
各公報間の相互類似度（距離）計算
非計量多次元尺度法により座標計算（2D）
50×50メッシュで公報密度計算
公報密度を高さに変換し3D表面描画

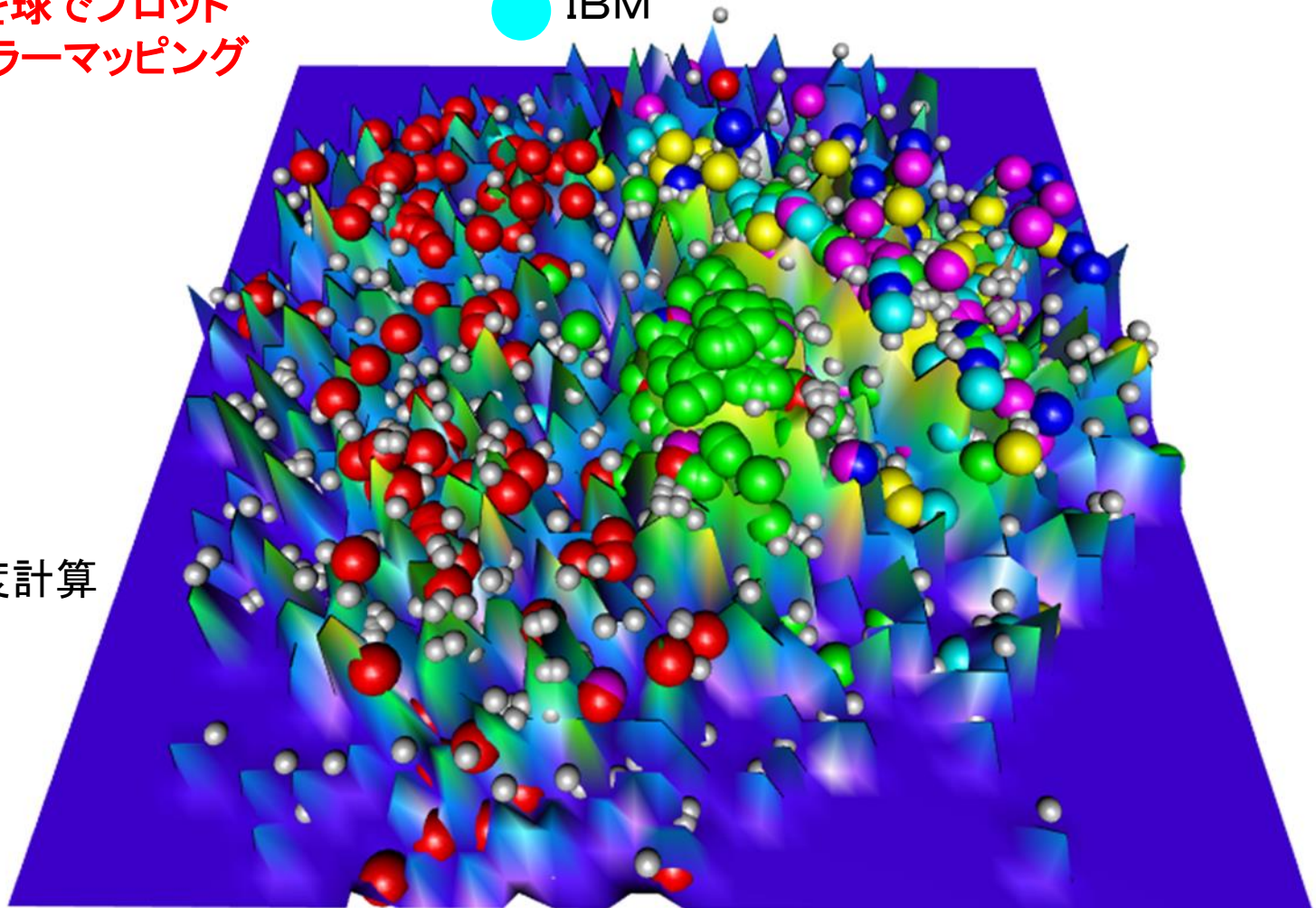


日本語の専門用語による公報間相互類似度計算Map

各公報より専門用語抽出
各公報間の相互類似度（距離）計算
非計量多次元尺度法により座標計算（2D）
50×50メッシュで公報密度計算
公報密度を高さに変換し3D表面描画
3D表面上に公報を球でプロット
特定の出願人をカラーマッピング

INFOPRO2016発表資料

- ソニー
- マイクロソフト
- クゥアルコム
- フィッシャーローズマウントシステムズ
- フィリップス
- IBM



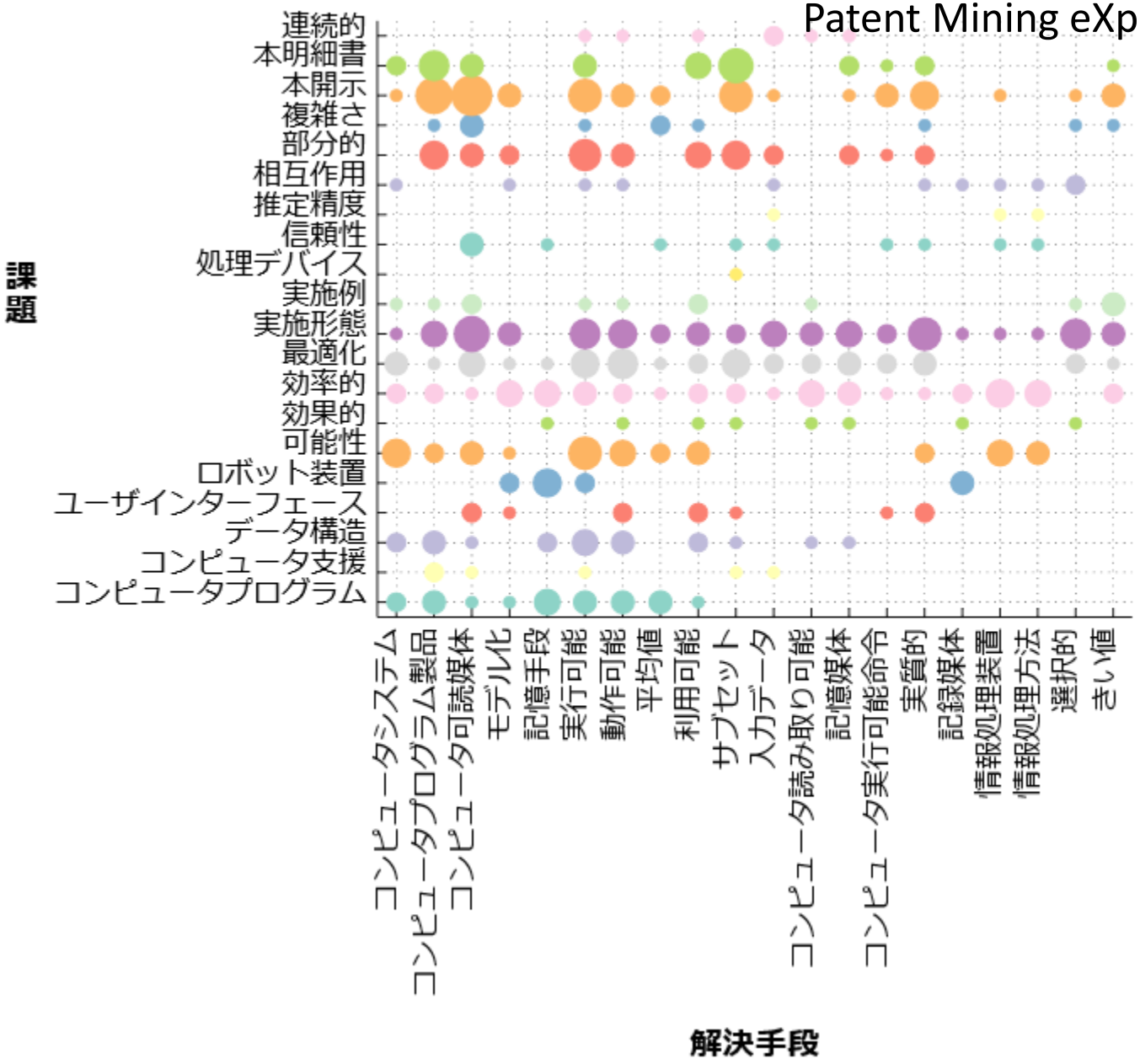
文書間相互類似度計算
文書数：1804
計算時間：92秒

Landscape mapの出願人別カラーマッピング

ソニー、マイクロソフトは
同様なクラスタリング傾向

PMXによる課題と解決手段

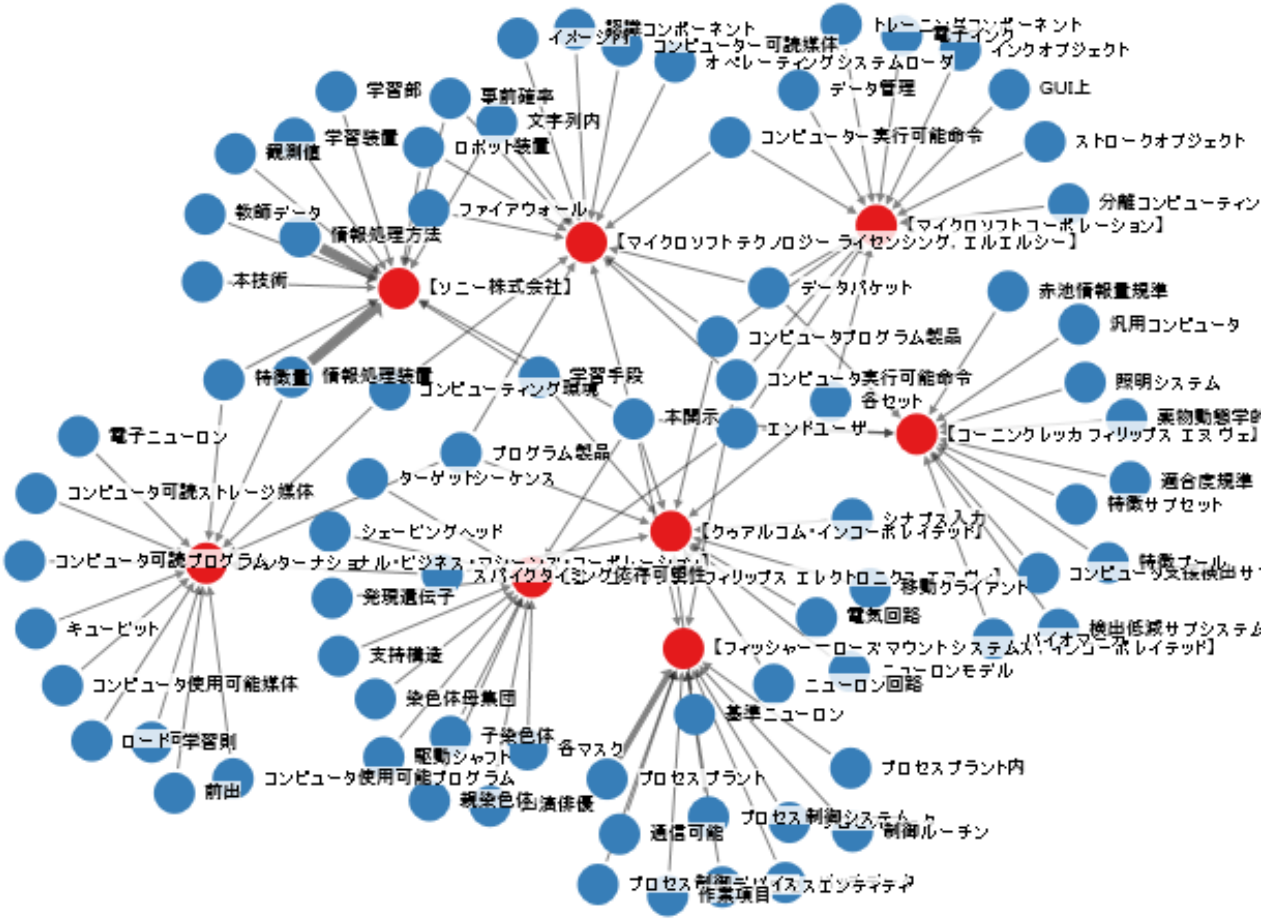
Patent Mining eXpress（PMX）



PMXによる技術特徴ネットワークグラフ

Patent Mining eXpress（PMX）

INFOPRO2016発表資料



動向調査パートの中間まとめ

①動向調査への教師データなし機械学習（特にクラスタリング）の応用
クラスタリングの特徴を理解して従来の解析手法と併用することで
実務上十分に有用である。

解析に当たっての注意点
・解析ツール（機能）を十分理解して使用することが重要
・解析したい内容に応じて各種ツールの特徴を使い分ける

解析ツール例
・書誌事項、KWの統計解析→パテントマップEXZ、Patent Mining eXpress（PMX）
・テキストマイニング　有償：Text Mining Studio(TMS)、無償：KH Coder
・データマイニング、機械学習　有償：Visual Mining Studio(VMS)、無償：R

クラスタリングの参考情報（今後検討予定）
・PLSA（確率的潜在意味解析法）：行（文書）と列（単語）を同時にクラスタリング
https://www2.deloitte.com/jp/ja/pages/deloitte-analytics/articles/analytics-plsa.html
・二項ソフトクラスタリング（VMS）
・トピックによるクラスタリング

トピックとは文（センテンス）の意味的内容で専門用語より大きなかたまり、
係り受け解析を利用して抽出できる。

教師データを用いた機械学習の先行技術調査フロー

INFOPRO2016発表資料

事例：即席麺

調査対象公報

826件

一部をサンプリング

学習データ

①機械学習

学習モデル構築

対象文献：A,T,S
1：名称+要約+請求項　めん+麺
2：名称+要約+請求項　即席+インスタント
3：公開・公表日：20010101:20101231
1*2*3　318件

318件をダウンロードして**引用文献**を抽出して1993年以降の公報を**追加**

計826件

10,100,500件の
3水準で検証

各公報の1:正解（当たり）／0:不正解（外れ）
フラグ（ラベル）

教師データ

教師データの
当たり（正例）は本願のみ入力！
外れ（負例）を3水準用意

826件の公報データ

②予測

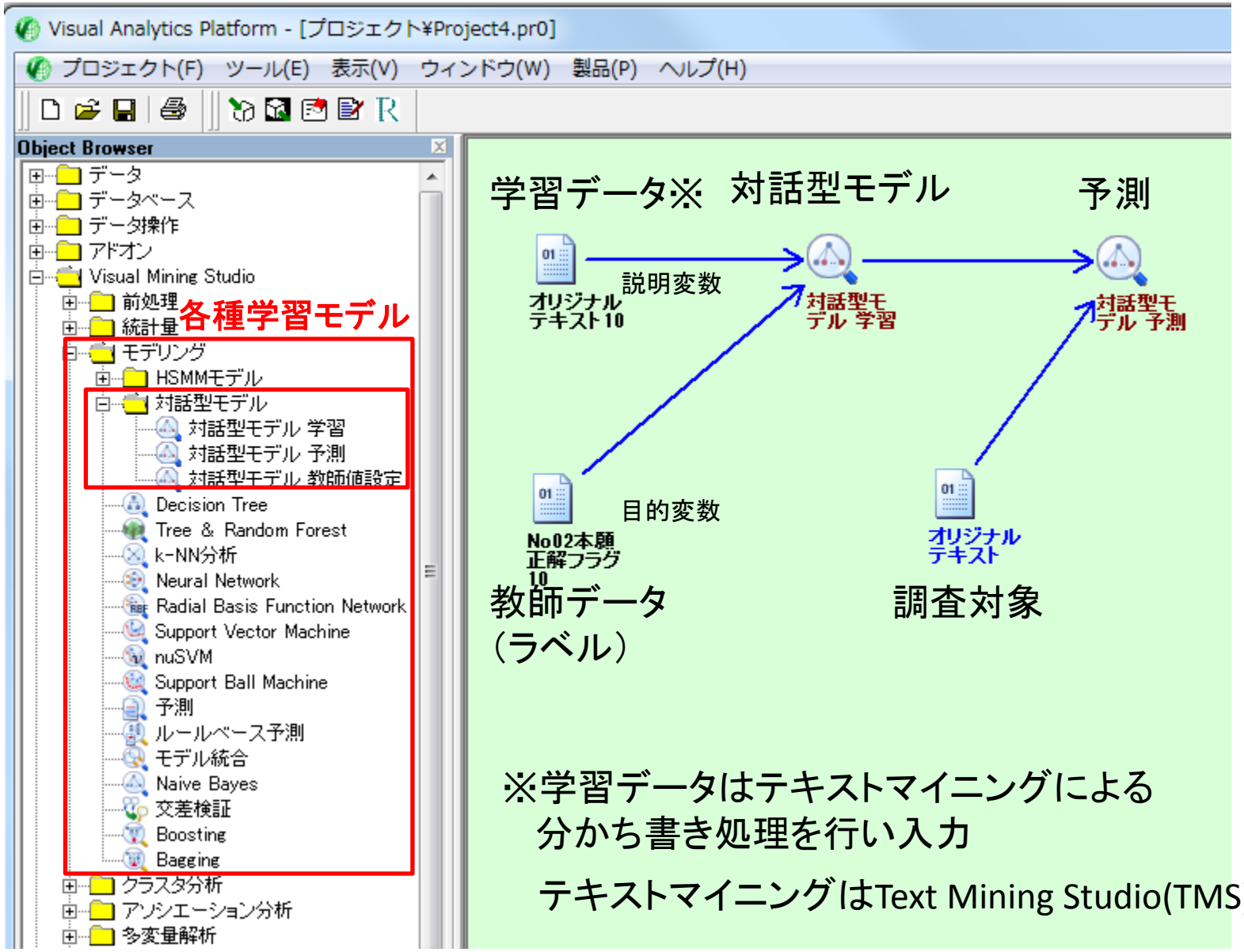
学習済モデル

審査官が引用しそうな
公報を0 or 1で予測（目的変数）
0:引用公報でない（外れ）
1:引用公報（当たり）

正解：審査官引用と答え合わせ

教師データを用いた機械学習ツールの設定画面

汎用データマイニングシステム：**Visual Mining Studio(VMS)**



※学習データはテキストマイニングによる分かち書き処理を行い入力

テキストマイニングはText Mining Studio(TMS)を使用

Text Mining Studio(TMS)の分かち書き出力例

INFOPRO2016発表資料

本願　Text Mining Studio(TMS)のテキストマイニング分かち書き出力例（デフォルト設定）

ファイルID	行ID	文章ID	単語ID	見出し語	原形	置換語	品詞	品詞詳細	係り先	述語属性	関係子
1	2	1	1	請求項	請求項	請求項	名詞	一般	2	なし	限定
1	2	1	2	1	1	1	名詞	数	3	なし	限定
1	2	1	3	炭酸カルシウム、	炭酸カルシウム	炭酸カルシウム	名詞	一般	10	なし	状況
1	2	1	4	燐酸カルシウム	燐酸カルシウム	燐酸カルシウム	名詞	一般	10	なし	状況
1	2	1	5	以下、	以下	以下	名詞	副詞可能	7	なし	状況
1	2	1	6	カルシウム剤と	カルシウム剤	カルシウム剤	名詞	一般	7	なし	現象
1	2	1	7	記す	記す	記す	動詞	自立	4	なし	注釈
1	2	1	8	及び	及び	及び	接続詞		9	なし	状況
1	2	1	9	ドロマイトから	ドロマイト	ドロマイト	名詞	一般	10	なし	状況
1	2	1	10	なる	なる	なる	動詞	自立	11	なし	限定
1	2	1	11	群から	群	群	名詞	一般	12	なし	状況
1	2	1	12	選ばれた	選ぶ	選ぶ	動詞	自立	20	なし	限定
1	2	1	13	少なくとも	少なくとも	少なくとも	副詞	一般	20	なし	状況
1	2	1	14	1種100重量	1種100重量	1種100重量	名詞	数	16	なし	限定
1	2	1	15	A	A	A	名詞	一般	14	なし	注釈
1	2	1	16	部に対し、	部	部	名詞	一般	20	なし	限定
1	2	1	17	加エデンプンを	加エデンプン	加エデンプン	名詞	一般	20	なし	現象
1	2	1	18	B	B	B	名詞	一般	17	なし	注釈
1	2	1	19	0．1～80重量	0．1～80重量	0．1～80重量	名詞	数	20	なし	限定
1	2	1	20	部含有させて	部含有	部含有	名詞	サ変接続	21	なし	状況
1	2	1	21	なることを	なる	なる	動詞	自立	22	なし	現象
1	2	1	22	特徴とする	特徴	特徴	名詞	一般	23	なし	限定
1	2	1	23	食品添加剤スラリー組成物。	食品添加剤スラリー組成物	食品添加剤スラリー組成物	名詞	サ変接続	-1	なし	なし

VMSに入力

注目特許（本願）P2009-258887　特開2010-29218　⇒　機械学習で審査官引用を予測する

【請求項1】炭酸カルシウム、燐酸カルシウム（以下、カルシウム剤と記す）及びドロマイトからなる群から選ばれた少なくとも1種（A）100重量部に対し、加エデンプン（B）を0．1～80重量部含有させてなることを特徴とする食品添加剤スラリー組成物。

分かち書き対象：要約+請求項
行ID:公報番号に相当
行IDと置換語をVMSに入力、説明変数として置換語を設定

目的変数：審査官引用を予測
0：引用しない　1：引用する

公報単位の機械学習と類似検索の比較結果

INFOPRO2016発表資料

事例：即席麺

対象文献：A,T,S
1：名称+要約+請求項　めん+麺
2：名称+要約+請求項　即席+インスタント
3：公開・公表日：20010101:20101231
1*2*3　**318件**

→ 318件をダウンロードして引用文献を抽出して1993年以降の公報を追加
計826件

注目特許（本願）：**特開2010-29218**
引用文献：**特開平7－111879|特開平6－125741|特開平6－197736|特開平6－245720|特開平11－113532**|
（特開昭61－242562を除く上記5件を正解として機械学習により予測を試みる

予測
0：外れ
1：当たり

教師データ数と予測結果

正解行ID	教師データ数 10	100	500
特開2010-29218	2（本願）	1	1	0
特開平11-113532	595	1	1	1
特開平7-111879	755	1	0	0
特開平6-245720	773	1	1	1
特開平6-197736	779	1	1	0
特開平6-125741	782	1	1	0
	0個数	48	516	806
当たりと予測→	1個数	778	310	20
	計	826	826	826

正解数	6	5	2
正解率	0.8%	1.6%	10.0%
漏れ率	0%	17%	67%

類似検索順位

HYPAT-i 請求項1	HYPAT-i 全請求項	NRI 請求項1	NRI 全請求項
		1	1
－	－	－	－
－	－	－	－
－	－	－	－
8	6	180	－
4	14	－	－

上位300位まで確認　上位1万位まで確認
－：圏外

優秀

- 教師データ数増加により**正解率（精度）向上**
- 教師データ数増加により**正解数は減少**
- 教師データ数増加により**漏れ増加**

文（センテンス）単位の機械学習結果とDB検索結果

教師データ数と予測結果（文単位）

正解行ID	教師データ数 126文	1323文	5797文
特開2010-29218 2本願	3	2	0
特開平11-113532 595	6	5	1
特開平7-111879 755	6	1	0
特開平6-245720 773	5	3	3
特開平6-197736 779	15	6	1
特開平6-125741 782	5	3	0
0個数	3786	7603	8663
1個数	5008	1191	131
計	8794	8794	8794

←文の合計8794

正解数	6	6	3
正解率	0.1%	0.5%	2.3%
漏れ率	0%	0%	50%

文書単位の概算値

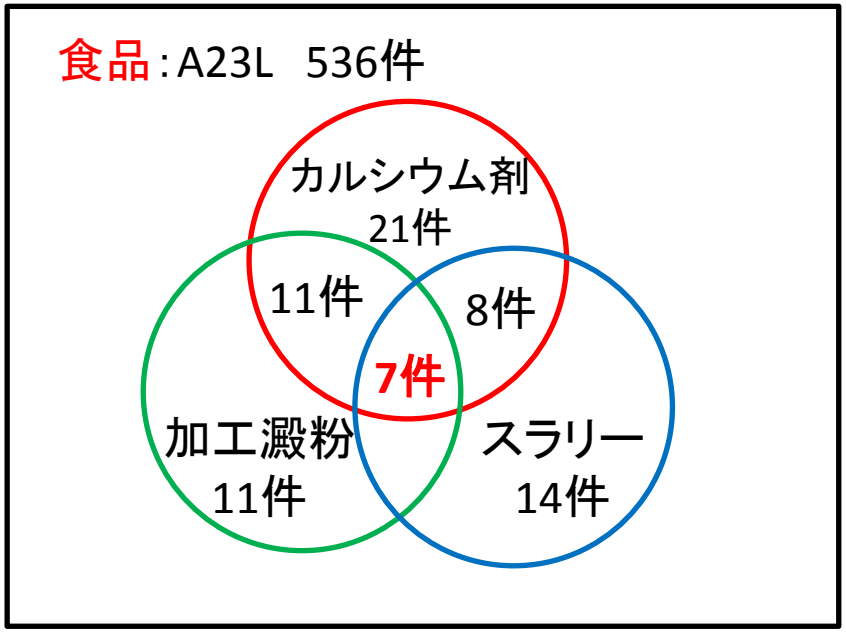
母集団：即席麺826件

INFOPRO2016発表資料

構成要件数	構成要件
4	カルシウム剤、加工デンプン、食品、スラリー
2	カルシウム剤、食品
3	カルシウム剤、食品、スラリー
3	カルシウム剤、デンプン、食品、スラリー（糊状）
3	カルシウム剤、食品、スラリー
3	カルシウム剤、食品、スラリー

DB検索結果：7件

本願特開2010-29218
特開2002-186458
特開2001-186863
特開2001-178412
引用特開平7-111879
引用特開平6-197736
引用特開平6-125741



母集団：即席麺826件

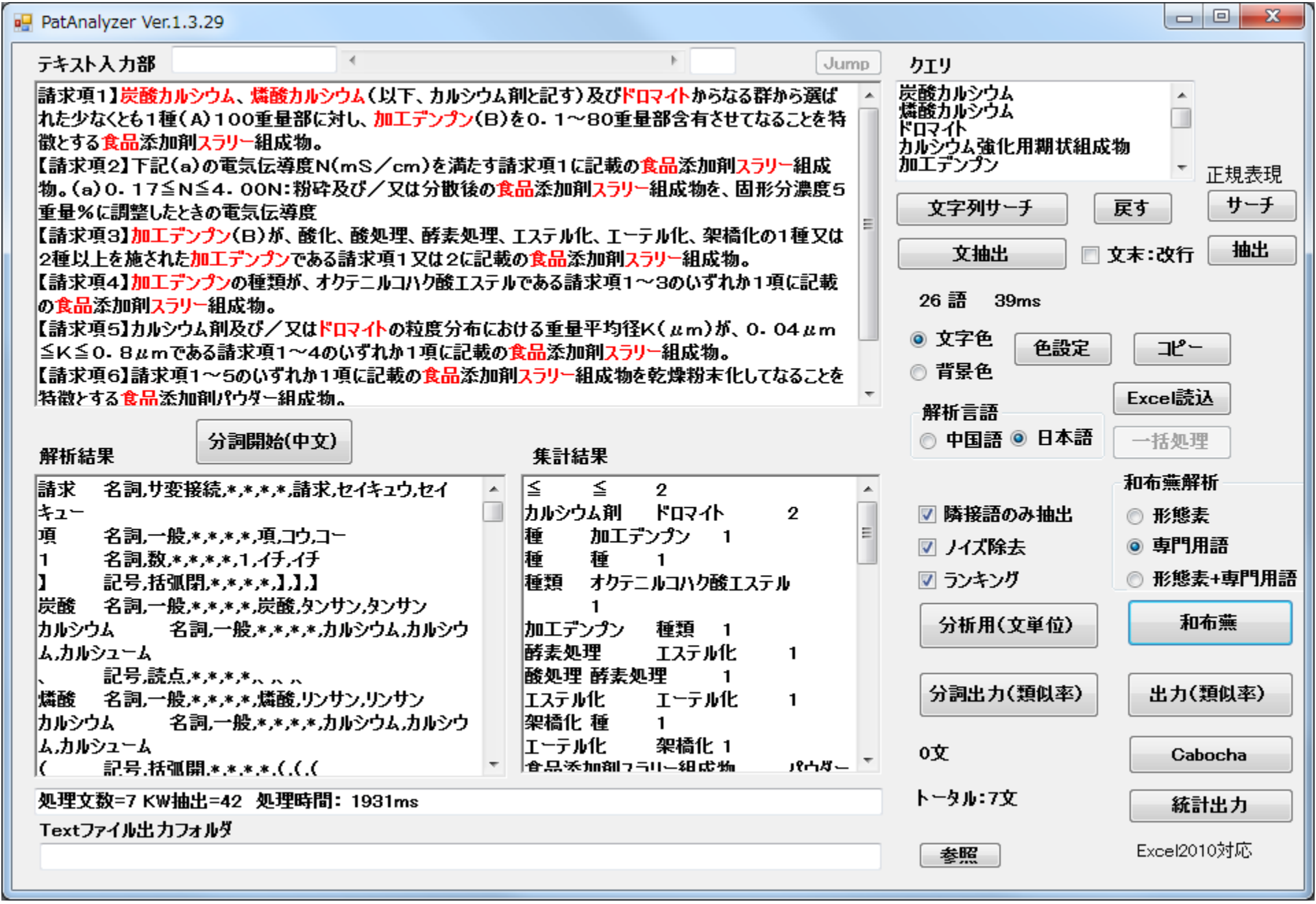
各構成要素のブーリアン演算

カルシウム剤
炭酸カルシウム
燐酸カルシウム
リン酸カルシウム
ドロマイト
4B018MD04 ・・・カルシウム

加工澱粉
加工澱粉
加工デンプン
加工でんぷん

スラリー
スラリー

PatAnalyzer画面



まとめと考察

機械学習を利用した効率的な特許調査方法を実務ベースに重きを置いて
①動向調査と、②先行技術調査について検討した。

まとめ

①動向調査への教師データなし機械学習の応用
書誌事項の統計解析（パテントマップソフト等）と併用することで実務上十分に有用である。

②先行技術調査への教師データあり機械学習の応用
- 教師データ（正解）の準備が課題
- 教師データを公報（文書）単位とすると審査官引用等があるものは準備は容易だが機械学習の精度は良くない
- 教師データを文あるいは段落単位とすると機械学習の精度は上がるが教師データの準備自体が課題　→　スコアリングツール作成を検討
- TF-IDFによる文書の（コサイン）類似度でなく新規性の観点に適合するように特徴語の重み付けを行うとスコアリング精度が向上すると考えられる

考察

特徴語の重みを機械学習により調整して類似度計算を行うとさらにスコアリング精度改善の余地が大きいと考えられる

今後の予定

- 新規性の観点に適合した機械学習を利用した新規性評価関数の最適化検討

ノーフリーランチ定理（NFL定理）

Wikipedia

https://ja.wikipedia.org/wiki/ノーフリーランチ定理

YEARBOOK2017

ノーフリーランチ定理（no-free-lunch theorem、NFLT）は、物理学者 David H. Wolpert と William G. Macready が生み出した組合せ最適化の領域の定理である。その定義は以下のようになる。
コスト関数の極値を探索するあらゆるアルゴリズムは、全ての可能なコスト関数に適用した結果を平均すると同じ性能となる
― Wolpert and Macready、1995年

この定理は「**あらゆる問題で性能の良い汎用最適化戦略は理論上不可能**であり、**ある戦略が他の戦略より性能がよいのは、現に解こうとしている特定の問題に対して特殊化（専門化）されている場合のみ**である」ということを立証している（Ho and Pepyne、2002年）。

工学者や最適化の専門家にとって、この定理は、**問題領域の知識を可能な限り使用して最適化すべき**だということを示しており、**領域を限定して特殊な最適化ルーチンを作成すべきである**ことを示している。

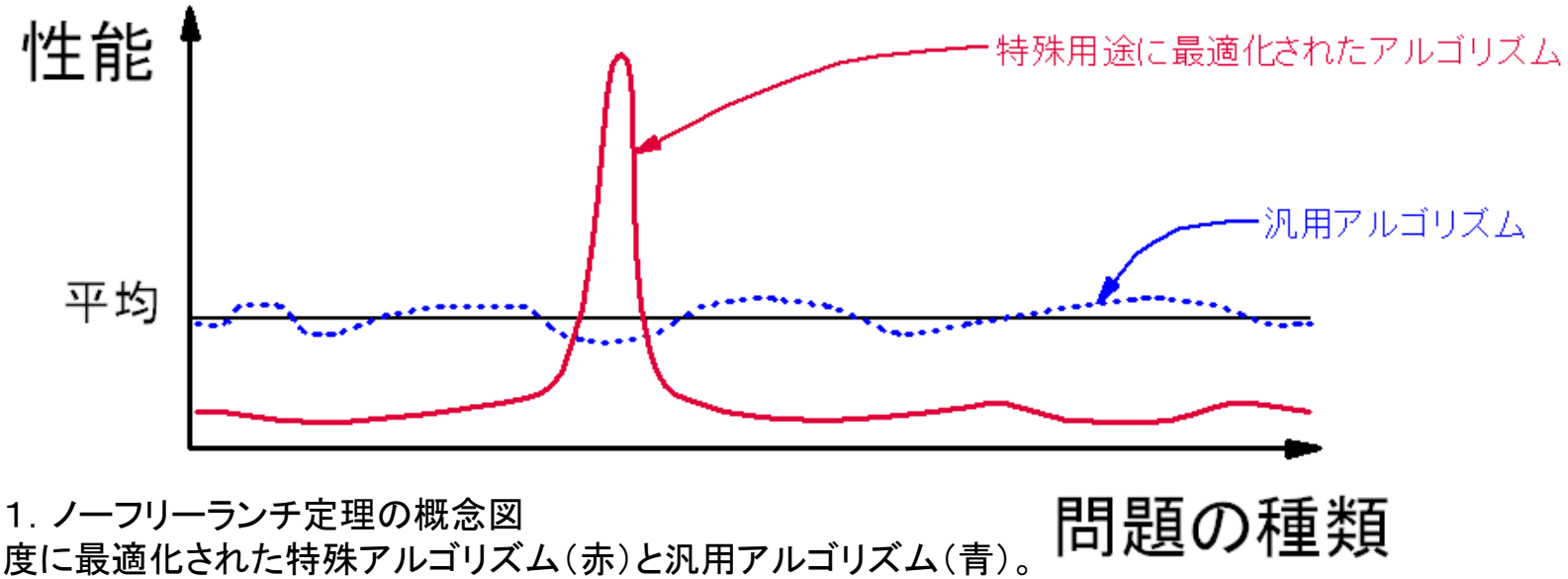


図1．ノーフリーランチ定理の概念図
高度に最適化された特殊アルゴリズム（赤）と汎用アルゴリズム（青）。
どちらも平均すれば同程度の性能となることに注意。

先行技術調査の流れ（進め方）

出願したい明細書から構成要素を分析する

特許検索競技大会2016
フィードバックセミナー資料p35

明細書を熟読して発明内容を理解し、検索式作成のための
構成要素を決定する

↓

予備検索の実行

特許分類（FI、Fターム、IPC）、キーワードの検討
海外の場合（IPC,CPC）

↓

検索戦略立案、検索式作成

検索式に使用する特許分類、キーワードの抽出
多観点の検索式の検討

↓

スクリーニング過程を詳細に検討し、
機械学習を応用した支援方法（ツール）検討

検索実行、**スクリーニング**

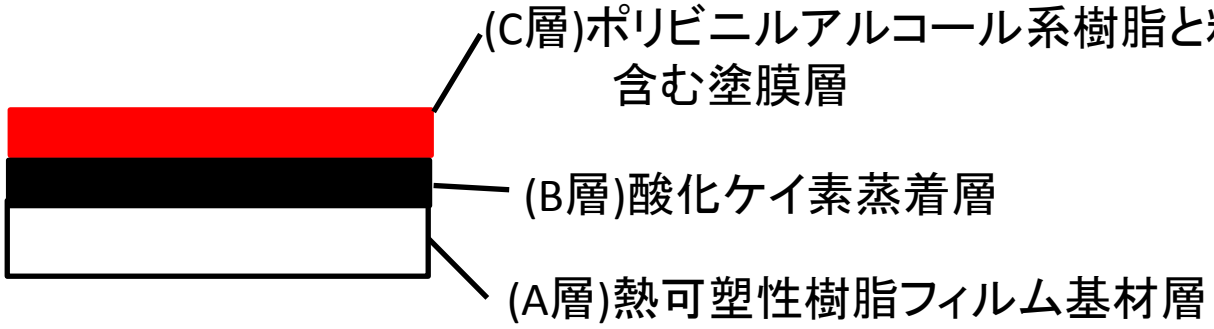
優先順位を決め、効率的にスクリーニングを行う
スクリーニング結果に応じて、検索戦略を再検討

図2．先行技術調査の流れ

先行技術調査の事例検討

特許検索競技大会2016　化学・医薬分野
出題内容:【問2】問題文概要(2/3)

【特許請求の範囲】
【請求項1】
熱可塑性樹脂フィルム基材層(A層)、酸化ケイ素蒸着層(B層)、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層(C層)が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。

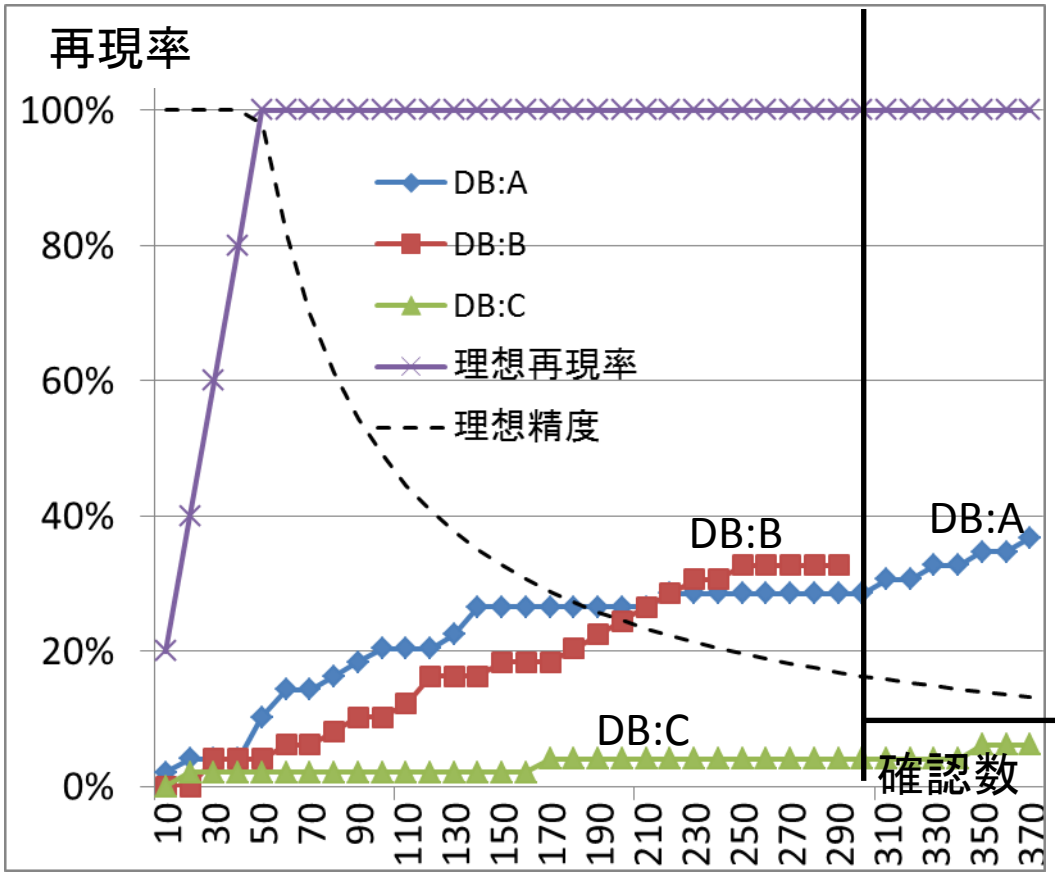


ガスバリア性包装用フィルム

図3．特許検索競技大会2016の化学・医薬分野の問2

商用データベースの概念(類似)検索の再現率比較

YEARBOOK2017



正解順位

No,	A	B	C
1	10	22	11
2	14	23	170
3	41	51	347
4	43	71	
5	47	84	
6	53	105	
7	59	116	
8	76	117	
9	81	145	
10	95	177	
11	129	182	
12	134	199	
13	140	208	
14	213	217	
15	309	226	
16	322	248	
17	342		
18	363		

確認数:300　正解数:49

精度	4.7%	5.3%	0.7%
再現率	28.6%	32.7%	4.1%
F値	0.08	0.09	0.01

図4. 商用データベースの概念(類似)検索の再現率比較

実験用データセットの作成

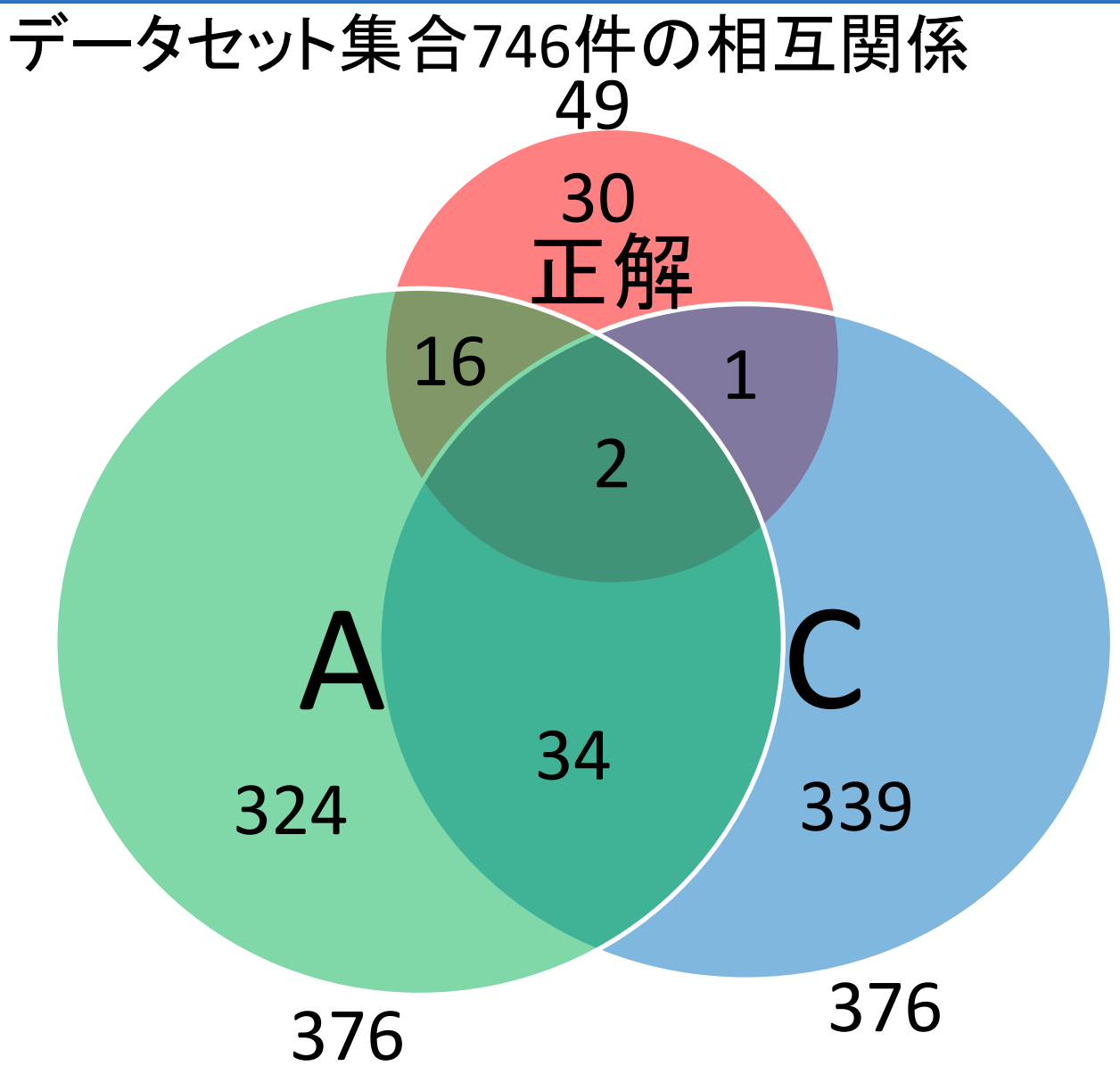


図5．データセット集合746件の相互関係

分かち書きと重み付けの再現率への影響

YEARBOOK2017

分かち書き（形態素、専門用語）と重み付け（TF 、TF・IDF）の再現率への影響

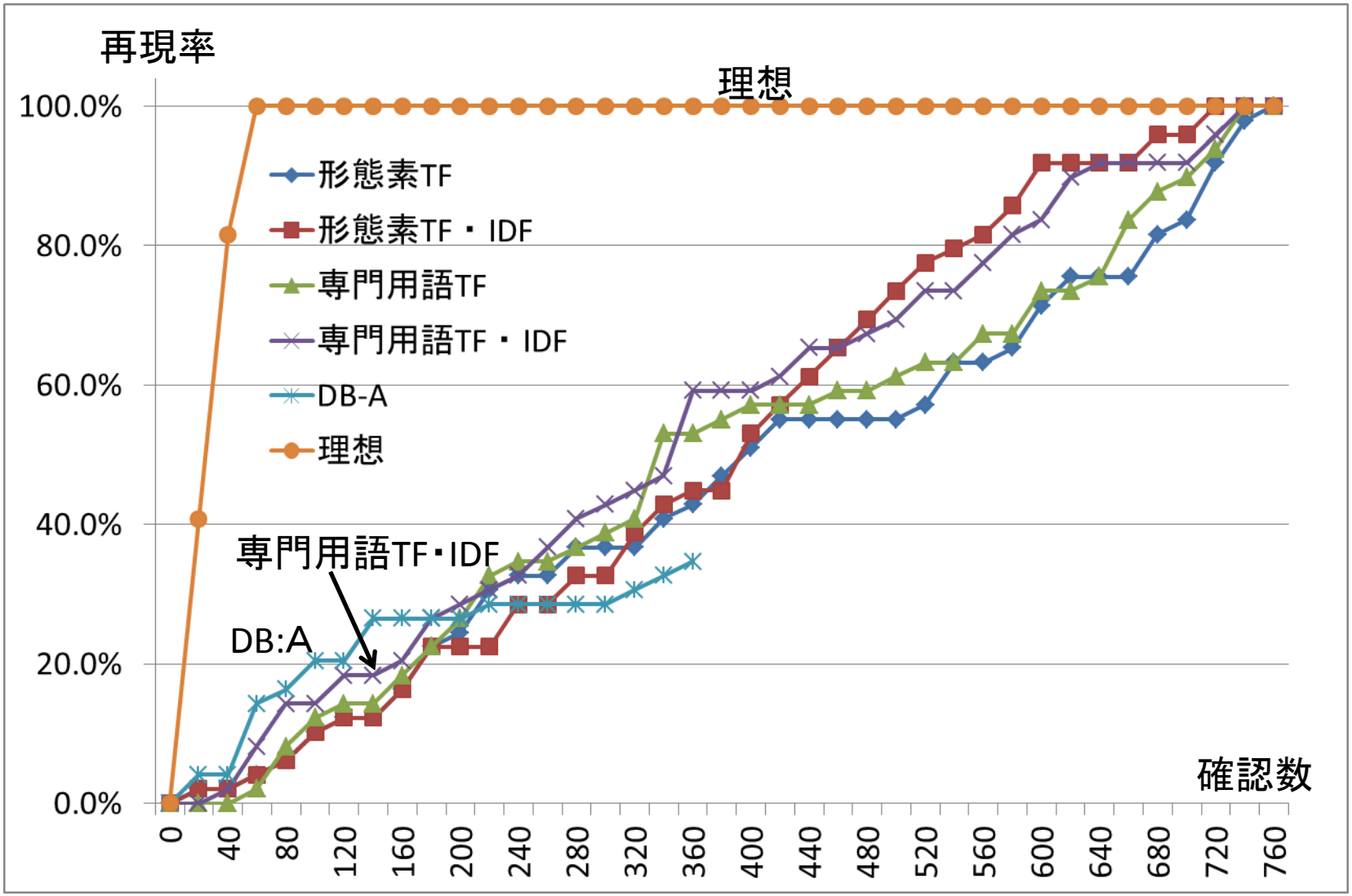


図6. 分かち書きと重み付けの再現率への影響

形態素と専門用語による分かち書き

YEARBOOK2017

熱可塑性樹脂フィルム基材層、酸化ケイ素蒸着層、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。

熱	名詞,一般,*,*,*,*,熱,ネツ,ネツ
可塑	名詞,一般,*,*,*,*,可塑,カソ,カソ
性	名詞,接尾,一般,*,*,*,性,セイ,セイ
樹脂	名詞,一般,*,*,*,*,樹脂,ジュシ,ジュシ
フィルム	名詞,一般,*,*,*,*,フィルム,フィルム,フィルム
基	名詞,一般,*,*,*,*,基,モト,モト
材	名詞,接尾,一般,*,*,*,材,ザイ,ザイ
層	名詞,接尾,一般,*,*,*,層,ソウ,ソー
、	記号,読点,*,*,*,*,、,、,、

図7．形態素解析（MeCab）による分かち書き（一部）

熱可塑性樹脂フィルム基材層
酸化ケイ素蒸着層
ポリビニルアルコール系樹脂
粘土鉱物
塗膜層
他
層
積層
特徴
ガスバリア性包装用フィルム

図8．専門用語による分かち書き

N-グラムの文字数Nと重み付けの影響

Nグラムの文字数Nと重み付け（2値、重み TF）の再現率への影響

YEARBOOK2017

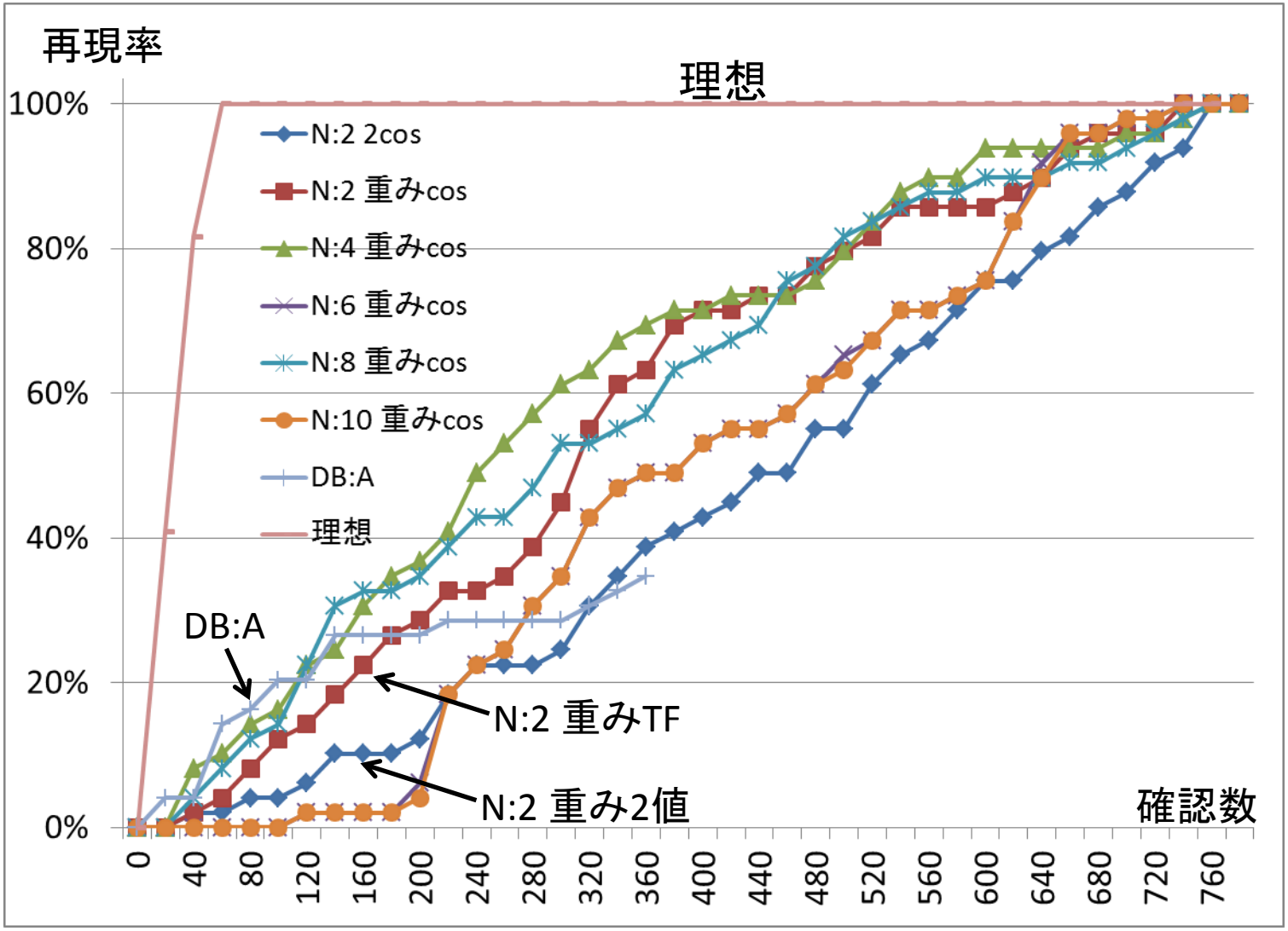


図9. N-グラムの文字数Nと重み付けの影響

構成要素分析（検索競技大会の模範解答例）

YEARBOOK2017

熱可塑性樹脂フィルム基材層、酸化ケイ素蒸着層、ポリビニルアルコール系樹脂と粘土鉱物を含む塗膜層が他の層を介して又は介さずにこの順に積層されてなることを特徴とするガスバリア性包装用フィルム。

正解例と解説:【問2】(1)構成要素分析

(1)調査依頼された請求項1に対して、検索すべき技術の構成要素(概念)を記述しなさい。

記号	構成要素(概念)
a	熱可塑性樹脂フィルム基材層
b	酸化ケイ素蒸着層
c	ポリビニルアルコール系樹脂を含む塗膜層
d	塗膜層に粘土鉱物を含む
e	他の層を介してまたは介さずにこの順に積層
f	ガスバリア性
g	包装用フイルム

※構成要素の分け方は本例に限定しない

図10. 構成要素分析（検索競技大会の模範解答例）

Fタームと形態素TF類似度による評価関数

YEARBOOK2017

要素	b1 酸化ケイ素	b2 蒸着	c PVA	d 粘土鉱物	f ガスバリア	g 包装用フィルム
FI	B32B9/00@A		B32B27/30,102			
Fターム	4F100AA20	4F100EH66	4F100AK21 4F100AK69	4F100AC03 4F100AD01	4F100JD02	4F100GB15

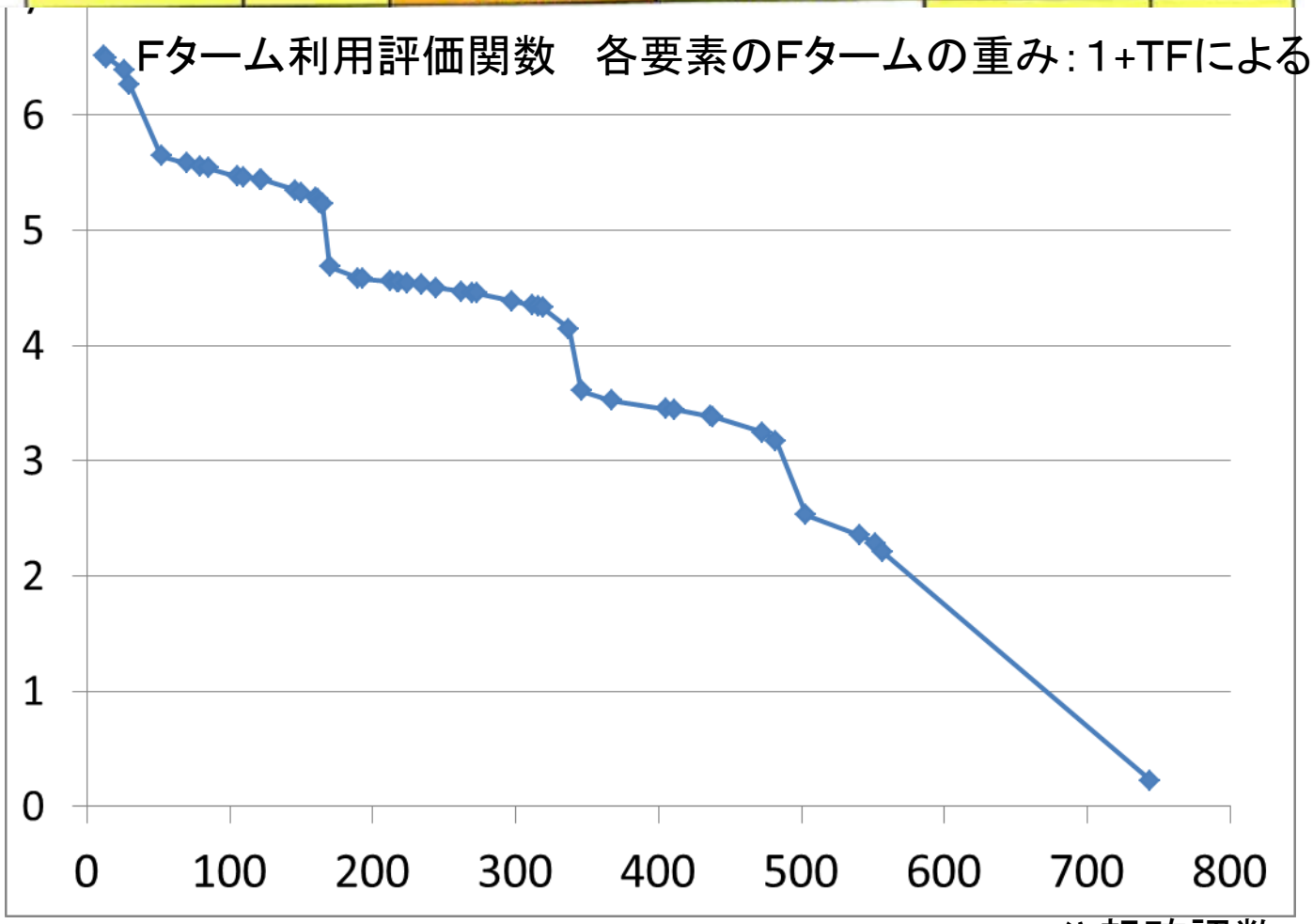


図11．Fタームと形態素TF類似度による評価関数

評価関数とフィルターの影響

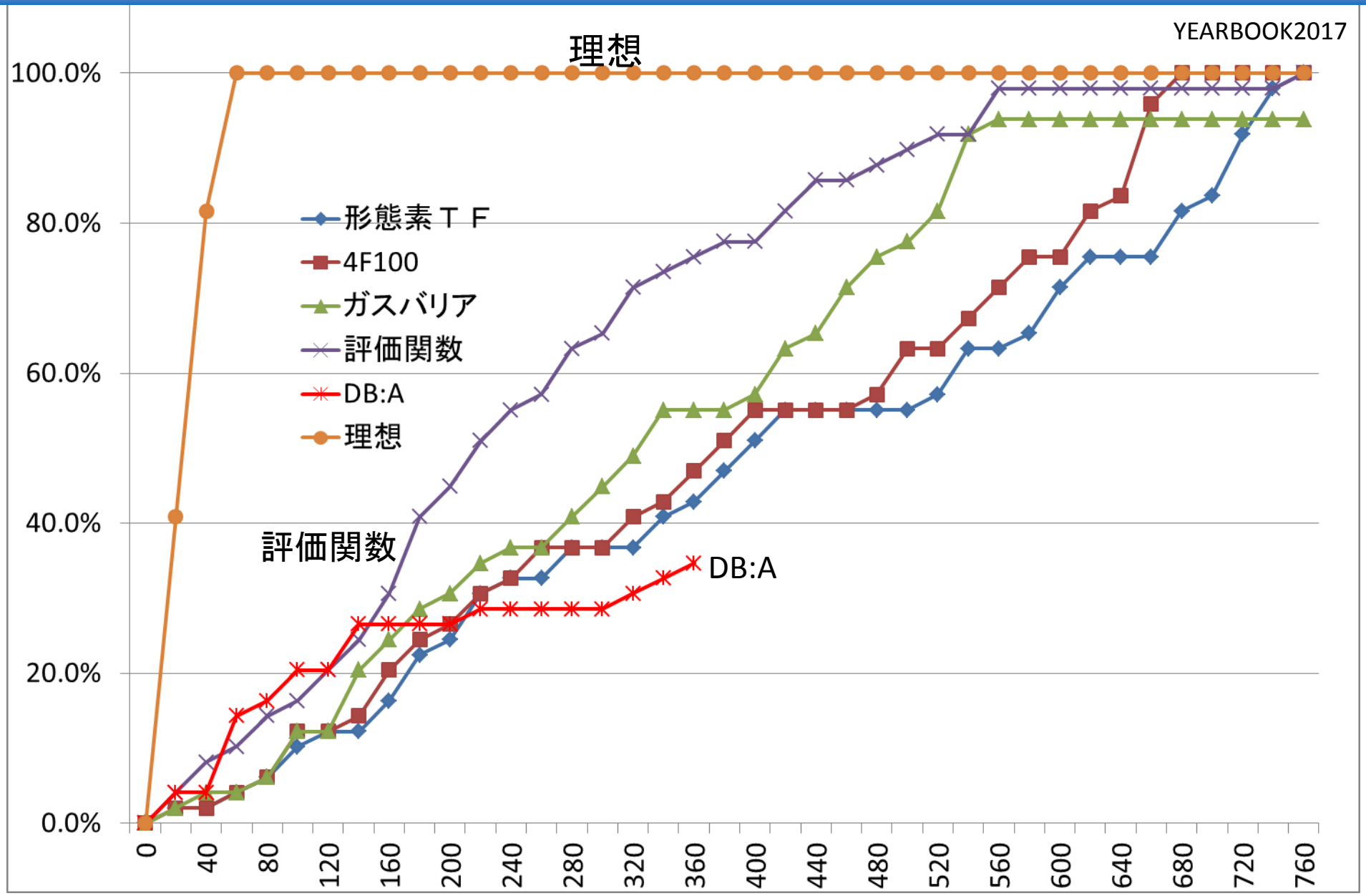


図12. 評価関数とフィルターの影響

doc2vecによる文書のベクトル化処理の概要

YEARBOOK2017

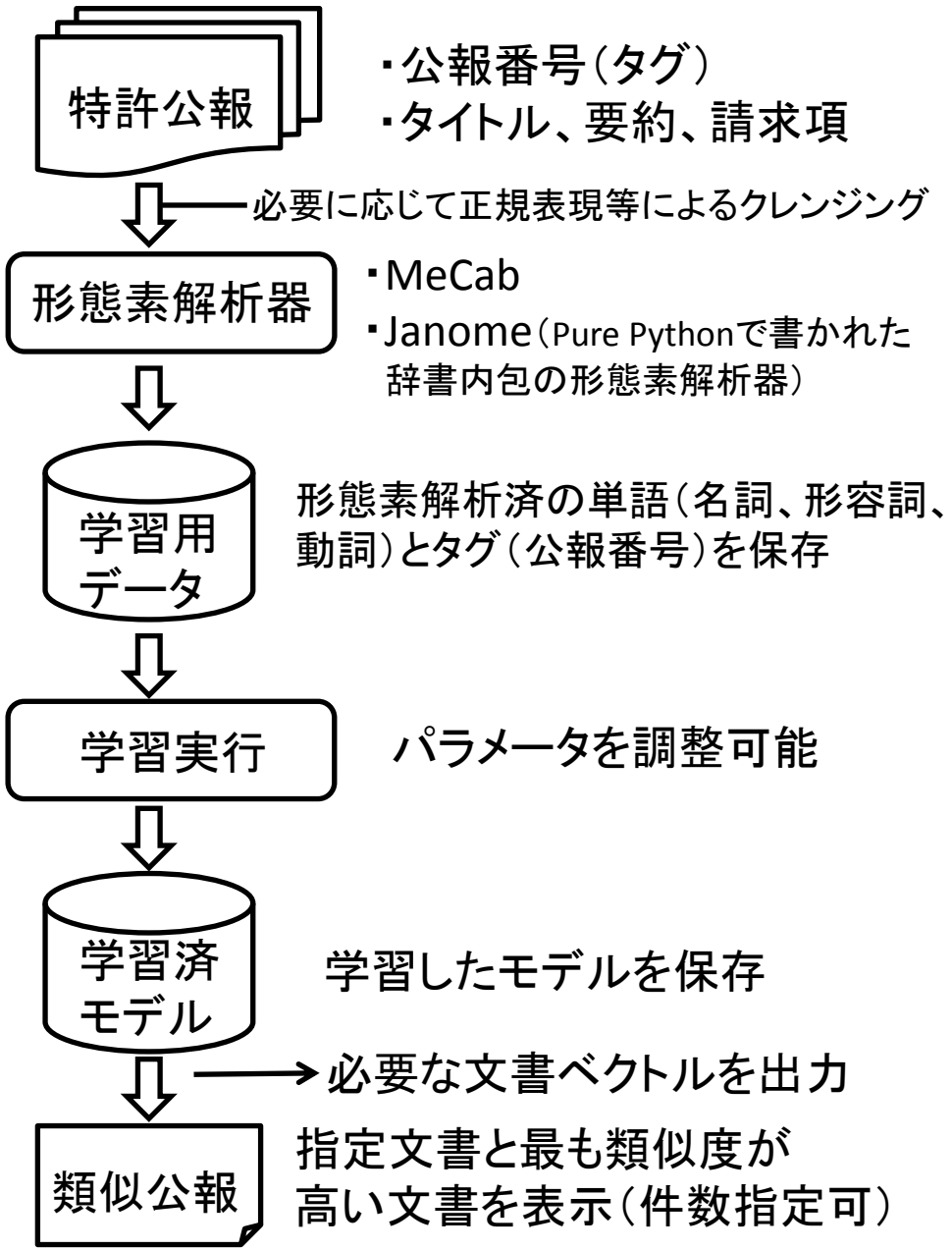


図13．Doc2vecによる文書のベクトル化処理の概要

文書の分散表現ベクトルの学習モデルと再現率

YEARBOOK2017

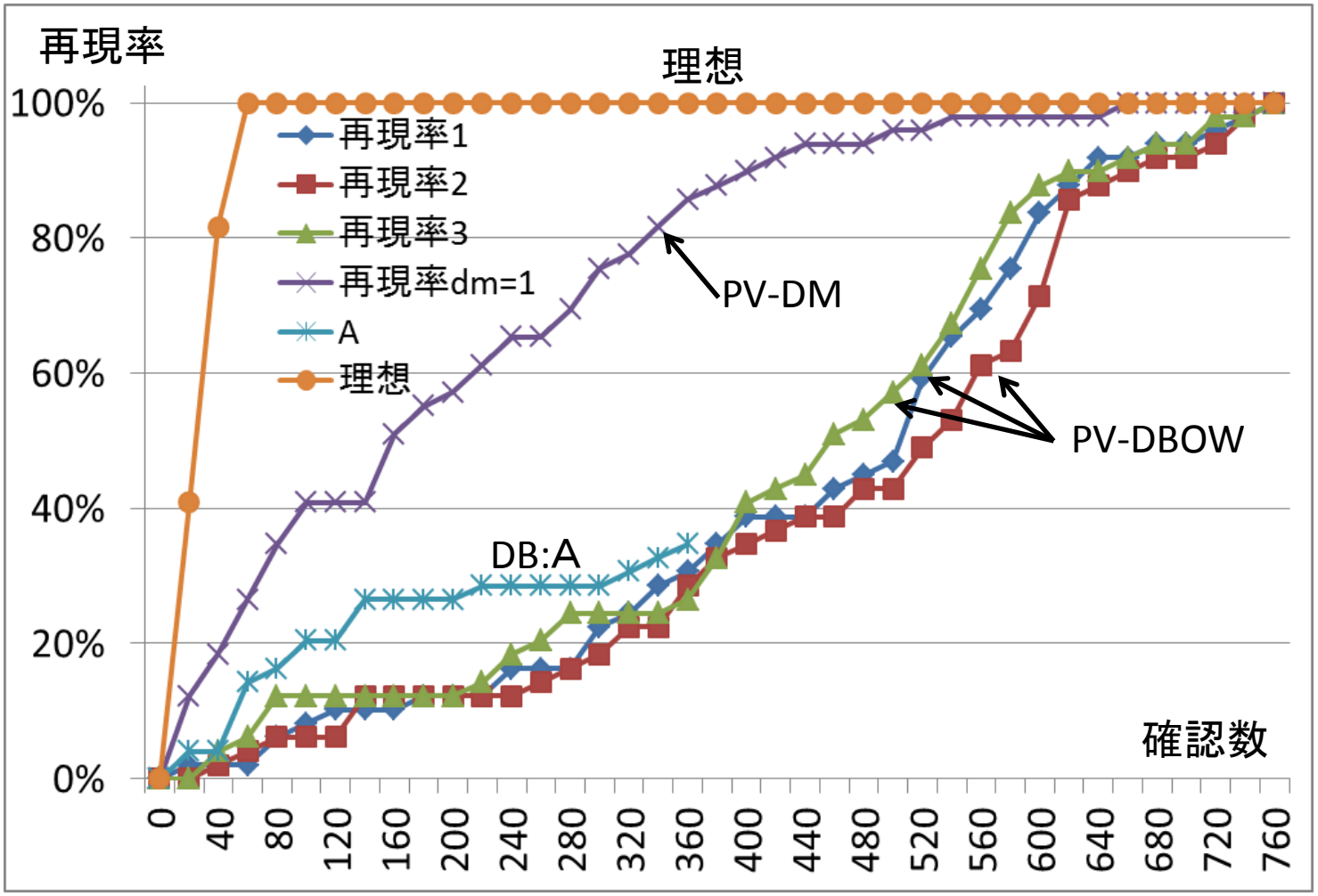


図14. 文書の分散表現ベクトルの学習モデルと再現率

文書の分散表現ベクトルの次元数（Size）の影響

YEARBOOK2017

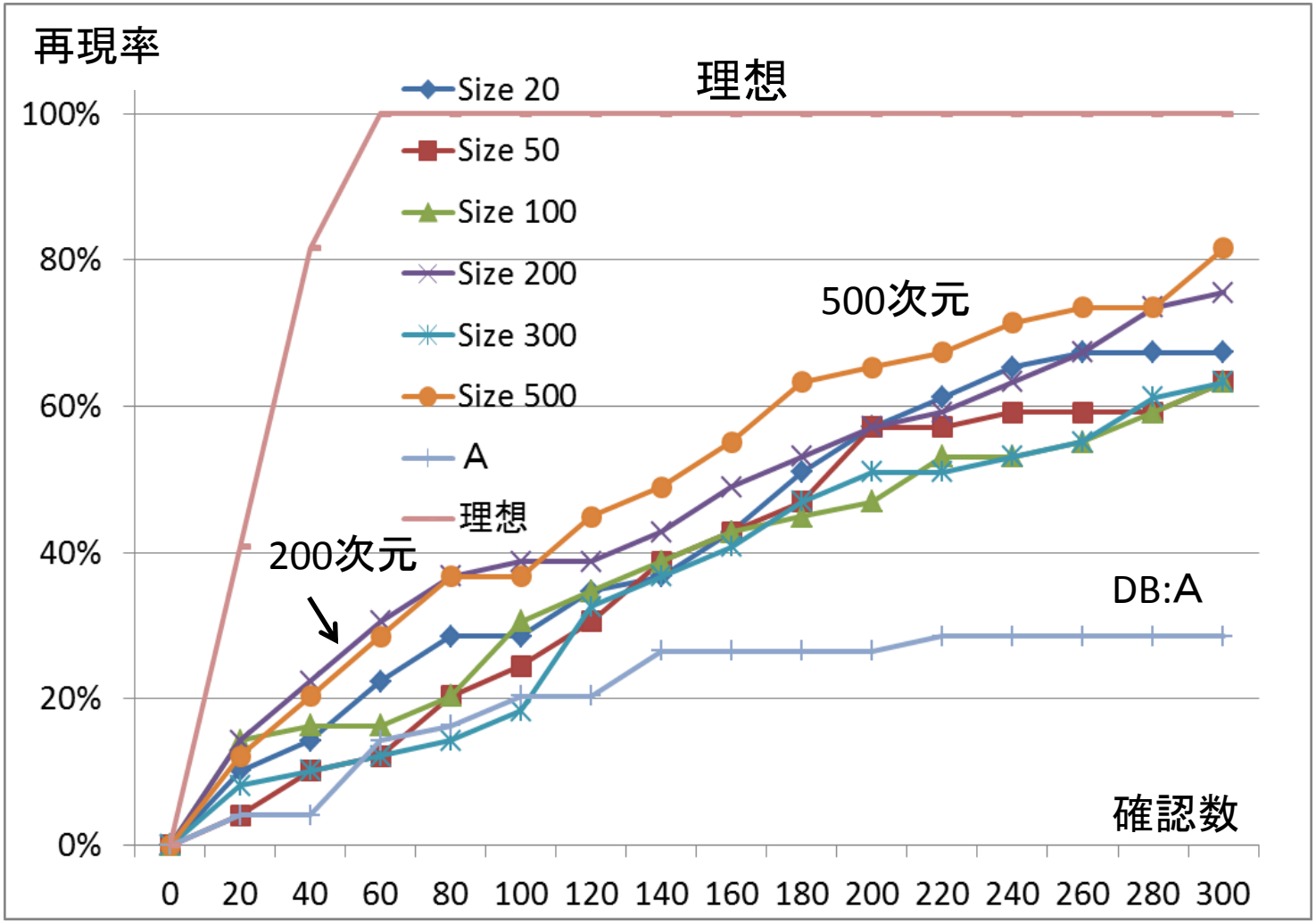


図15．分散表現ベクトルの次元数（Size）の影響

非計量多次元尺度法による各公報の可視化

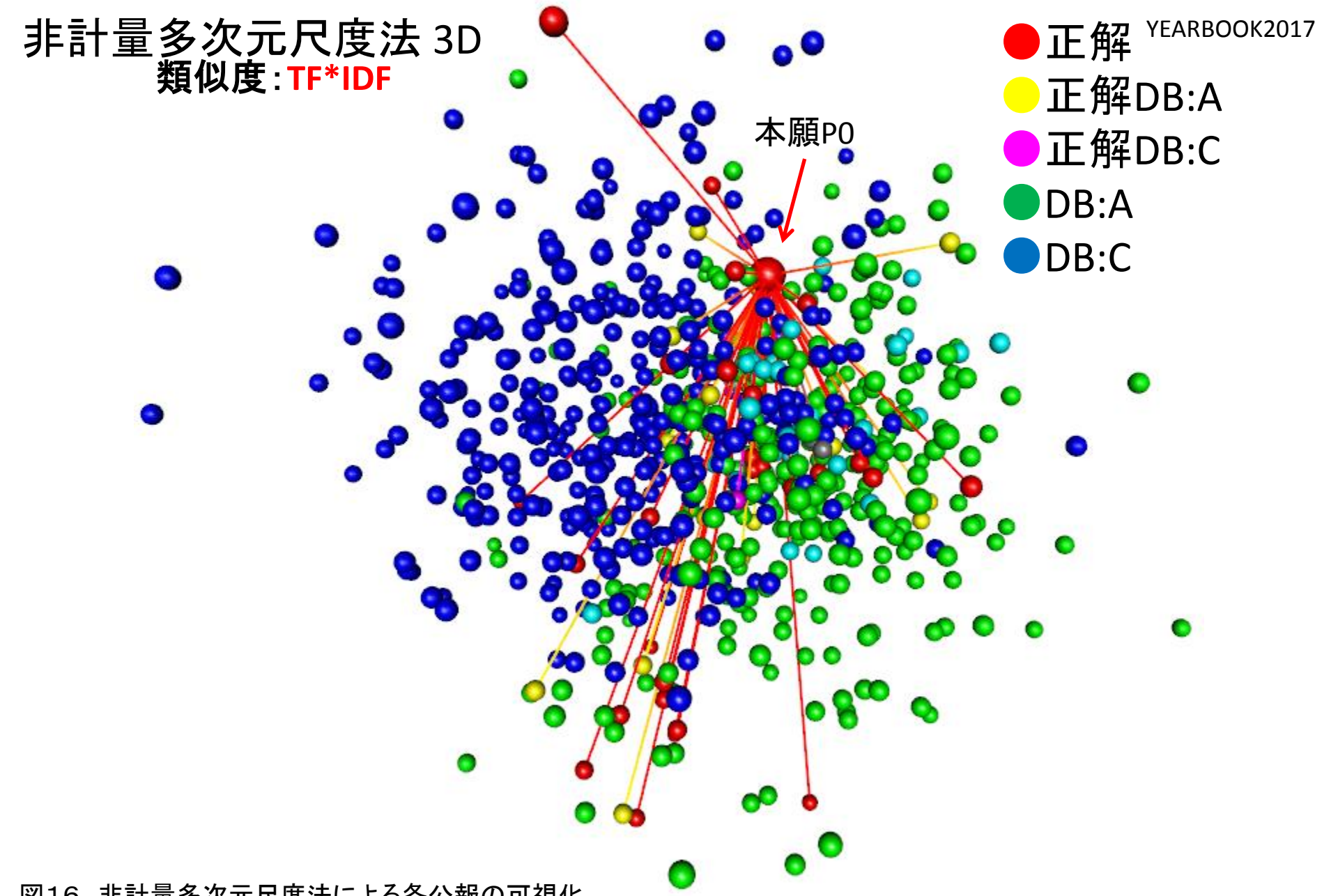


図16. 非計量多次元尺度法による各公報の可視化

doc2vecの類似度による各公報の可視化

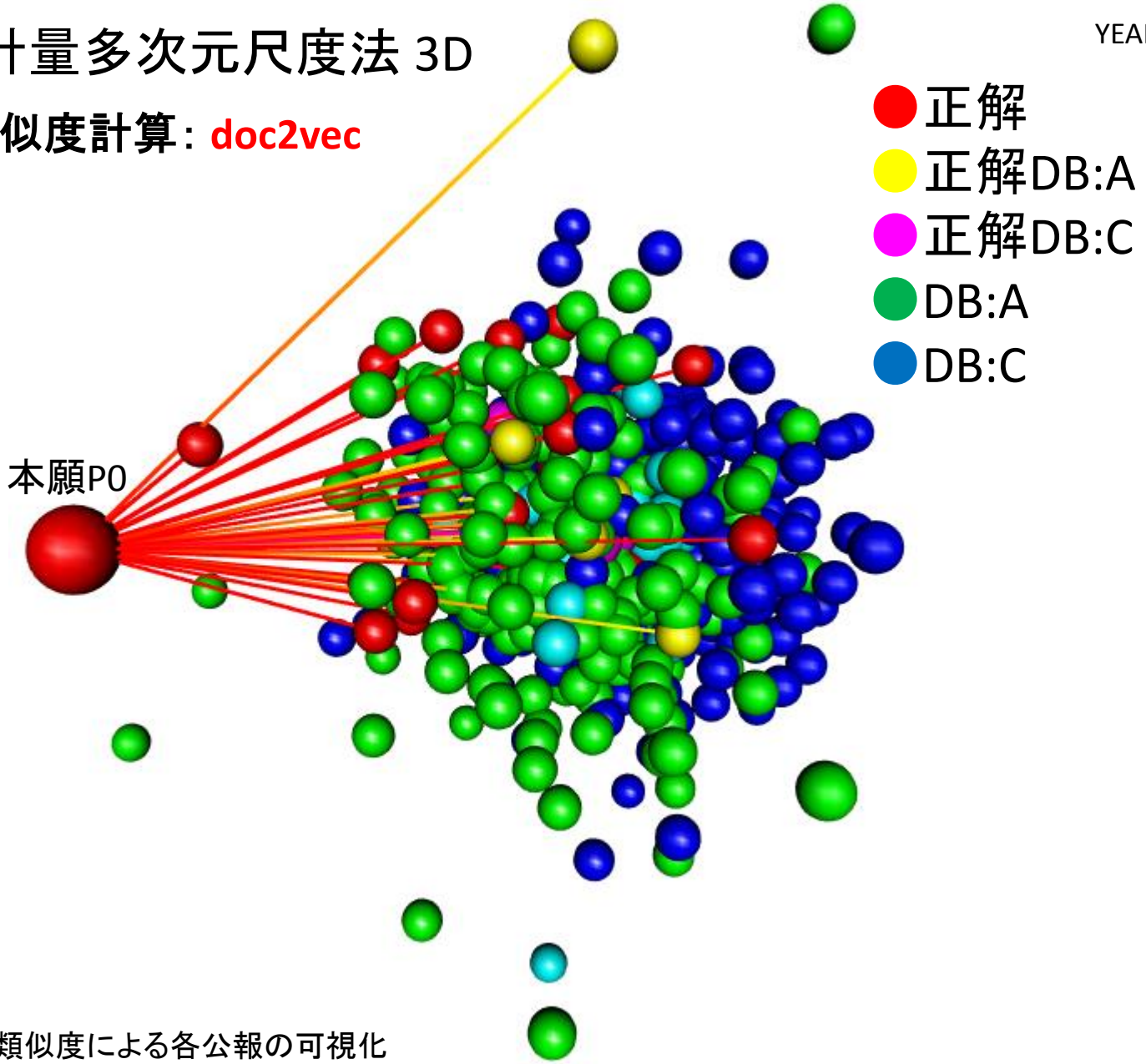


図17．doc2vecの類似度による各公報の可視化

word2vecによる「粘土」の類似語抽出

YEARBOOK2017

word2vec「粘土」の類似語　　形態素　　専門用語抽出

順位	類似語	類似度	順位	頻度	専門用語	順位	頻度
1	スメクタイト	0.774	555	26	スメクタイト	1655	7
4	サポナイト	0.646	2101	4	サポナイト	4655	2
5	ヘクト	0.637	2099	2	ヘクトライト	4656	2
7	スチーブン	0.630	2100	2	スチーブンサイト	4703	2
8	ナイト	0.615	1448	4	カオリナイト	2669	4
9	マイカ	0.614	1449	4	マイカ	3441	3
11	モンモリロナイト	0.599	359	53	モンモリロナイト	246	52
12	カオリ	0.597	1635	3	カオリナイト	2669	4
14	タルク	0.587	1446	4	タルク	2691	4
16	ゼオライト	0.561	1175	7	ゼオライト	1652	7
17	セリ	0.554	2184	4	セリサイト	5112	2

図18．Word2vecによる「粘土」の類似語抽出

専門用語抽出（続き）

専門用語	順位	頻度
水素型スメクタイト	1657	7
合成スメクタイト	1979	6
スメクタイト族	3864	2
スメクタイト群粘土鉱物	4002	2
スメクタイト粘土鉱物	4740	2
合成マイカ	7890	1
カオリン	7203	1

図19．専門用語抽出（続き

主な粘土鉱物(Wikipedia)

カオリナイト（高陵石）
スメクタイト
モンモリロン石（モンモリロナイト）
絹雲母（セリサイト）
イライト
海緑石（グローコナイト）
緑泥石（クロライト）
滑石（タルク）
沸石（ゼオライト）

https://ja.wikipedia.org/wiki/粘土鉱物

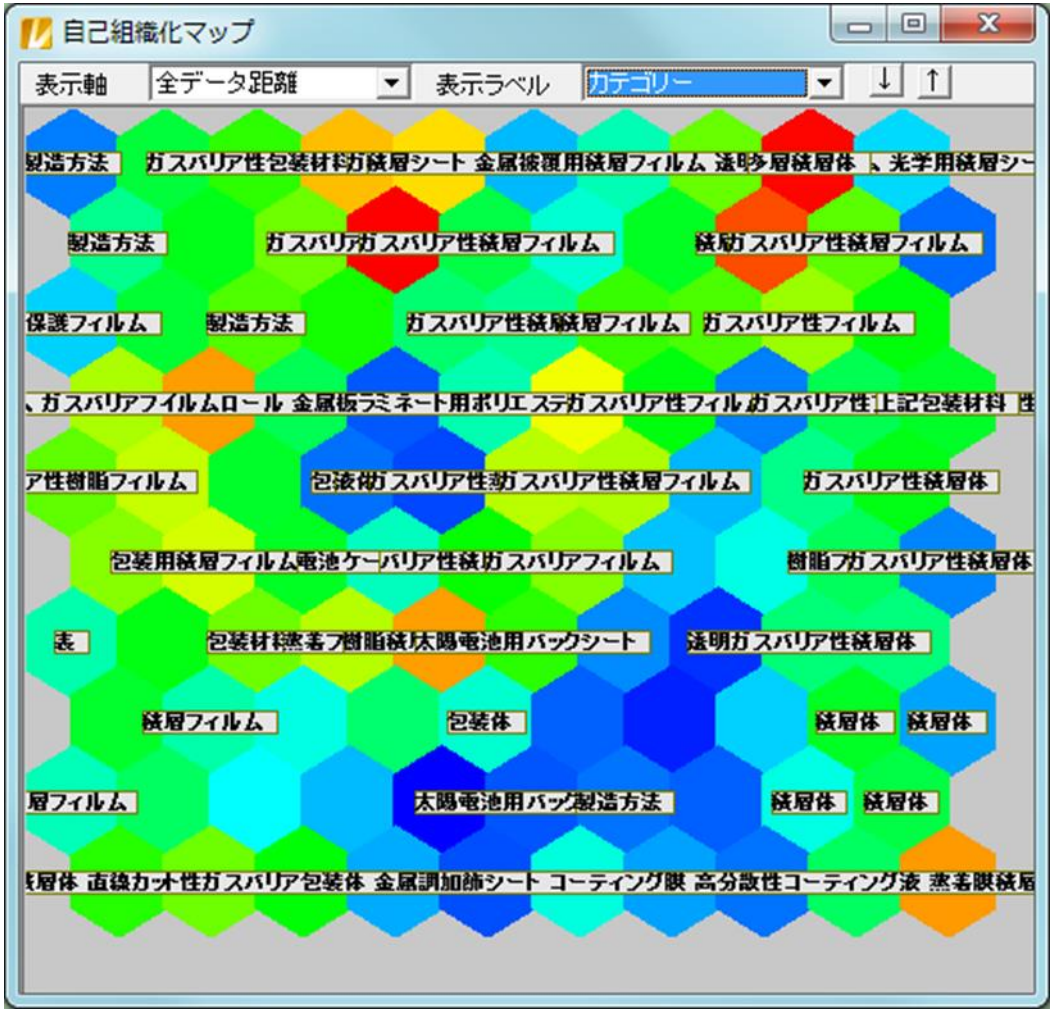
図20．主な粘土鉱物

word2vecを使用すると文脈に「粘土」の記載のない文からも具体的な粘土鉱物を学習しており検索クエリの拡張支援ツールとして有用である

Visual Mining Studio（VMS）の自己組織化マップ

YEARBOOK2017

多次元データの**自己組織化マップ**による可視化



発明のカテゴリーから、自己組織化マップ（SOM）を生成

図21．Visual Mining Studio（VMS）の自己組織化マップ

BayoLinkによるベイジアンネットワーク

YEARBOOK2017

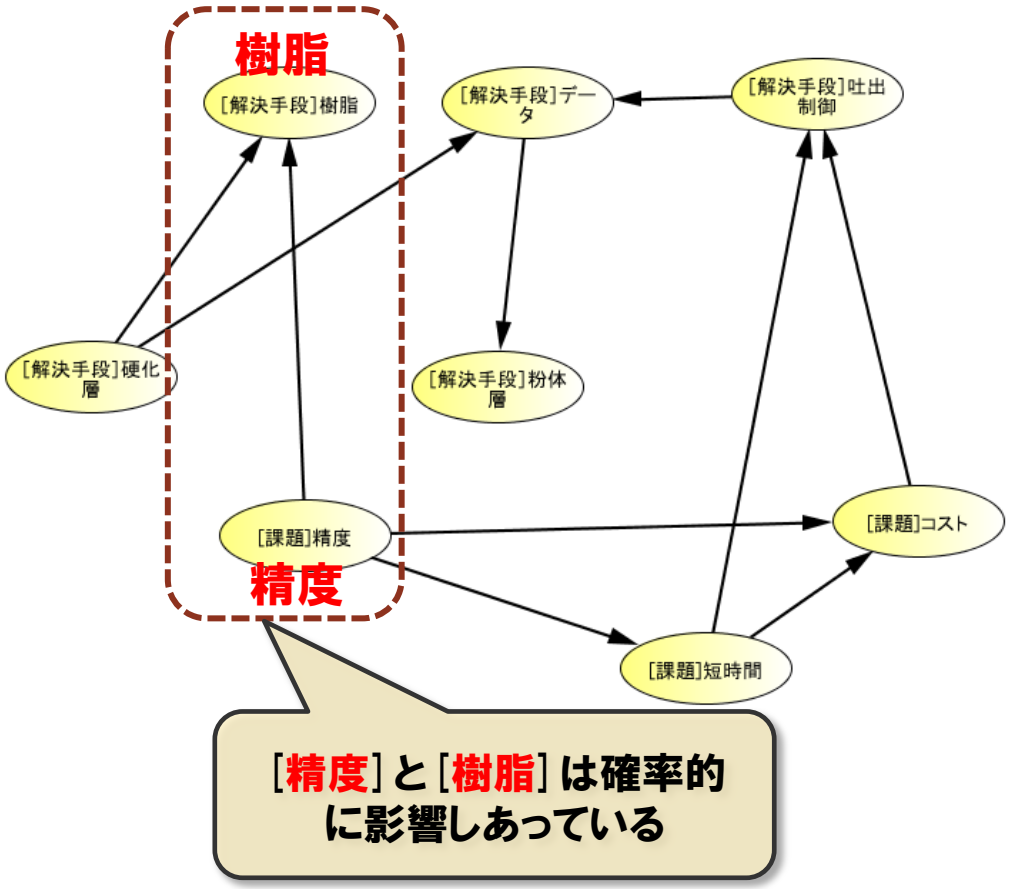


図22．BayoLinkによるベイジアンネットワーク

まとめ

本報では先行技術調査を念頭に特許検索競技大会2016の化学・医薬分野の
問2（ガスバリア性包装用フィルム）を例題として選択しデータセットを作成して
前半ではスクリーニング過程の再現率曲線に影響を与える要因を実験的に検討した。

後半は教師なし機械学習を用いて単語の分散表現で文書の固定長ベクトルが
得られるdoc2vecの学習モデルを使用して公報の類似度を計算する手法を検討した。
その結果単語の出現頻度と出現順序を考慮したモデルPV-DMを使用すると非常によい
類似度計算ができることがわかった。
公報の類似度計算精度が向上すると特許調査において効率的なスクリーニングが
可能となる。可視化や技術動向調査への応用も可能である。

word2vecのような機械学習のフリーライブラリを用いると単語の分散表現学習は非常に
簡単であるが特許調査の精度を上げるには前処理の形態素解析が重要になる。
知財分野では新語の発生頻度も高く形態素解析用辞書の更新や専門用語辞書の
活用も重要である。

謝辞

免責
本報告は2016-2017年の「アジア特許情報研究会」のワーキングの一環として
報告するものである。本報の内容は筆者の私見であり所属機関の見解ではない。

謝辞
最後に大変有用な各種ツールに関し機械学習の初心者である筆者を様々な形で
サポートしていただいたNTTデータ数理システムの多くの皆様に感謝申し上げます。