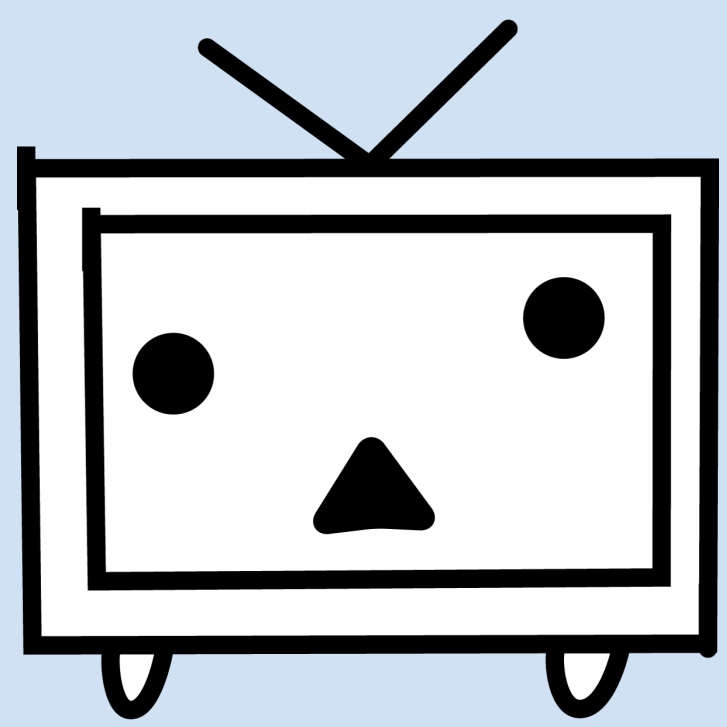




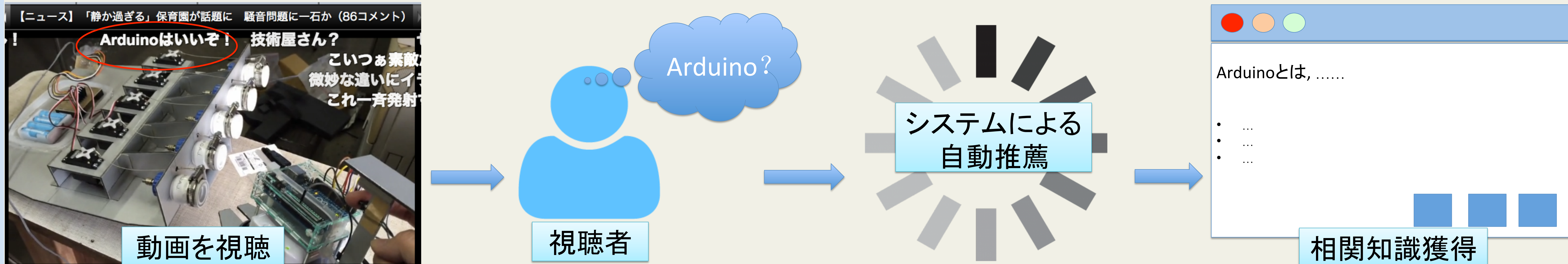
ニコニコ動画のコメントを利用した 相関知識推薦システムの提案

電気通信大学
劉英楠 沼尾雅之



研究の概要

近年ニコニコ動画や、ニコニコ生放送などの、コメント付きの動画サービスが発展してきた。投稿されたコメントに、様々な有用な関連情報が入っているので、これらをデータマイニングにより抽出し、視聴者に関連サイトへのリンクを紹介する推薦するシステムを提案する。



コメント形式

- ニコニコ動画のコメントは、JSON形式で構造化表現されている。

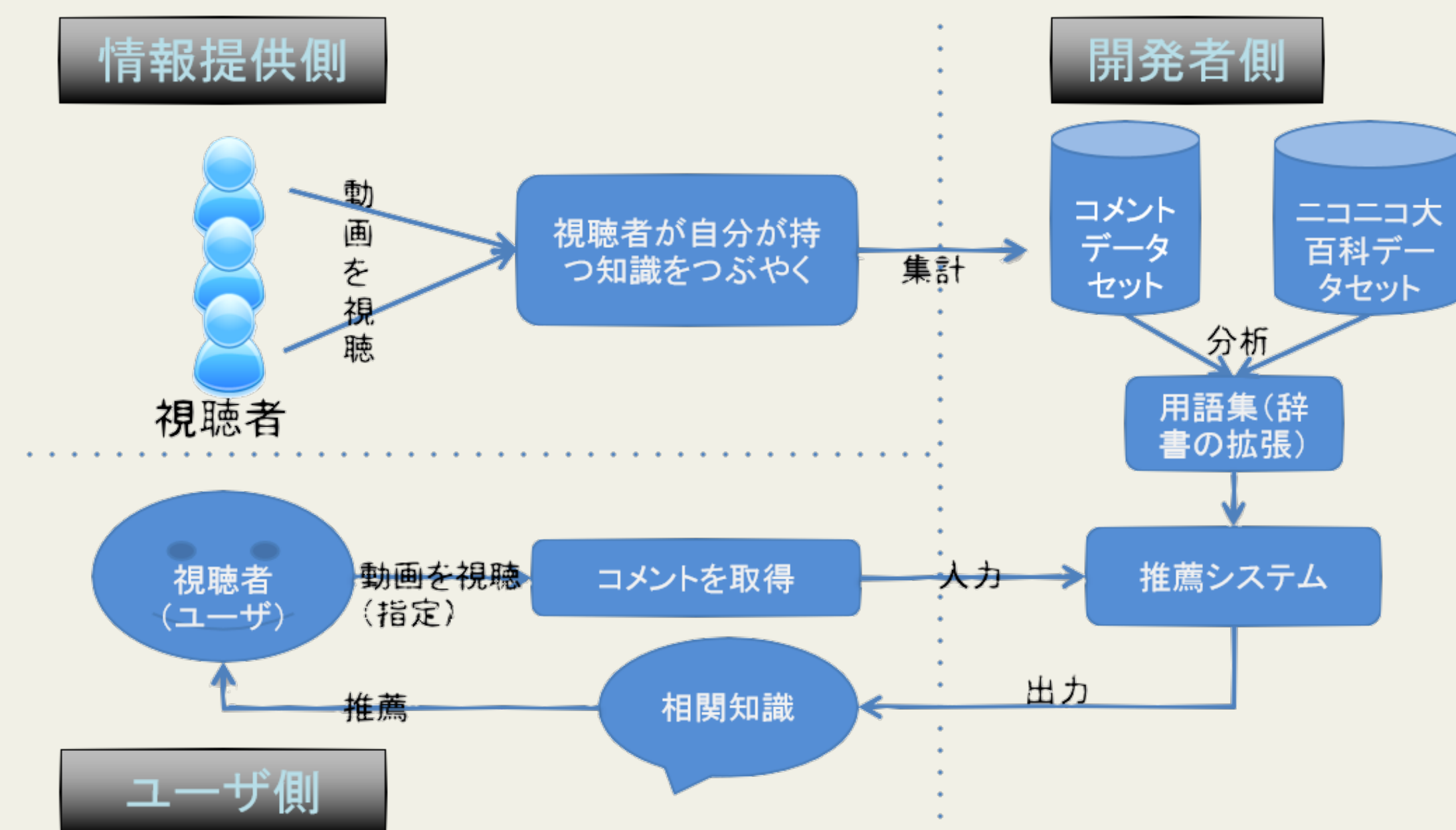
```
{  
  "command": "shita",  
  "comment": "抱きしめて 星空に",  
  "date": 1184420666,  
  "no": 50,  
  "vpos": 8041  
}
```

キー	説明	解説
"data"	コメントの投稿日時	2007.7.14, 22:44:26に投稿された
"no"	コメントごとに振られる番号	50番目つけられたコメント
"vpos"	再生位置	80.41sに表示される
"comment"	コメント本文	"抱きしめて 星空に"
"command"	フォントや位置を指定	下側に出現する

- 本研究では、主に再生位置とコメント本文を中心に解析を行う。

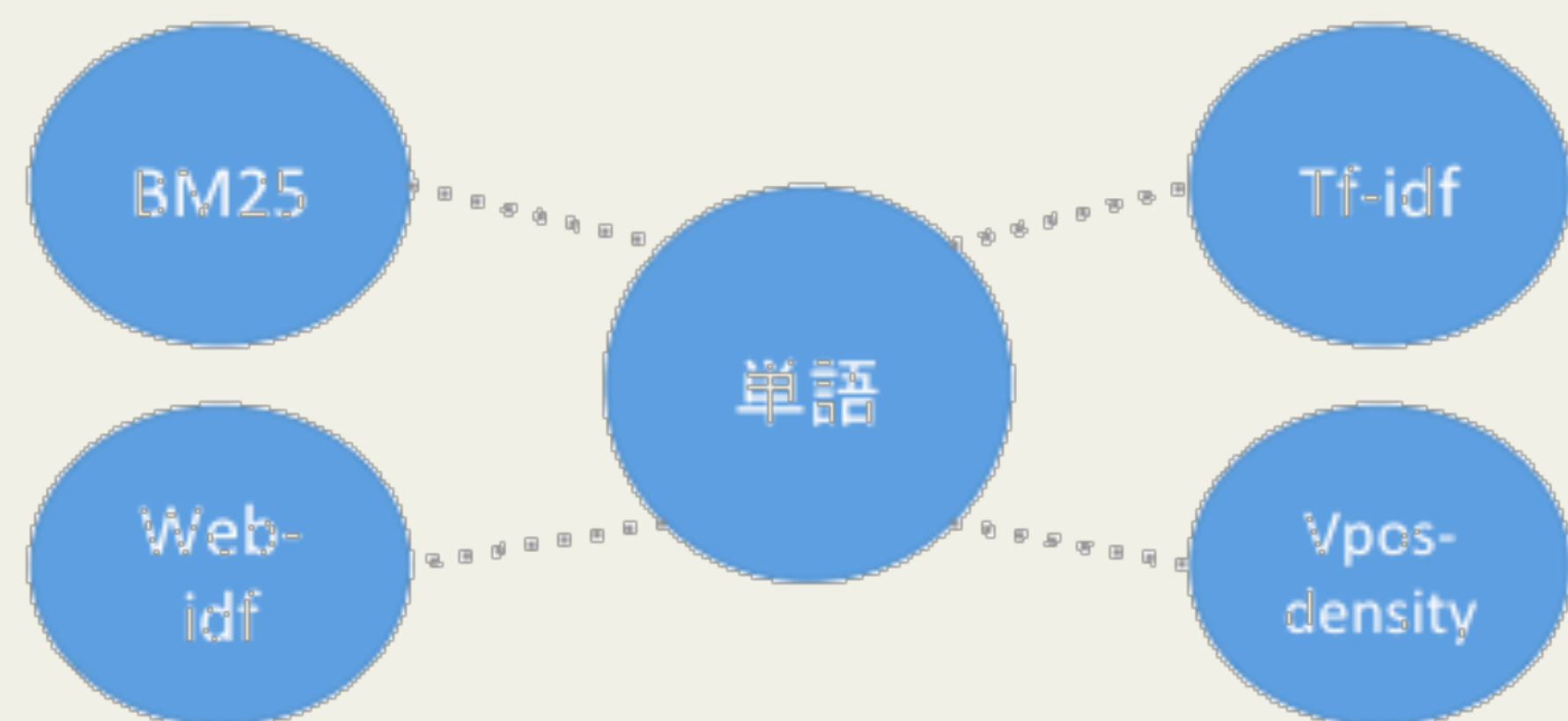
提案プロセス

- ニコニコ動画のコメントを集めたデータセットを解析し、評価用の辞書を作る。
- 特定の動画のキーワードを抽出。
- 抽出したキーワードが関連知識であるかどうかを判断。
- ユーザーに関連知識(リンク)を推薦する。



単語の評価基準

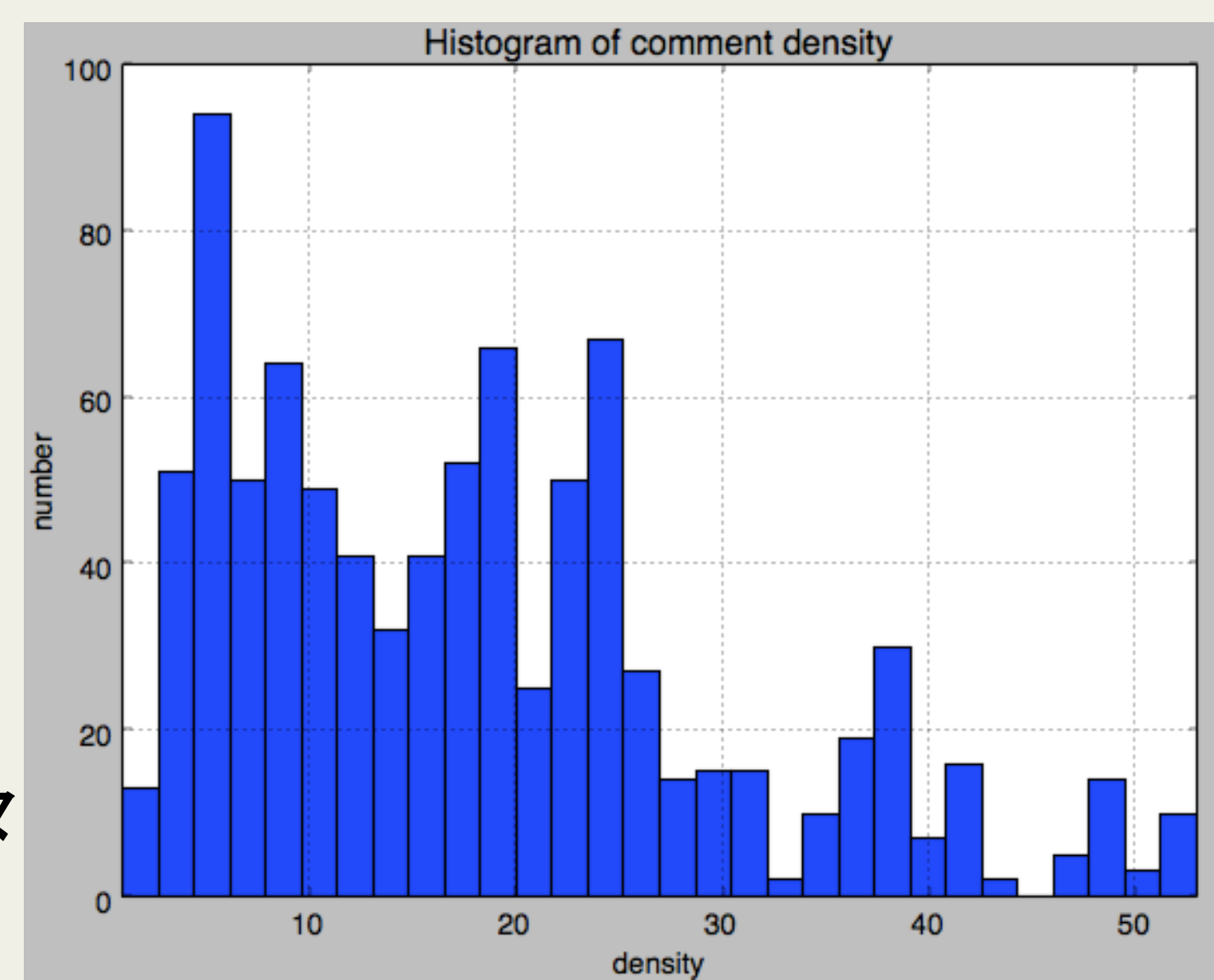
- 多様な基準で単語の重要度を評価する。
- BM25
- Tf-idf
- Vpos-density
- Web-idf



予備実験

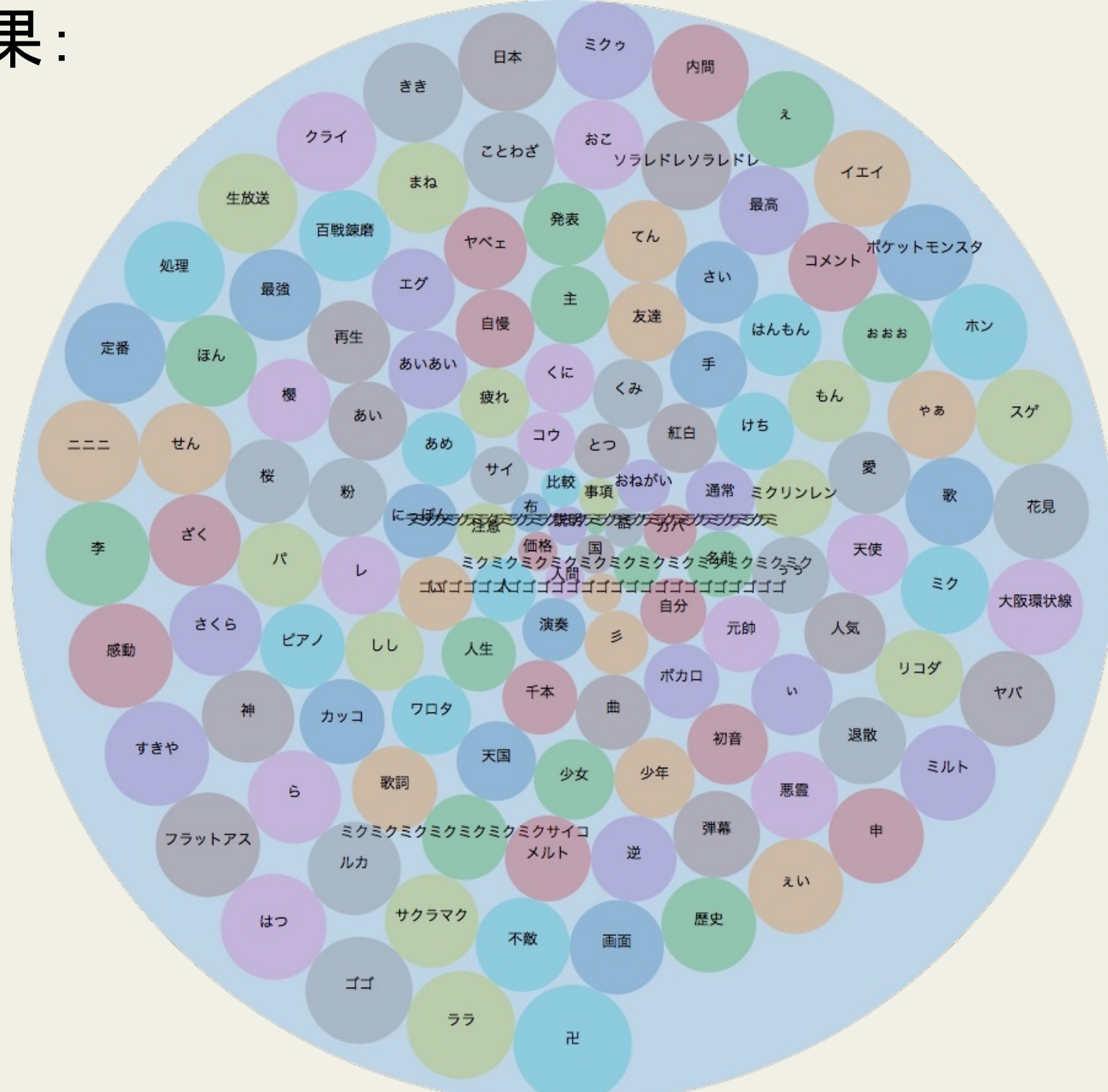
動画sm15630734,『初音ミク』千本桜『オリジナル曲PV』
コメント1000件に対して以下の評価を行った：

- 動画内でのBM25, Tf-idf, Vpos-density(右図)の計算。
(横軸: 同じスクリーンに出現するコメントの件数(Vpos-density).
縦軸: このコメント密度のコメント数.)
- 実験的な, Web-idfの取得.Google検索エンジンにリクエストを投げ、ヒット数を抽出。
- BM25, Tf-idf, Vpos-densityを使用し, 抽出した個々の単語に, 点数をつけ, 可視化してみた。



予備実験結果

予備実験で得られたキーワードと点数を可視化した結果：



Javascriptの
ライブラリ,
D3.jsを利用

*円の半径が
大きければ,
得点が高い。

実験および今後の課題

単一動画だと、抽出する単語は動画に依存しすぎることが分かり、計算の「全文書」を「データセット全体」に拡大したい。つまり、データセットを「全文章」としたTf-idf BM25の計算を行いたい。現状、以下のことができた：

- tf値を動画ごとに計算し、出力結果をJSON形式で保存。
- idfを計算するために単語ごとの出現回数をカウントした。

今後の課題：

- Web-idfの獲得方法の検討。
- MeCabの辞書を「ニコニコ大百科」のインデックスで拡張し、拡張前の結果と比較する。
- 単語のポイント付け方法の検討。
- 実験結果の評価方法を考える。

長期的目標：

- Hadoopを実装し、コメントファイルへのアクセスを加速させたい。
- ウェブアプリの形に仕上げ、サービスとして提供してみたい。

