A31 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

アジア特許情報研究会
〇安藤俊幸　花王株式会社
桐山　勉　はやぶさ国際特許事務所

目次

検討概要

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

One hotベクトルと分散表現ベクトル

One hotベクトル

One hotベクトルとは1種類の単語に1次元を割り当てたベクトルである。
下記は本願P0と文書1の単語頻度を用いたOne hotベクトル（一部抜粋）である。
狭義のOne hotベクトルは単語の有無を0,1で表わす。

	層	樹脂	フィルム	性	他	鉱物	塗	膜	積層	包装	用	特徴	ガスバリア	粘土	基	材	熱	可塑	酸化	・・・
本願P0	4	2	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	・・・
文書1	9	1	7	4	0	0	0	0	3	6	0		0	0	0	0	0	0	0	・・・

後述のデータセット746件　専門用語：8626次元、専門用語+形態素：9974次元

分散表現ベクトル

分散表現ベクトルは固定長の実数で表現されるベクトルである。
下記は本願P0と文書1の200次元の固定長分散表現ベクトルの一部抜粋である。

	v1	v2	v3	v4	v5	v6	・	・	・	v200
本願P0	−4.38.E−03	−2.51.E−03	−1.29.E−02	−5.03.E−03	−1.64.E−02	9.95.E−03				7.27.E−03
文書1	1.36.E−03	1.31.E−03	6.57.E−03	−2.74.E−04	−9.00.E−03	−1.42.E−02				−1.70.E−03

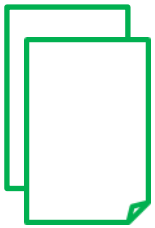
使用データベース／解析ツール

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

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

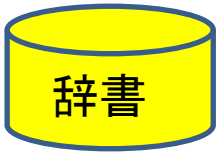


日本語検索
・NRI2
中国語検索
・日本版CNIPR
・Orbit（中国語）

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
・
・

One hotベクトル

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）

INDEX

マイニング

・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

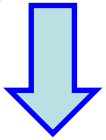
PatAnalyzer（自作ツール）画面

先行技術調査の流れ（進め方）

特許検索競技大会2016
フィードバックセミナー資料p35

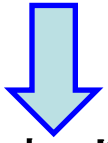
出願したい明細書から構成要素を分析する

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



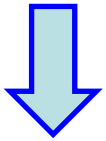
予備検索の実行

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

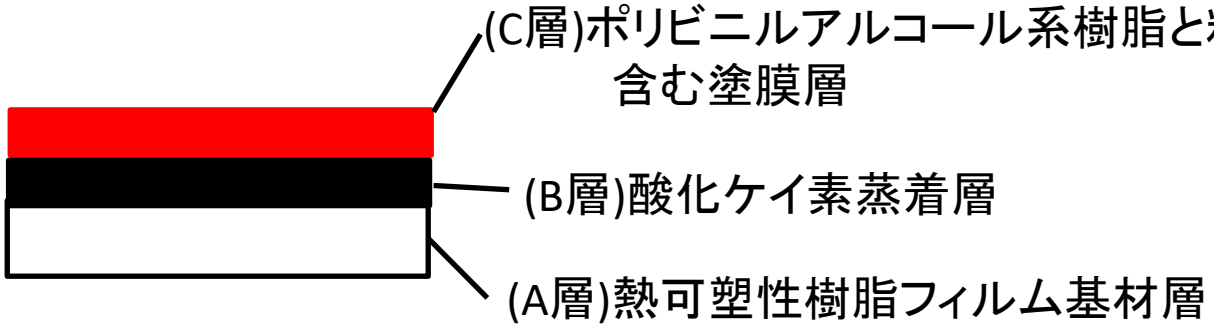
検索実行、**スクリーニング**

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

先行技術調査の事例検討

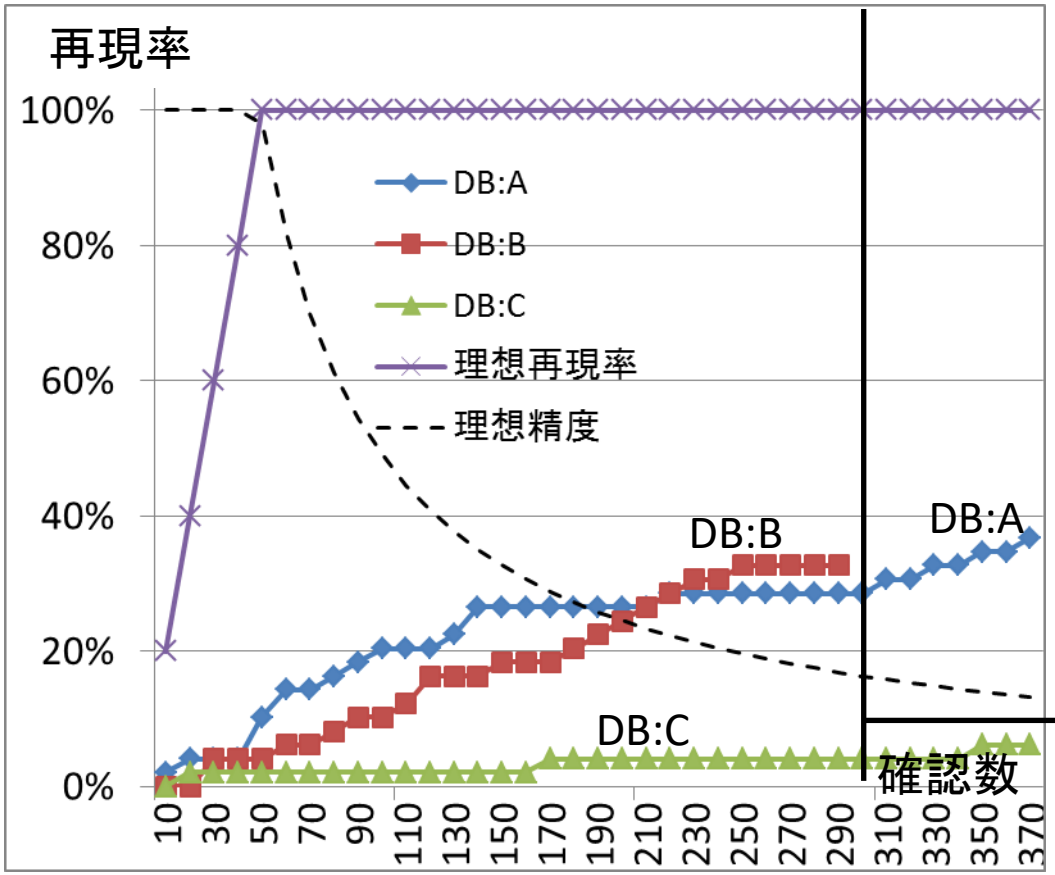
特許検索競技大会2016　化学・医薬分野
出題内容:【問2】問題文概要(2/3)

【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

商用データベースの概念(類似)検索の再現率比較



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

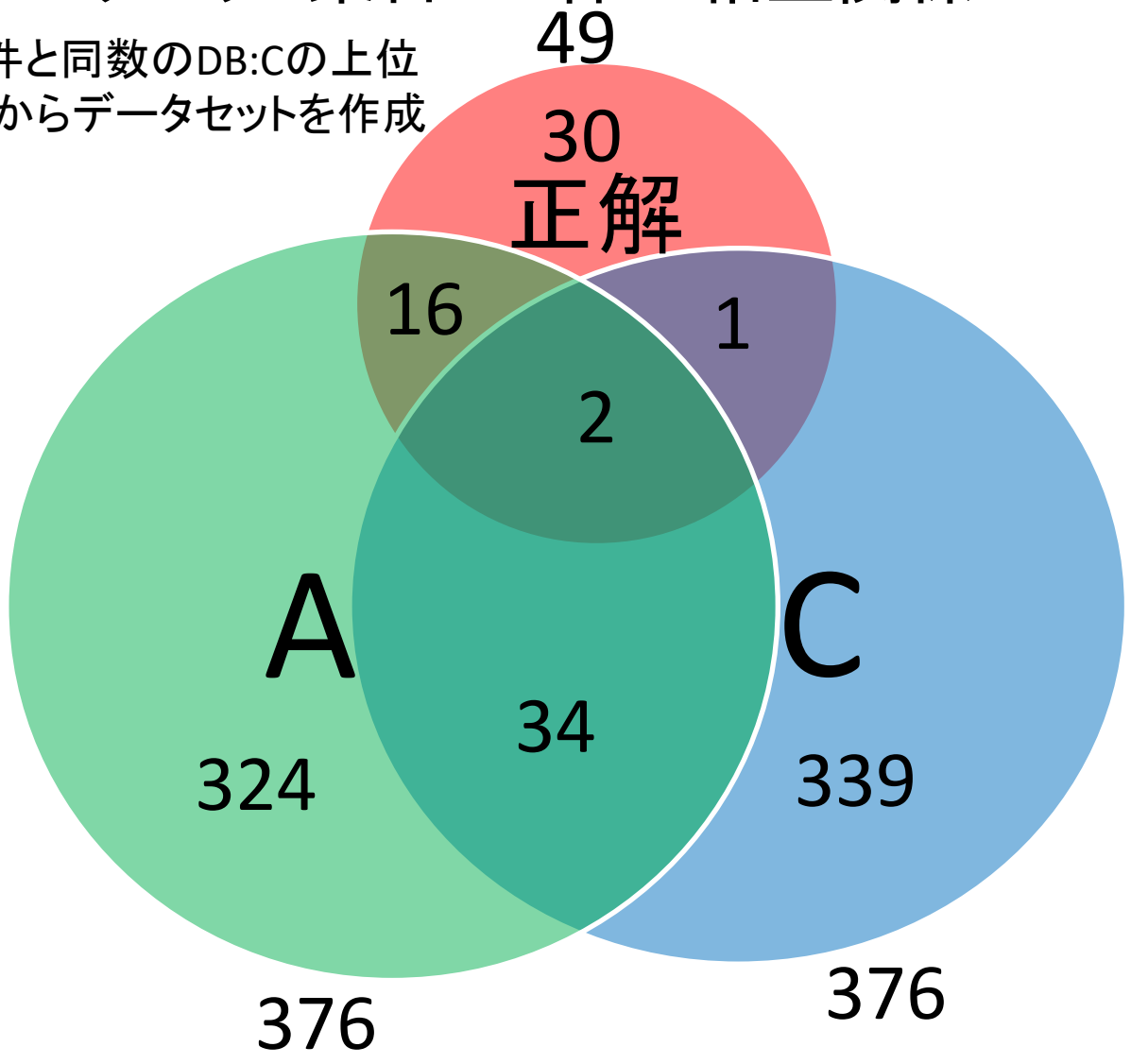
精度	4.7%	5.3%	0.7%
再現率	28.6%	32.7%	4.1%
F値	0.08	0.09	0.01

横軸は公報確認数、縦軸は再現率。破線は理想的な場合の精度。
先行技術調査では確認数少ない(100以下)場合の精度が重要。

実験用データセットの作成

データセット集合746件の相互関係

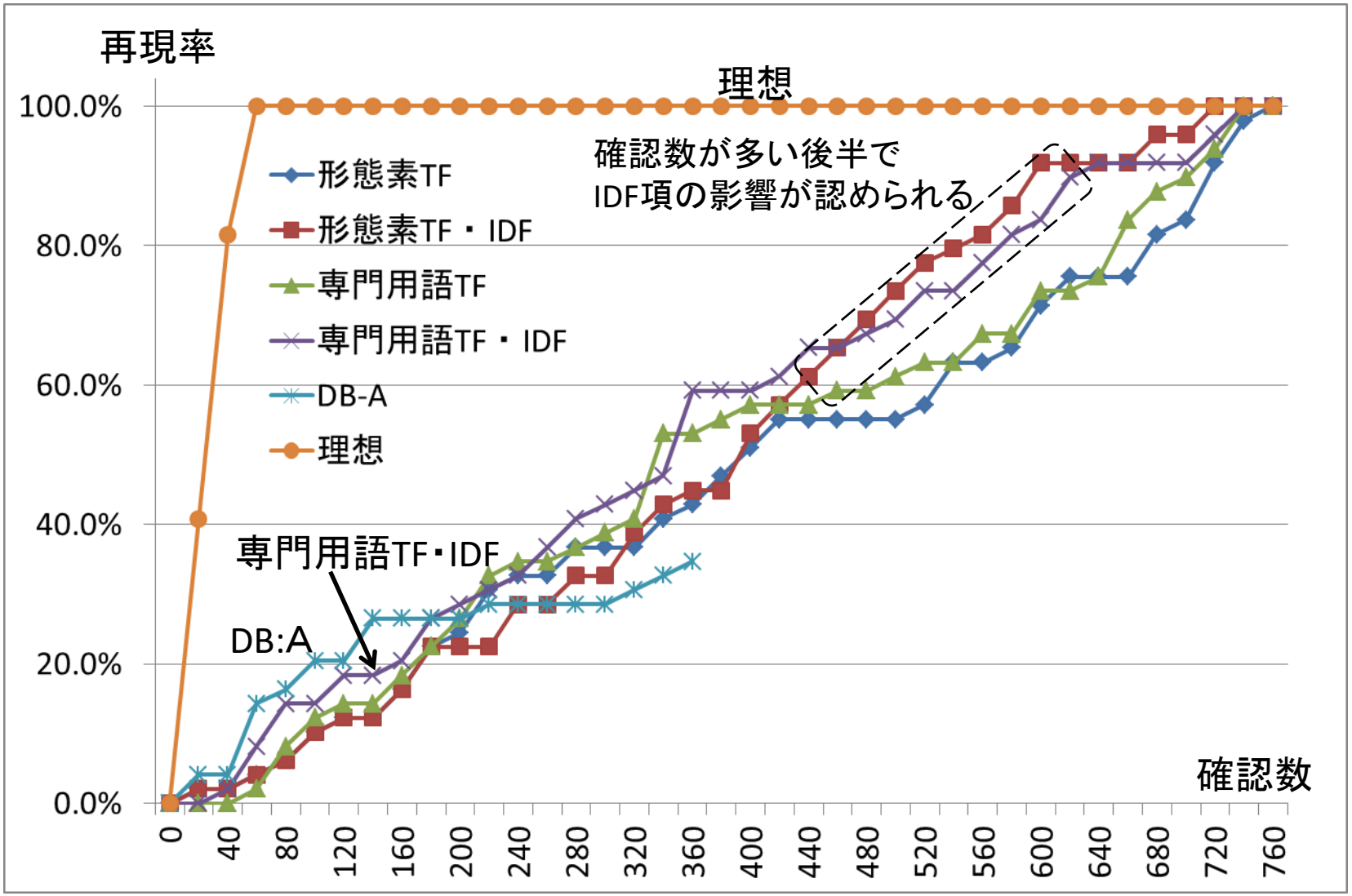
DB:Aの上位376件と同数のDB:Cの上位
公報と正解公報からデータセットを作成



分かち書きと重み付けの再現率への影響

One hotベクトル

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



全体として分かち書き（形態素、専門用語）と重み付け（TF 、TF・IDF）の差は意外に少ない

形態素と専門用語による分かち書き

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）

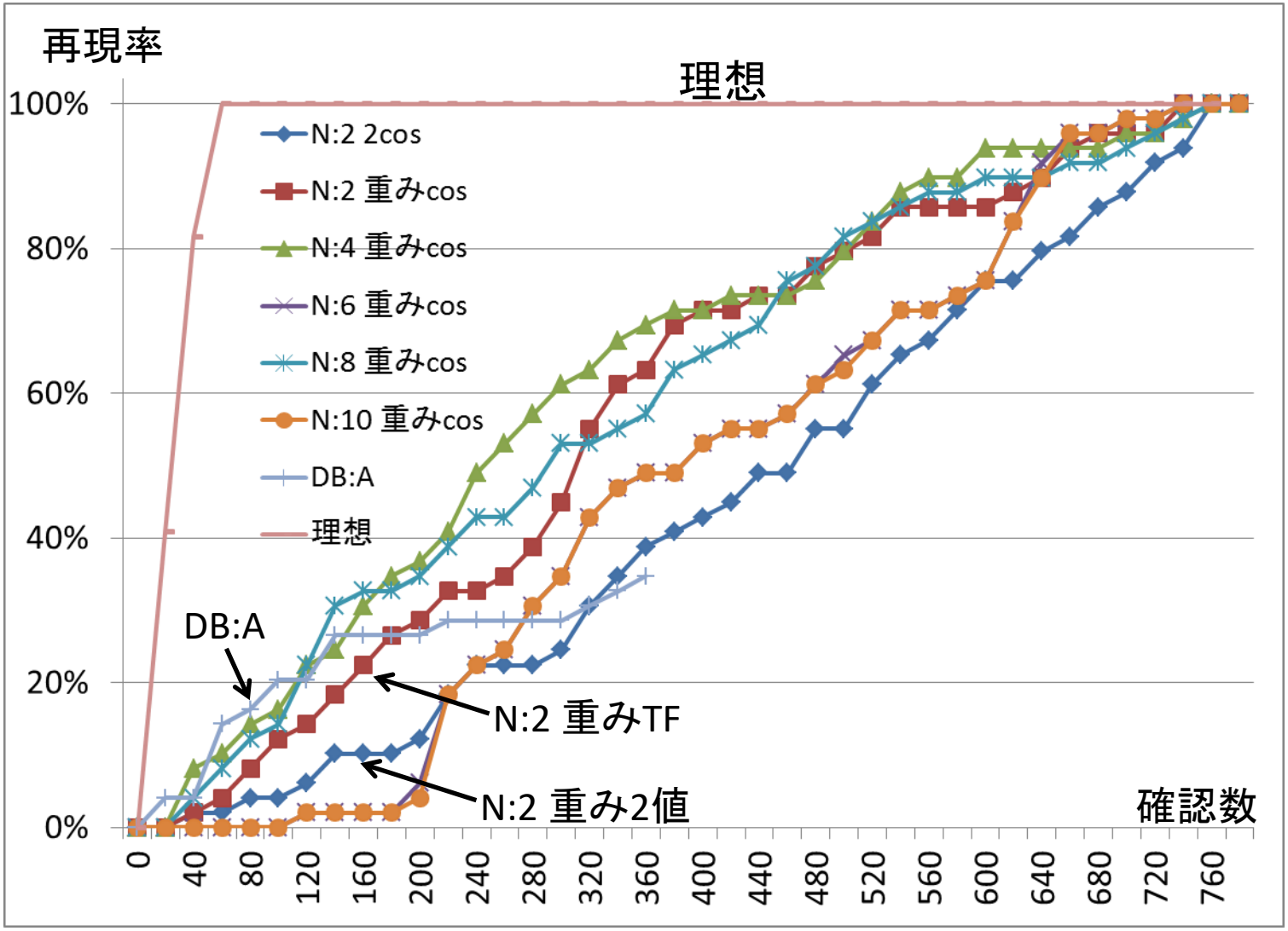
熱可塑性樹脂フィルム基材層
酸化ケイ素蒸着層
ポリビニルアルコール系樹脂
粘土鉱物
塗膜層
他
層
積層
特徴
ガスバリア性包装用フィルム

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

One hotベクトル

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



N:4付近が良さそう、重みとしてTFと2値ではTFの方が良い

構成要素分析（検索競技大会の模範解答例）

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:【間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。

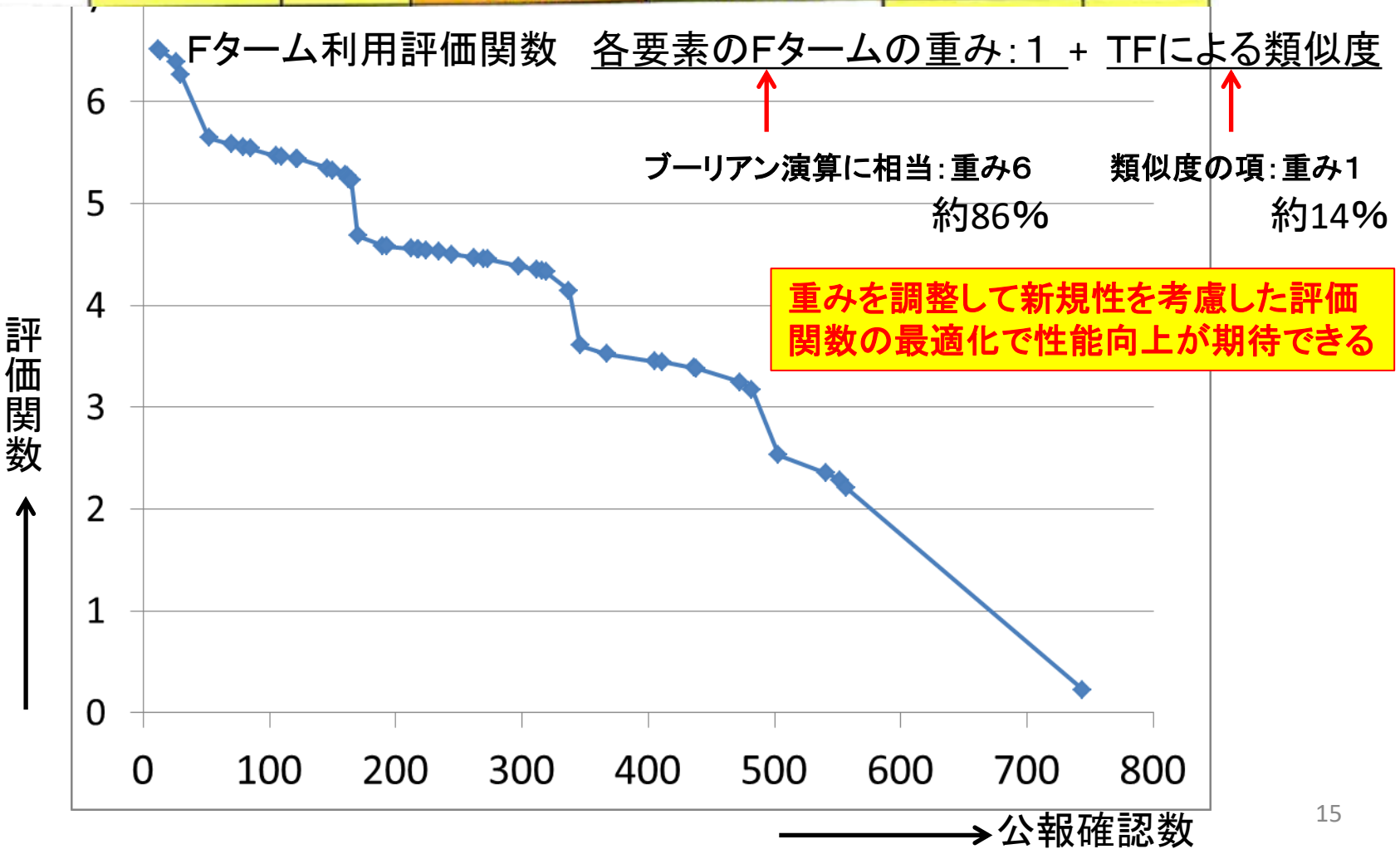
記号	構成要素(概念)
a	熱可塑性樹脂フィルム基材層
b	酸化ケイ素蒸着層
c	ポリビニルアルコール系樹脂を含む塗膜層
d	塗膜層に粘土鉱物を含む
e	他の層を介してまたは介さずにこの順に積層
f	ガスバリア性
g	包装用フイルム

※構成要素の分け方は本例に限定しない

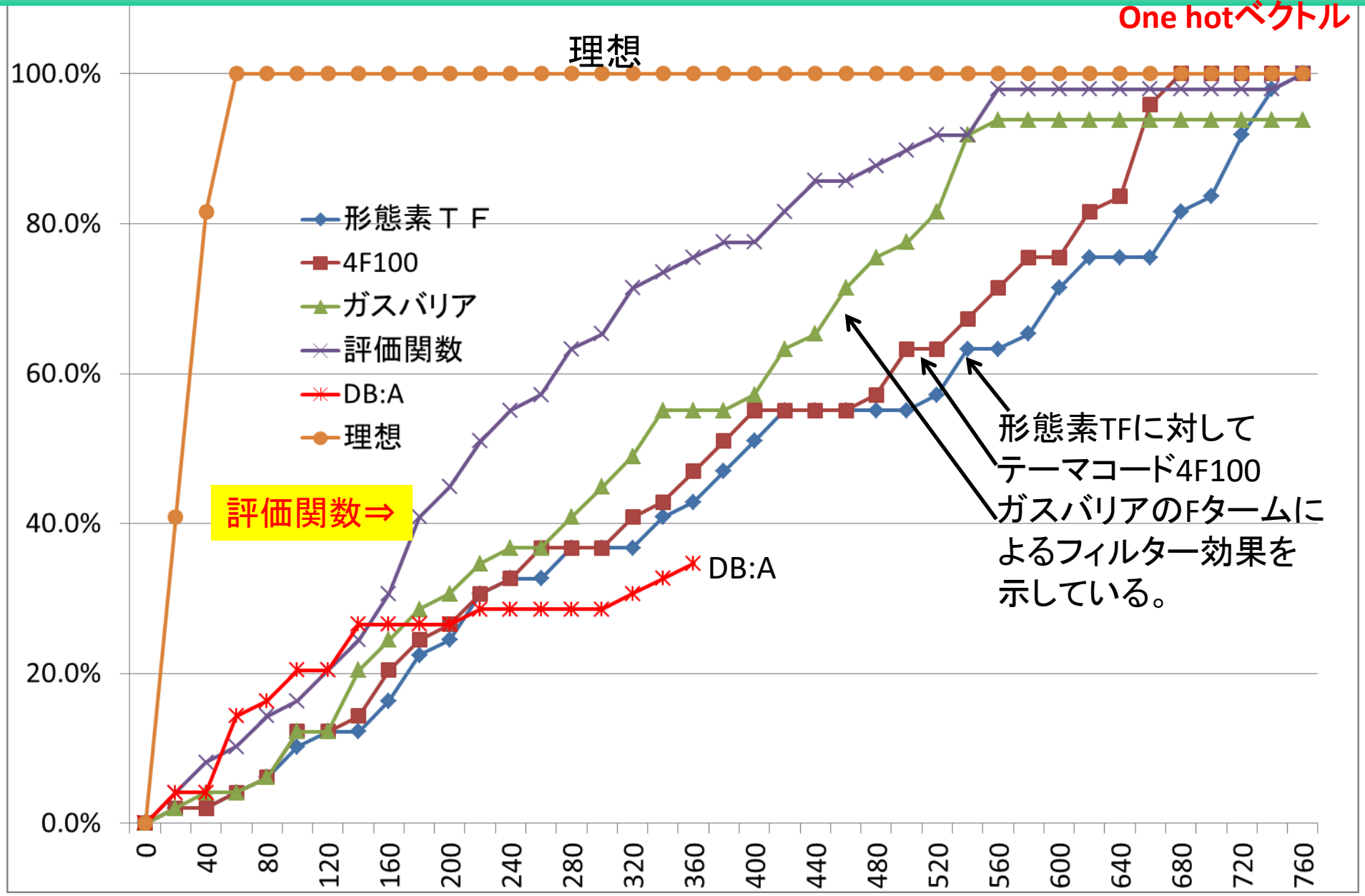
Fタームと形態素TF類似度による評価関数

One hotベクトル

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



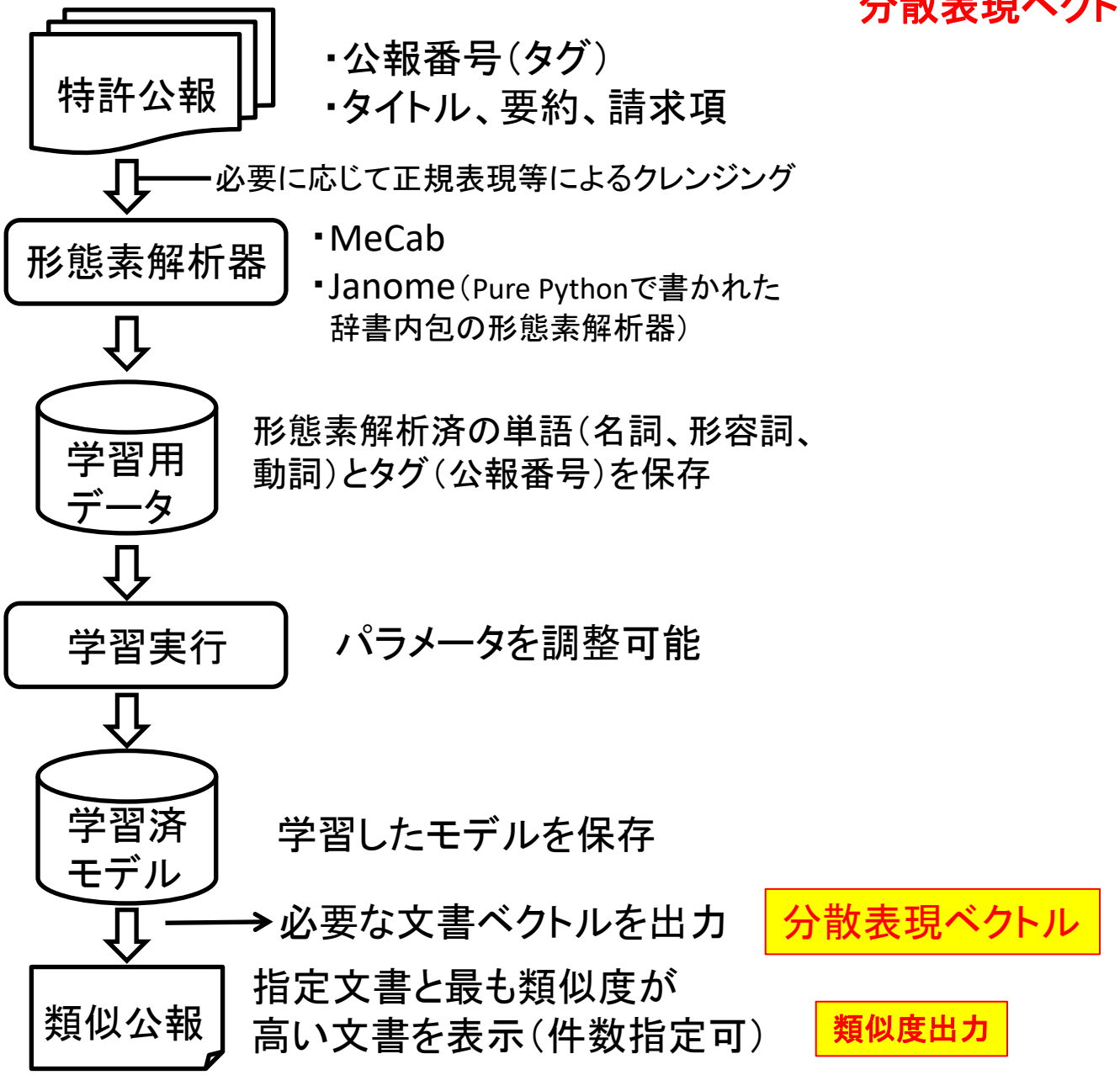
評価関数とフィルターの影響



評価関数が良い結果を示している。

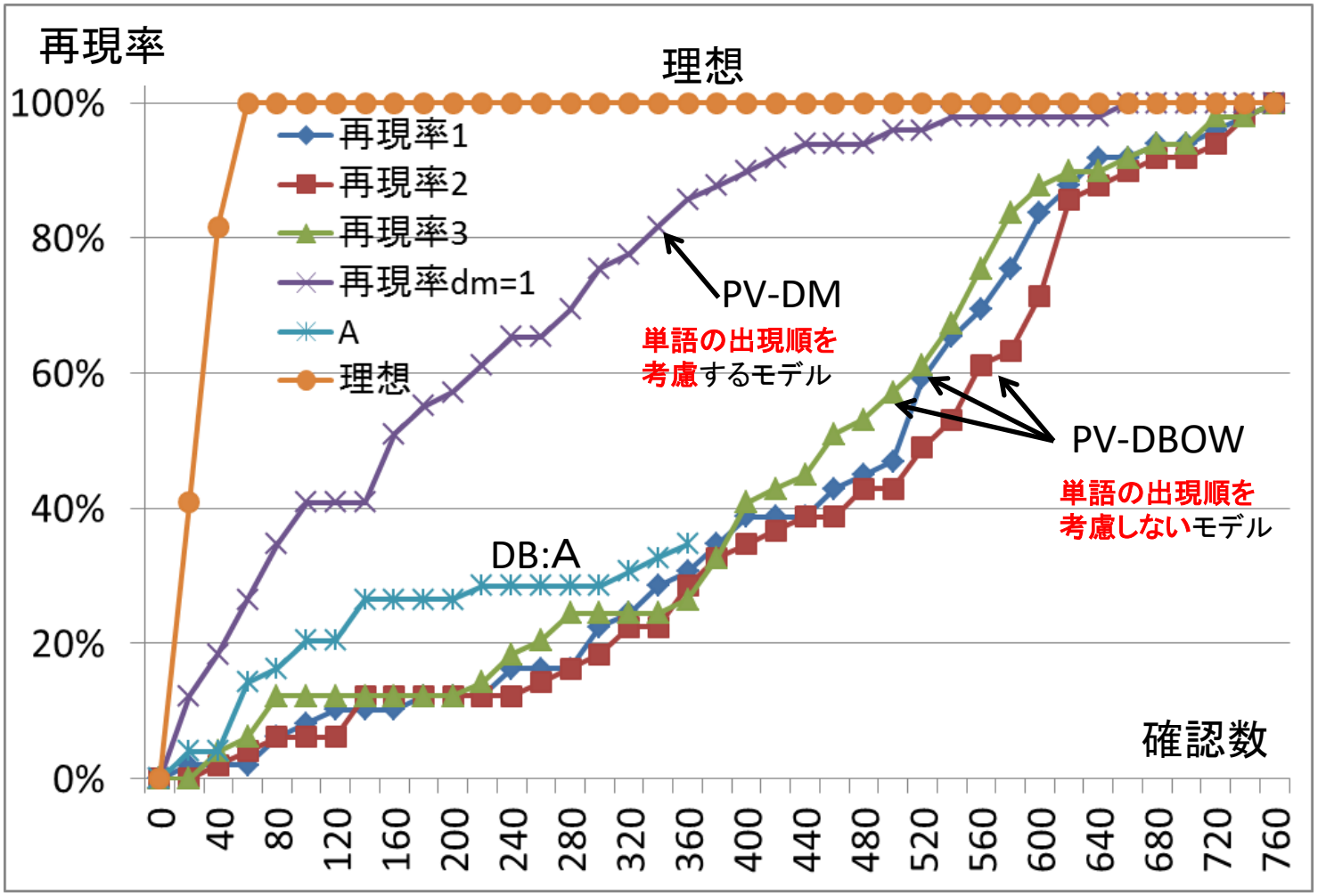
doc2vecによる文書のベクトル化処理の概要

分散表現ベクトル



文書の分散表現ベクトルの学習モデルと再現率

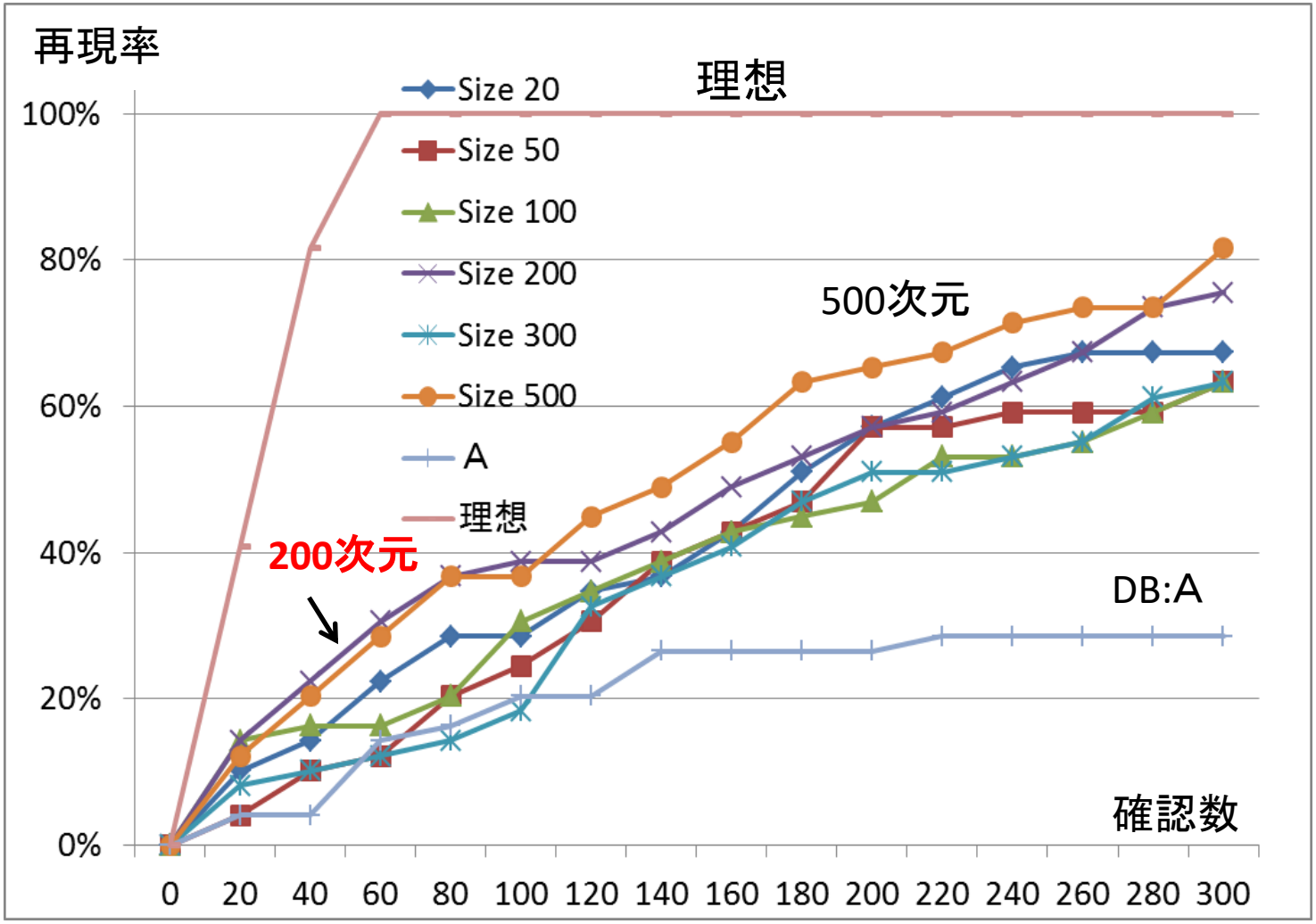
分散表現ベクトル



単語の出現順を考慮するモデルPV-DMが良い結果を示している

文書の分散表現ベクトルの次元数（Size）の影響

分散表現ベクトル



確認数が少ない前半では200次元が良い結果を示している

word2vecによる「粘土」の類似語抽出

分散表現ベクトル

word2vec「粘土」の類似語

順位	類似語	類似度
1	スメクタイト	0.774
4	サポナイト	0.646
5	ヘクト	0.637
7	スチーブン	0.630
8	ナイト	0.615
9	マイカ	0.614
11	モンモリロナイト	0.599
12	カオリ	0.597
14	タルク	0.587
16	ゼオライト	0.561
17	セリ	0.554

黄色セルは形態素解析による分かち書きに失敗しているが類似語として上位に存在している

One hotベクトル

形態素

順位	頻度
555	26
2101	4
2099	2
2100	2
1448	4
1449	4
359	53
1635	3
1446	4
1175	7
2184	4

専門用語抽出

専門用語	順位	頻度
スメクタイト	1655	7
サポナイト	4655	2
ヘクトライト	4656	2
スチーブンサイト	4703	2
カオリナイト	2669	4
マイカ	3441	3
モンモリロナイト	246	52
カオリナイト	2669	4
タルク	2691	4
ゼオライト	1652	7
セリサイト	5112	2

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

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）

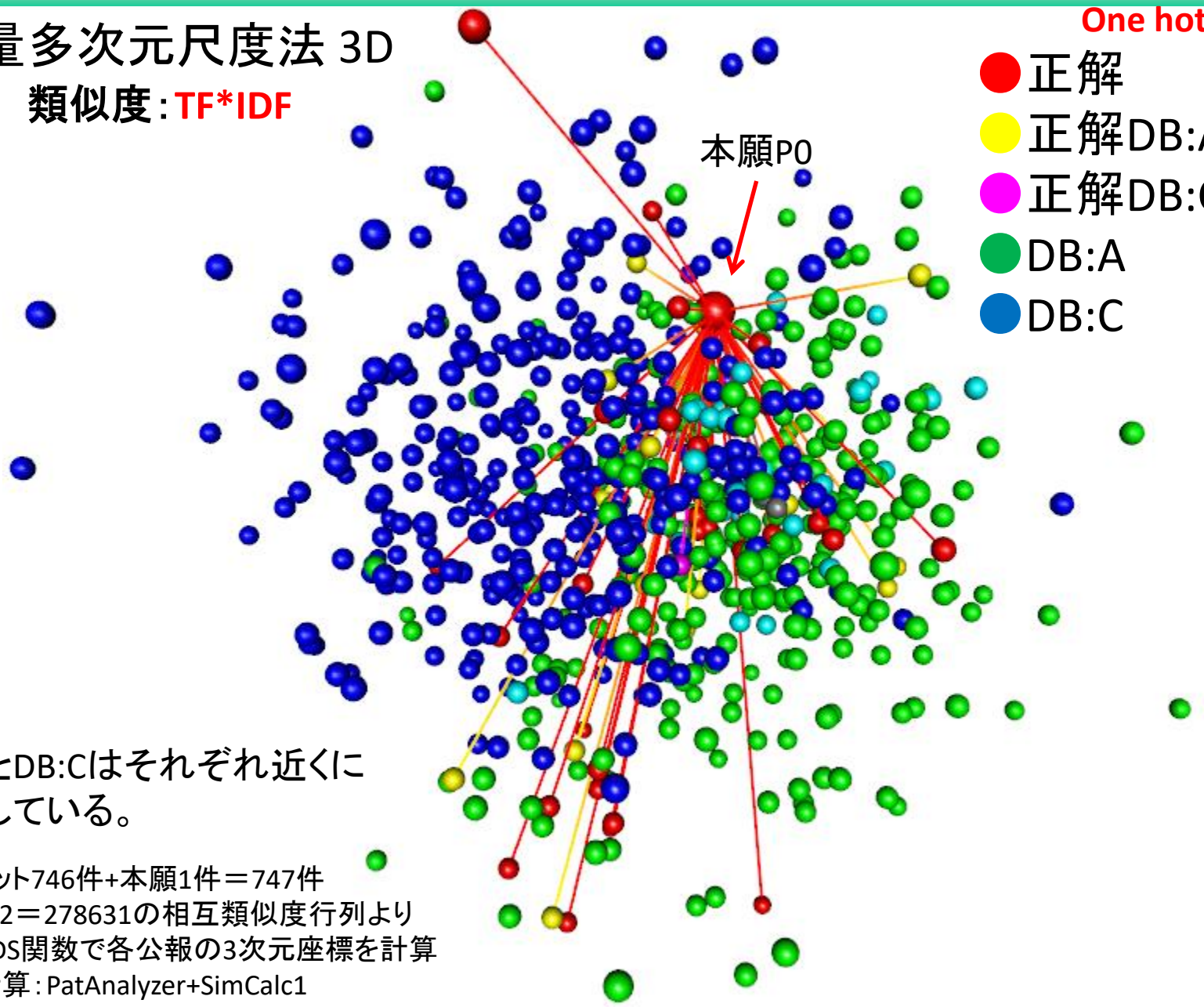
https://ja.wikipedia.org/wiki/粘土鉱物

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

非計量多次元尺度法による各公報の可視化

非計量多次元尺度法 3D

類似度：TF*IDF



DB:AとDB:Cはそれぞれ近くに
分布している。

データセット746件+本願1件＝747件
747*746/2＝278631の相互類似度行列より
RのisoMDS関数で各公報の3次元座標を計算
類似度計算：PatAnalyzer+SimCalc1

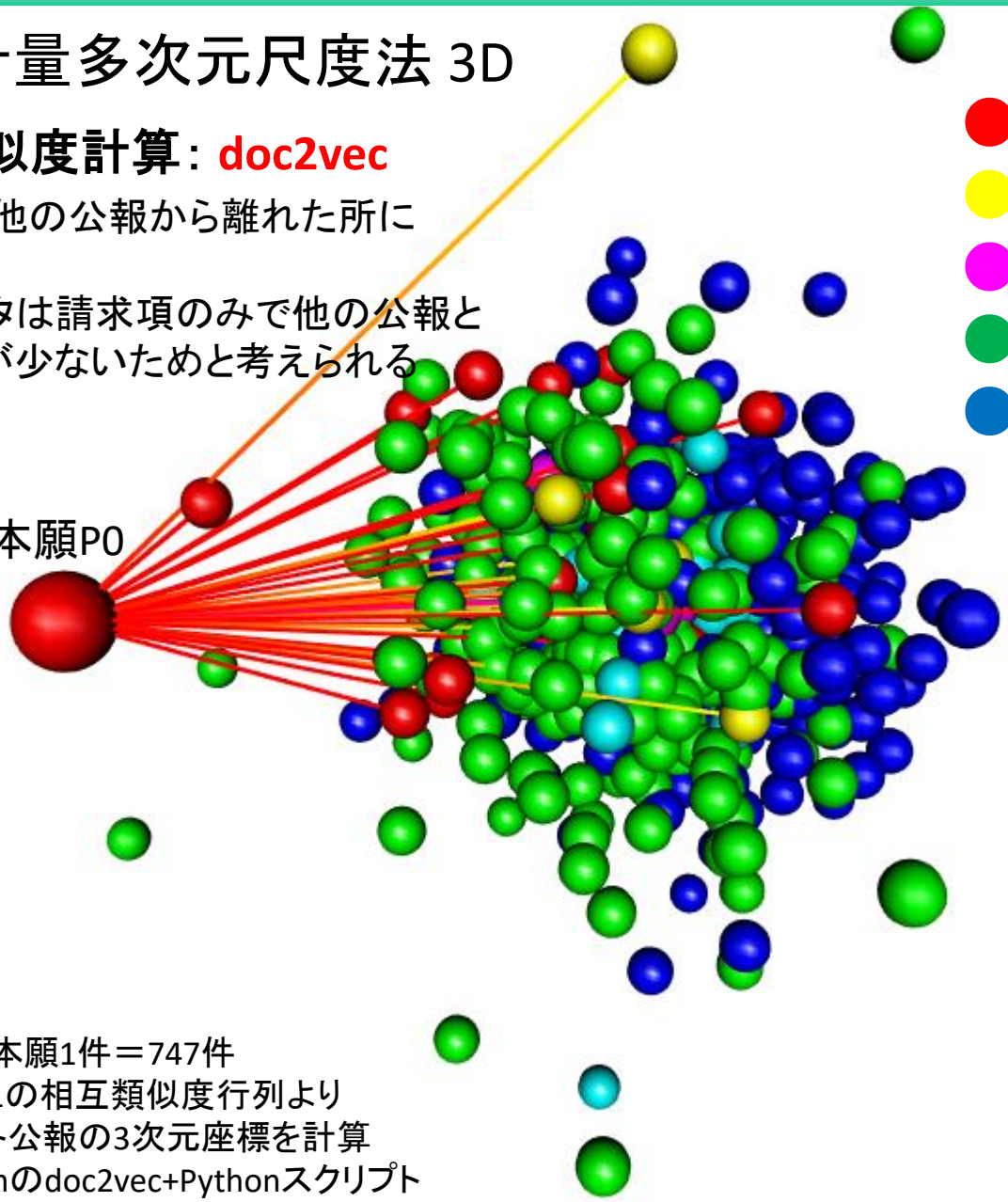
doc2vecの類似度による各公報の可視化

非計量多次元尺度法 3D

分散表現ベクトル

類似度計算：doc2vec

本願P0公報のみ他の公報から離れた所に
位置している。
本願の入力データは請求項のみで他の公報と
比較して情報量が少ないためと考えられる



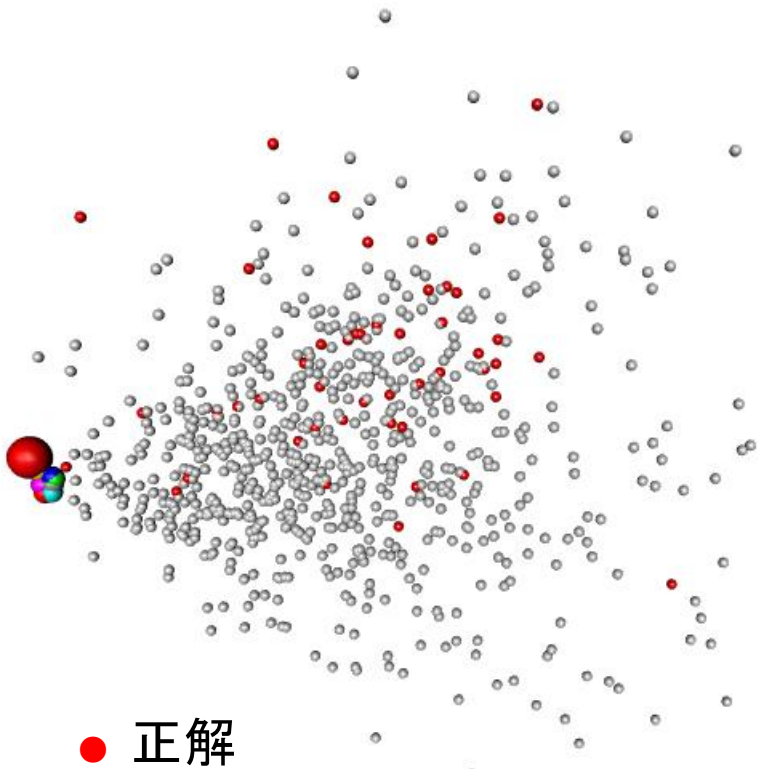
データセット746件+本願1件＝747件
747*746/2＝278631の相互類似度行列より
RのisoMDS関数で各公報の3次元座標を計算
類似度計算：Gensimのdoc2vec+Pythonスクリプト

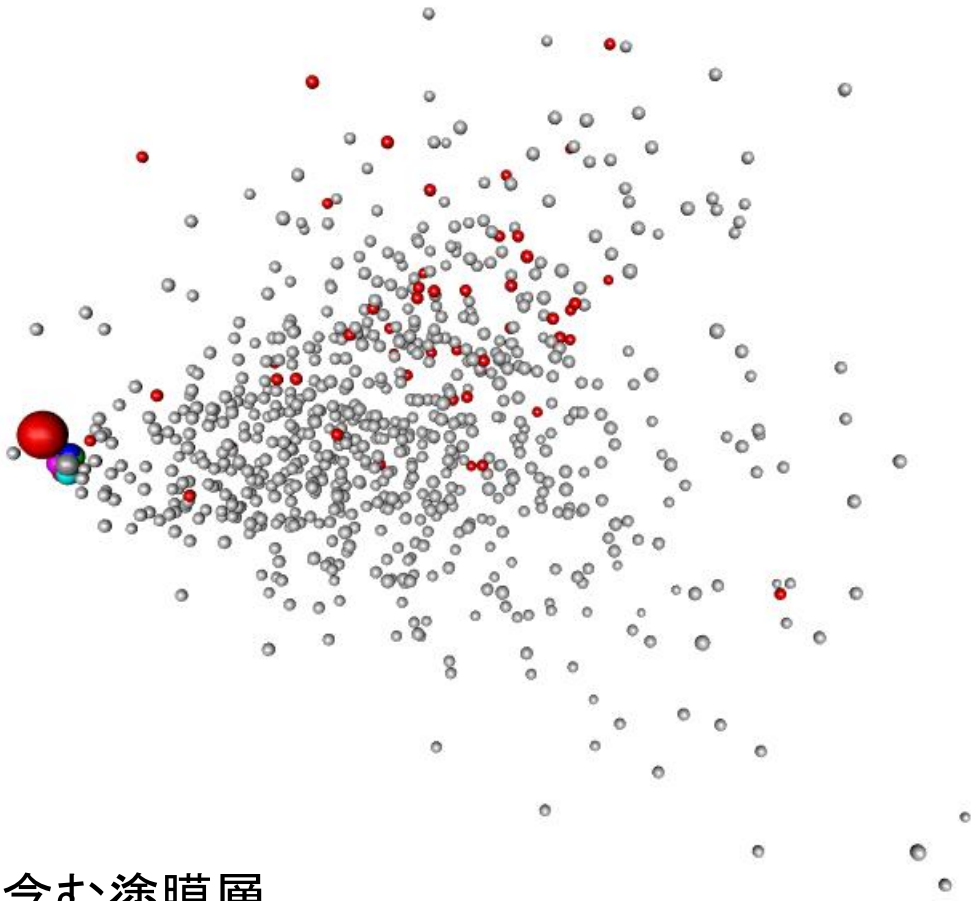
doc2vecによる分散表現ベクトルの次元圧縮検討

次元圧縮：PCA：主成分分析 2D

次元圧縮：PCA 3D

分散表現ベクトル





- 正解
- 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

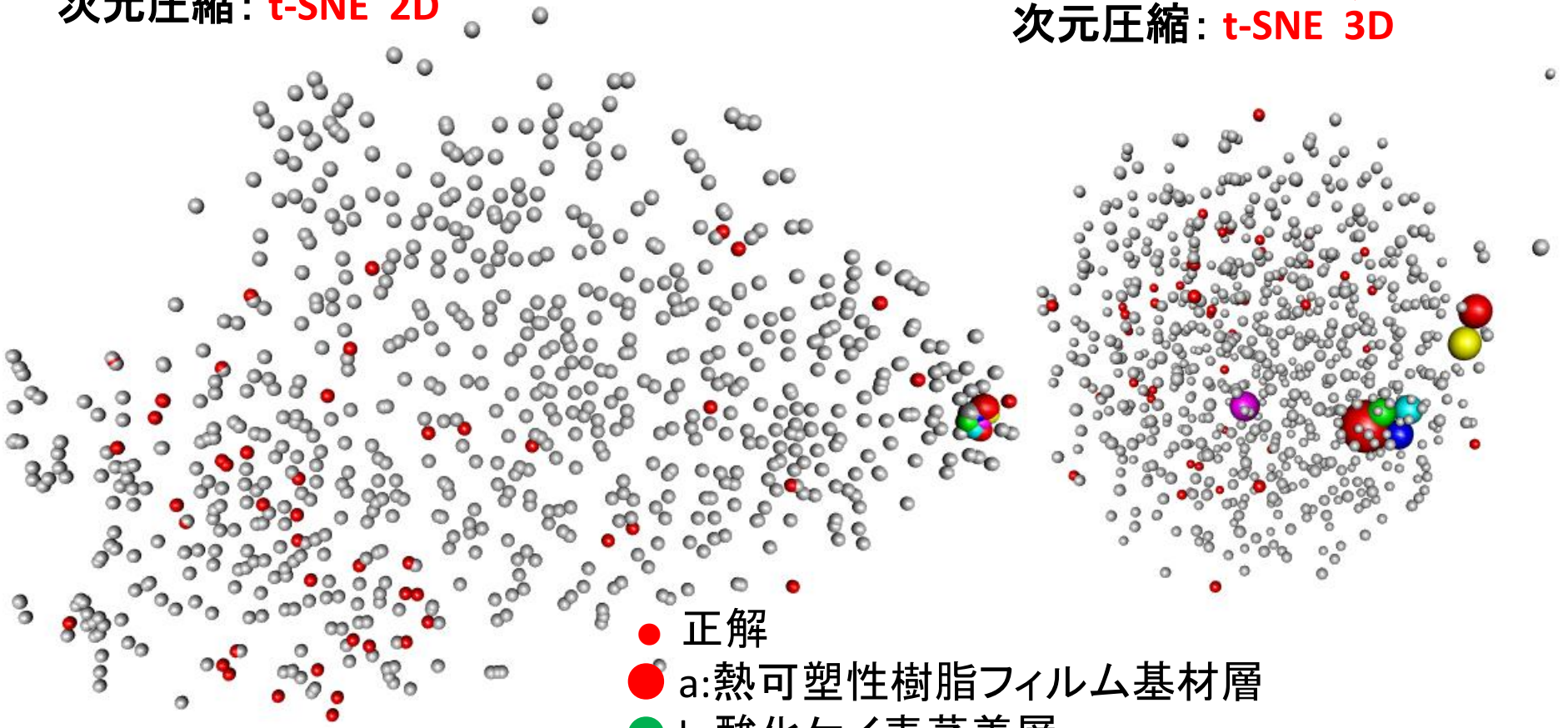
データセット746件+本願1件＝747件
747件×200次元の行列より
scikit-learnのPCAで次元圧縮
分散表現学習:Gensimのdoc2vec

doc2vecによる分散表現ベクトルの次元圧縮検討

分散表現ベクトル

次元圧縮：t-SNE 2D

次元圧縮：t-SNE 3D



- 正解
- 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

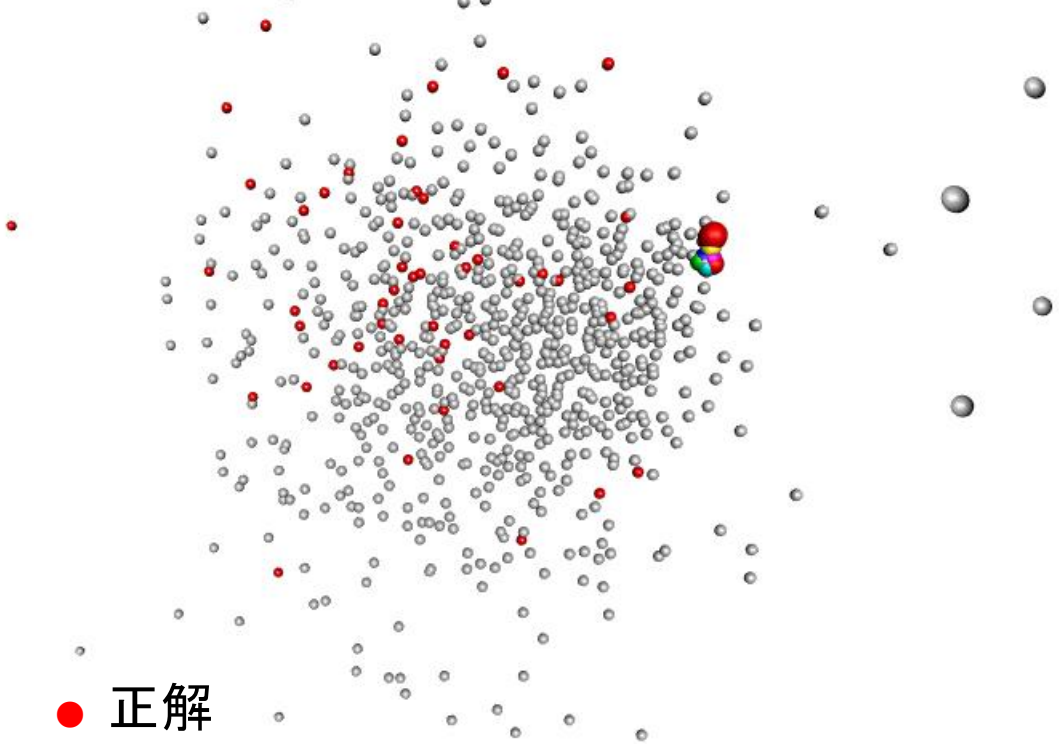
データセット746件+本願1件＝747件
747件×200次元の行列より
scikit-learnのt-SNEで次元圧縮
分散表現学習:Gensimのdoc2vec

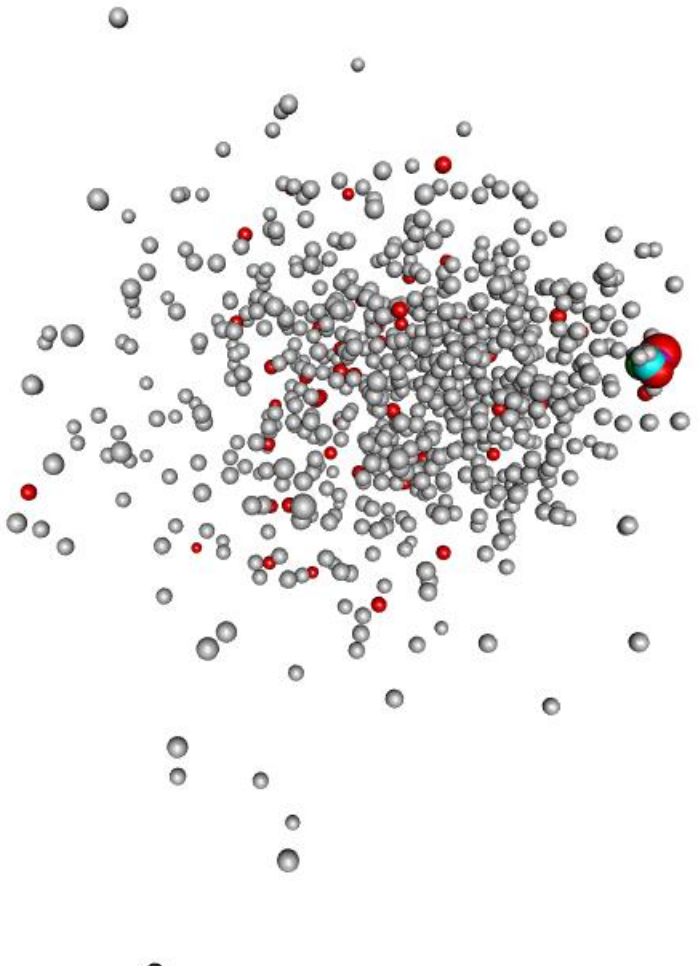
doc2vecによる分散表現ベクトルの次元圧縮表示

分散表現ベクトル

次元圧縮: MDS:多次元尺度法 2D

次元圧縮: MDS 3D





- 正解
- 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

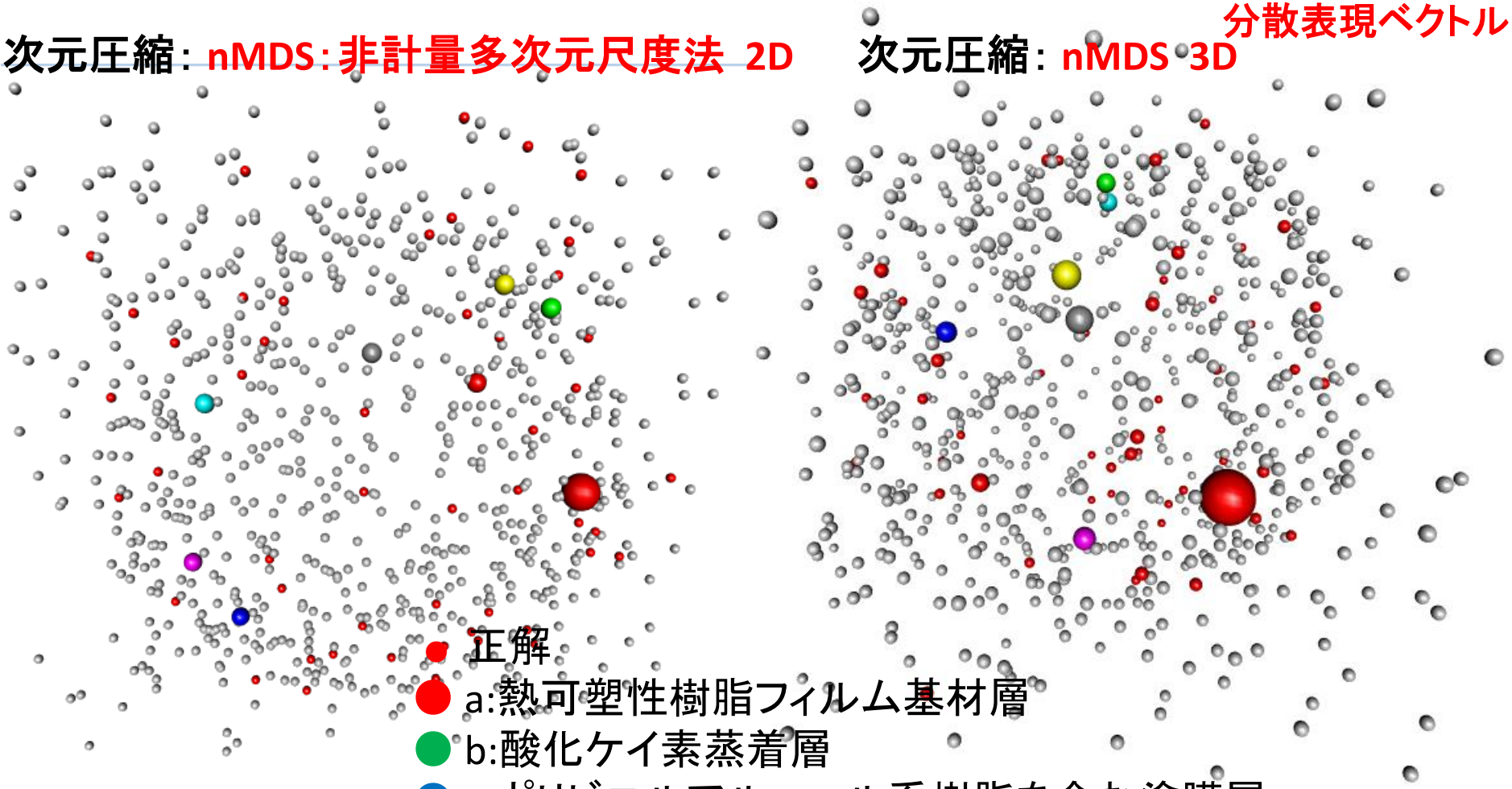
データセット746件+本願1件＝747件
747件×200次元の行列より
scikit-learnのMDSで次元圧縮
分散表現学習:Gensimのdoc2vec

doc2vecによる分散表現ベクトルの次元圧縮検討

分散表現ベクトル

次元圧縮：nMDS：非計量多次元尺度法 2D

次元圧縮：nMDS 3D



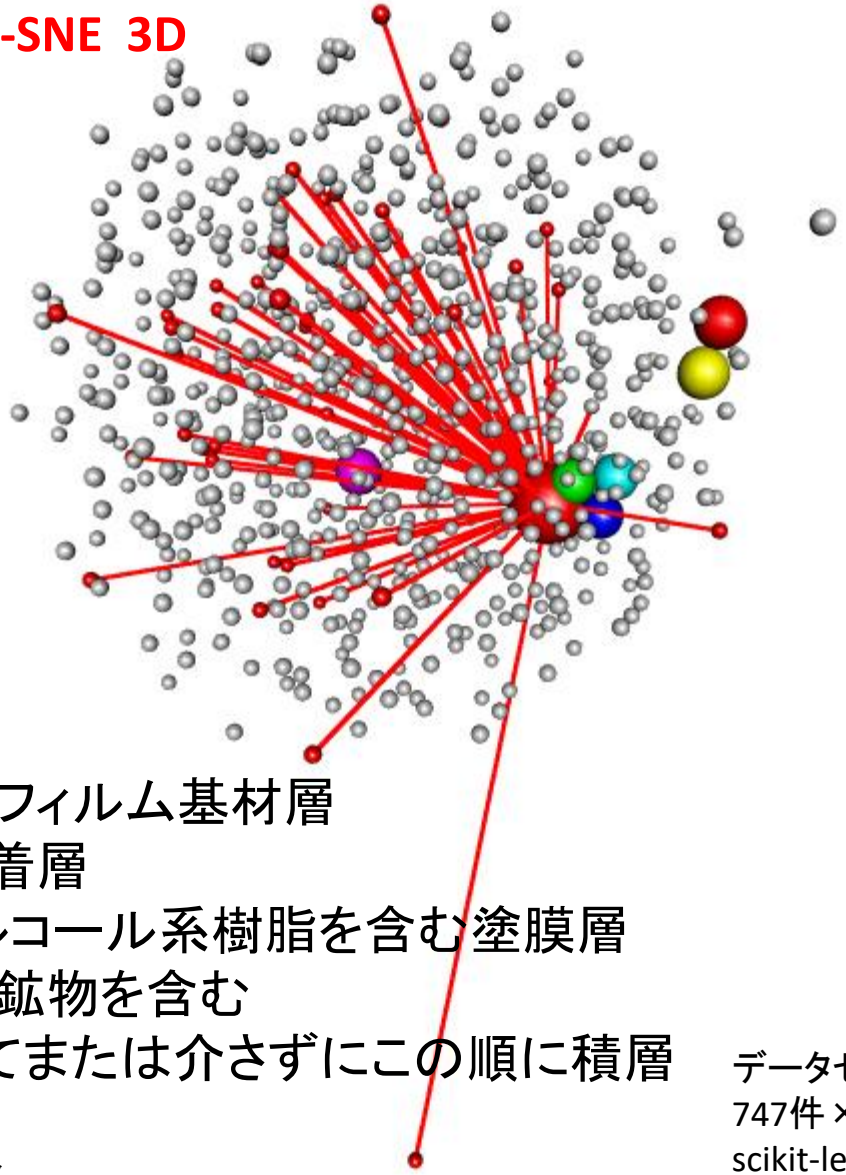
- 正解
- a:熱可塑性樹脂フィルム基材層
- b:酸化ケイ素蒸着層
- c:ポリビニルアルコール系樹脂を含む塗膜層
- d:塗膜層に粘土鉱物を含む
- e:他の層を介してまたは介さずにこの順に積層
- f:ガスバリア性
- g:包装用フイルム

747件×200次元の行列より
scikit-learnのnMDSで次元圧縮
分散表現学習:Gensimのdoc2vec

doc2vecによる分散表現ベクトルの次元圧縮表示

分散表現ベクトル

次元圧縮: t-SNE 3D



- 正解
- 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

データセット746件+本願1件＝747件
747件×200次元の行列より
scikit-learnのt-SNEで次元圧縮
分散表現学習:Gensimのdoc2vec

まとめ

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）
　②重み付けの影響（ TF/ TF-IDF）
　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/フィルター）
①②の影響は意外に小さく③の評価関数の性能が良く、更に最適化で向上が期待できる

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
　　単語の出現順を考慮したPV-DMモデルが良い結果を示した
　②Word2Vecによる単語の分散表現学習⇒文脈に合った類似語が学習できている

3．可視化検討
　①次元圧縮（PCA/t-SNE/MDS/nMDS）
　　t-SNEが良い結果を示した

結論

下記①②の方法を組み合わせて特許調査の精度と効率を向上可能
①母集団のサンプルサイズ減縮（余分なノイズ公報を除去） ←フィルタリング
②意味のある情報と無駄な情報を識別して意味のある情報を使用しての次元圧縮が有効
③公報の分散表現ベクトルの2ないし3次元への圧縮は動向調査への様々な応用が可能

今後の展望

本報で検討した分散表現ベクトルを更に教師データ有りの機械学習の入力データとすることも可能である。更なる精度、再現率向上には教師データ有りの機械学習と組み合わせることが必須と考える。教師データ有りの機械学習としては評価関数を用いて構成要素によって重みを変える、Fタームと形態素の類似度の寄与率を変える等々いろいろ考えられる。重み付けの調整や識別を利用することで改善の余地は大きいと考える。評価関数をどこまでチューニングできるか興味深い。
特許調査の精度を上げるには前処理の形態素解析による「分かち書き」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

「謝辞」

本報告は2017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す。研究会のメンバーの皆様には様々な協力をしていただきました。ここに改めて感謝申し上げます。

ご清聴、ありがとうございました。

参考資料

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

word2vecとdoc2vec

word2vec

word2vecは、2層から成り、テキスト処理を行うニューラルネットワークである。
分かち書きしたテキストコーパスを入力すると、単語の特徴量ベクトル（feature vector）が出力される。
word2vecは、ディープ・ニューラル・ネットワークではない。
テキストをディープニューラルネットワークが処理できる数値形式に変える。
word2vecの有用性は、類似語のベクトルをベクトル空間に配置することである。
word2Vecでは各単語を200次元くらいの空間内におけるベクトルとして表現する。
それぞれの単語を200個の実数の組み合わせとして表現するため、このような手法は「分散表現」とも呼ばれている。

doc2vec

doc2vec は、word2vecの拡張であり、（単語ではなく）任意の長さの文書を数百次元の固定長ベクトルとして表現する手法である。doc2vec と呼ばれているが内部的には2 つの学習方法が実装されている。word2vec と同様にCBOW モデルを拡張したPV-DM（Paragraph Vector with Distributed Memory）モデルとSkip-gram モデルを拡張したPVDBOW（Paragraph Vector with Distributed Bag of Words）モデルの2 種類のニューラルネットワーク構造が組み込まれている。PV-DBOW は単語の順序を考慮しないシンプルなモデルで計算効率が良く、PV-DM は単語の出現頻度と出現順序を考慮したモデルでPV-DBOW と比べると少し複雑でより多くのパラメータが必要になる。

word2vec 、doc2vec の実行にはgensim（Python 用のトピックモデルライブラリ）を使用した。