

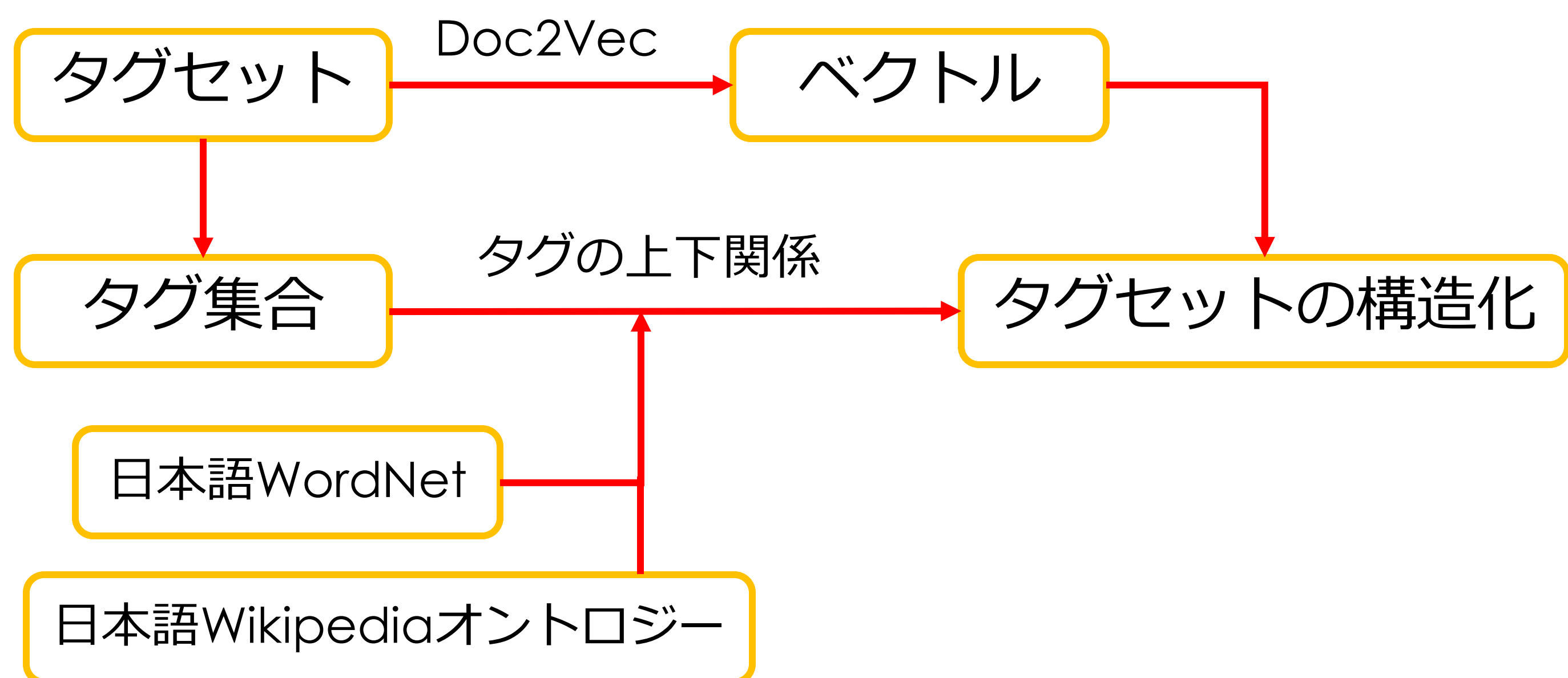
深層学習とパターンマイニングを利用したニコニコデータセットの分析に向けて

河野邊早, 鈴木宏明, 尾崎知伸(日本大学)

目的

- ニコニコ動画の動画の階層化を行う
- 前段階として、タグセットの階層化を行う

提案手法



データセットの確認

タグセットをカテゴリタグの一部で振り分けた。

- JWN:日本語WordNetで上位語または下位語が獲得できたタグ
- JWO:日本語Wikipediaオントロジーで上位語または下位語が獲得できたタグ
- 動画数: タグセットのサイズが2以上の動画の数

	動画数	タグの種類数	JWN	割合	JWO	割合
エンターテインメント	302672	19030	1935	10.17	6908	36.3
音楽	818934	43574	4348	9.98	13978	32.08
歌ってみた	532093	26888	1967	7.32	4868	18.1
演奏してみた	135329	8221	864	10.51	2715	33.03
踊ってみた	74355	5205	391	7.51	1037	19.92
VOCALOID	217538	11453	577	5.04	1819	15.88
動物	90785	3501	613	17.51	1173	33.5
料理	34149	2215	363	16.39	810	36.57
自然	122685	6862	997	14.53	2868	41.8
スポーツ	130018	8345	669	8.02	3312	39.69
歴史	26726	2167	354	16.34	975	44.99
政治	84090	5187	579	11.16	2093	40.35
科学	34026	2388	519	21.73	1065	44.6
技術部	20646	1455	196	13.47	499	34.3
作ってみた	22647	1475	207	14.03	541	36.68
アニメ	322090	20004	1261	6.3	6615	33.07
ゲーム	348913	109096	6151	5.64	19194	17.59
ラジオ	105579	6527	365	5.59	2711	41.54
描いてみた	43062	2693	191	7.09	698	25.92
Other	1376623	78535	5020	6.39	19598	24.95

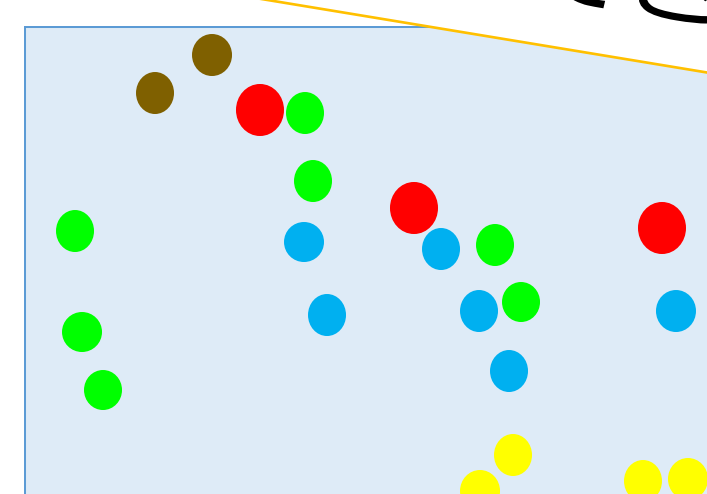
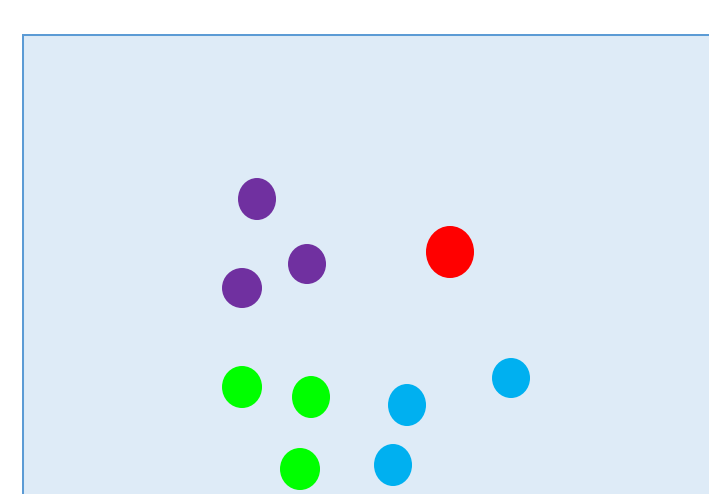
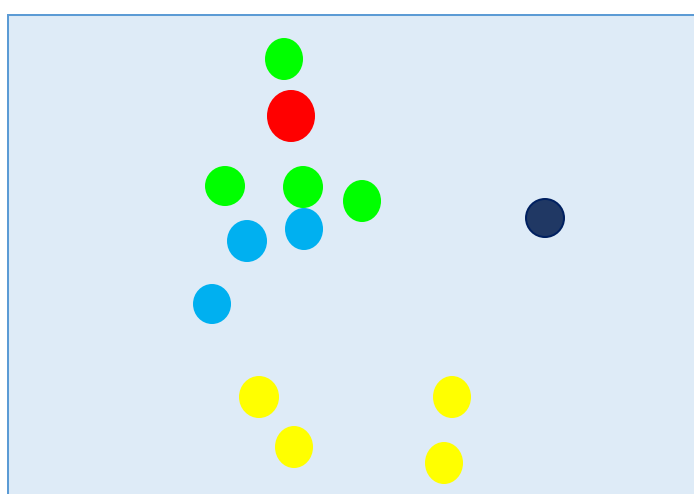
実験: タグセットのベクトル化

- タグセットをDoc2Vecでベクトル化
- 「アニメ」タグを持つタグセットのみを使用
- ベクトルをX-meansでクラスタリング
- クラスタ数は60

クラス	クラス中心のタグセット	クラス	クラス中心のタグセット
Cluster 0 タグセット数 (862)	アニメ, 宇宙戦艦ヤマト, 涼宮ハルヒの憂鬱		
Cluster 1 タグセット数 (915)	アイキャッチ, アニメ, ゆるゆり, 七森中ぐらぐら部, 吉川ちなつ, 蔵納京子, 船見結衣, 赤座あかり, アッカリン	Cluster 50 タグセット数 (3251)	fripSide, infinite_synthesis, stay_with_you, アニメ, 天元突破グレンラガン
Cluster 2 タグセット数 (2730)	アニメ, インタビュー, タイドライン・ブルー, 栗林みな実	Cluster 51 タグセット数 (2988)	ONE, PIECE, アニメ, ウィーアー! for NEW WORLD, きただにひろし
Cluster 3 タグセット数 (2018)	アニメ, 新きまぐれオレンジ☆ロード, 梶浦由記, 音楽, 鮎川まどか	Cluster 52 タグセット数 (9628)	アジカン MAD, アニメ, イヤホン推奨, インデックス, とある魔術の禁書目録, 一方通行, 上条当麻, 修正版, 打ち止め, 禁書良作 MAD
Cluster 4 タグセット数 (1806)	cm, アニメ, ツインタニシ, 機動戦士ガンダム 00	Cluster 53 タグセット数 (5583)	1st, 6 分で見えるガンダムシリーズ, アニメ, ウッディ, ガンダム, 機動戦士ガンダム, ガンダムシリーズダイジェスト, 大河原邦男, 安彦良和, 富野由悠季
Cluster 5 タグセット数 (1543)	Cloud_Age_Symphony, アニメ, エリキヤロ, 魔法少女リリカルなのは	Cluster 54 タグセット数 (5832)	アニメ, なのは VS シリーズ, 魔法少女リリカルなのは strikers
Cluster 6 タグセット数 (3255)	アニメ, 木村良平, 浅野真澄, 生天目仁美, 福圓美里, 織田信奈の野望	Cluster 55 タグセット数 (6374)	BGM バイバイ, アニメ, サムネイホイ, 不協和音, 俺の妹がこんなに可愛いわけがない, 新境地 OP
Cluster 7 タグセット数 (2765)	アニメ, スマイルプリキュア	Cluster 56 タグセット数 (6319)	Scratch on the Heart, SD ガンダム フォース, アニメ, ゴルドラン, だいたいあつてる, メイプルストーリー, エリナ, 主題歌差し替える MAD, 桜蘭高校ホスト部, 無編集
Cluster 8 タグセット数 (1468)	アニメ, じじい, チルノのパーフェクトさんすう教室, ローゼンメイデン, 水銀燈	Cluster 57 タグセット数 (13708)	アニソン, アニメ
Cluster 9 タグセット数 (2971)	Angel Beats!, Girls Dead Monster, アニメ, ガルデモ, 一番の宝物	Cluster 58 タグセット数 (5128)	アニメ, ロボットアニメーション集, 劇場版, 坂本真綾, 天空のエスカフローネ, 戦国シモン, 菅野よう子, 間智一, 高山みなみ, 高瀬夏
Cluster 10 タグセット数 (13752)	アニメ, ある意味持っていた, シャオメイ, もっと評価されるべき, らんま, らんま 1/2, 手描き, 手描きらんま, 水島版ファン向け, 親父の存在感	Cluster 59 タグセット数 (3746)	アニメ, ガンダム AGE, ジョジョの奇妙な冒険, よんでますよ, アザゼルさん, 井上和彦, 佐藤利奈, 夏目友人帳, 機動戦士ガンダム AGE, 銀魂
.....			

目的

頻出パターンマイニング



SIFT特徴量を利用
は Visual Words

ビジュアルパターン

{{ { 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60 } }}

画像から特徴量を抽出して組み合わせを考える
共通してみられる傾向として頻出パターンが読み取れる

- 膨大な数のパターンが抽出される
→ (意味のある) パターンだけを抽出したい
- 発見したパターンがどのような意味を持つのか
解釈するのが必ずしも容易ではない。

提案手法

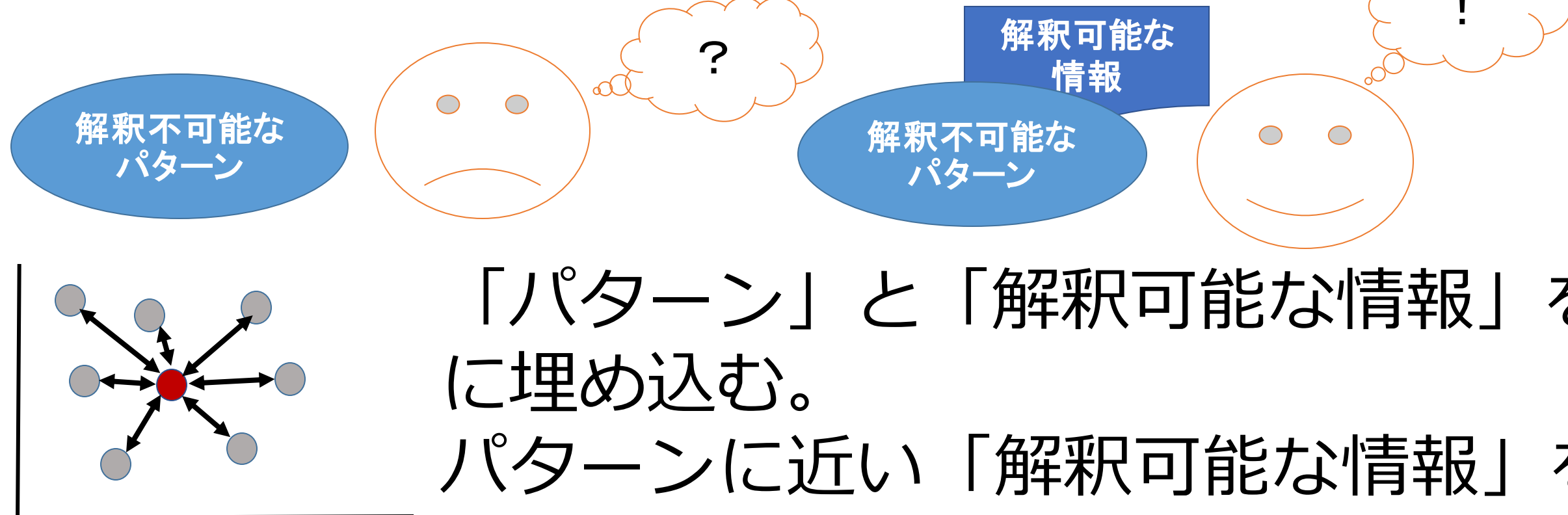
パターンの分散表現を用いた問題の解決

分散表現を利用してパターンをより高次元空間に写像

- 問題: 1 「大量パターン生成」の解決
→ 高次パターン (ベクトル) 空間での非階層的クラスタリングの適用
類似しているパターンの集まりを、一つに集約する
各クラスタから代表 (クラスタ中心) を抽出する
→ クラスタ数 = パターンの数

問題: 2 「パターンの理解困難性」の解決

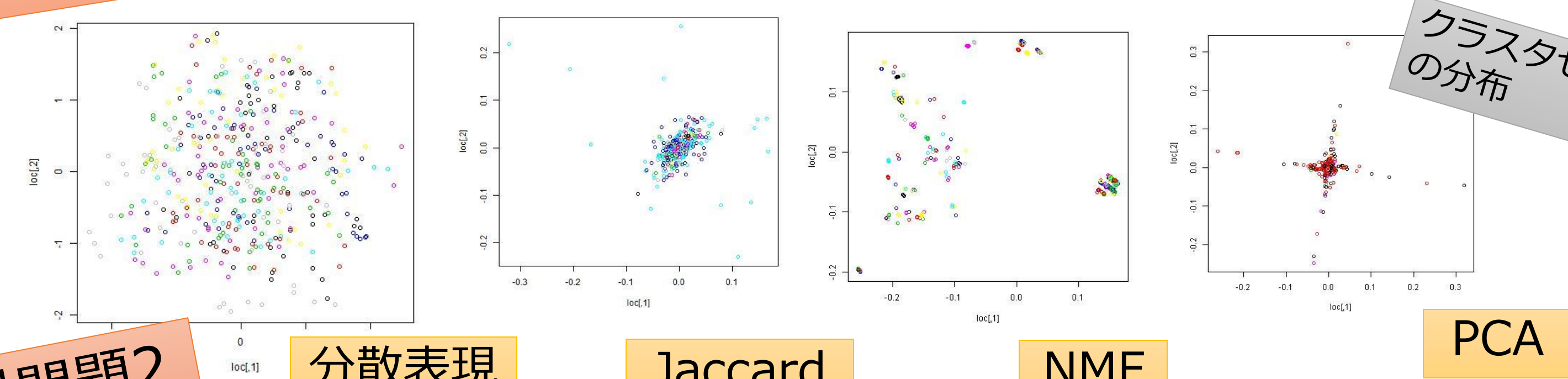
→ パターンの理解を補助するような
解釈可能な異なった情報を付与



実験

対問題1

NII より株式会社ダウンゴから提供を受けたニコニコデータを使用



対問題2

タグパターン	ビジュアルパターン
バスケットボール	P ₈₇ , P ₆₂₂ , P ₁₃₉₈
プロ野球, 野球	P ₈₇ , P ₂₈₅ , P ₁₁₂₉
野球, mlb	P ₈₇ , P ₆₂₂ , P ₁₃₉₈
読売ジャイアンツ	P ₆₂₂ , P ₁₃₉₈ , P ₁₄₈₈
阪神タイガース	P ₂₈₅ , P ₆₂₂ , P ₁₃₉₈
競艇	P ₈₇ , P ₆₂₂ , P ₁₃₉₈
格闘技, k-1	P ₈₇ , P ₁₃₂₉ , P ₁₃₉₈
サッカー, jリーグ	P ₈₇ , P ₂₈₅ , P ₆₂₂
車載動画, モータースポーツ	P ₈₇ , P ₆₂₂ , P ₁₃₉₈

ニコニコ動画のサムネ画像「スポーツ」タグをもつ8万件からランダムに選んだ2万件のデータ
Visual Words作成: 5000件の画像データ100 words作成
クラスタ数: 200
分散表現 GloVeを利用
次元数: 50

まとめ

「大量パターン生成」, 「解釈困難性」を緩和するため
分散表現技術のパターンマイニング分野への導入を行った。

今後の課題として更なる理解容易性のために

- 「加法構成性」に基づいた解釈
- 積層ノイズ除去自己符号化器の利用