ニュース記事を対象とした トピック解析を用いたバイアス発見方式

柳瀬 愛里†　木村 侑斗†　萩本 新平†　中田 亮佑†
中村 洋太†　本田くれあ†　仲程 凜太朗†　中西 崇文†

†武蔵野大学 データサイエンス学部 データサイエンス学科

第13回データ工学と情報マネジメントに関するフォーラム(DEIM) 2021
2021月3月1-3日 | DEIM2021オンライン会議ポータル H-14

目次

目次

研究背景 ： 現代のメディアにおけるトラブル

- **サイバーカスケード**

  インターネット上において，多くの人が自分と同じ考えの意見を見つけ，同調し合い，反対の立場の意見を無視・排除し，各々の主義主張を極端に先鋭化させることで，議論の収束先が絞り込まれてしまう現象．

  - **エコチェンバー**

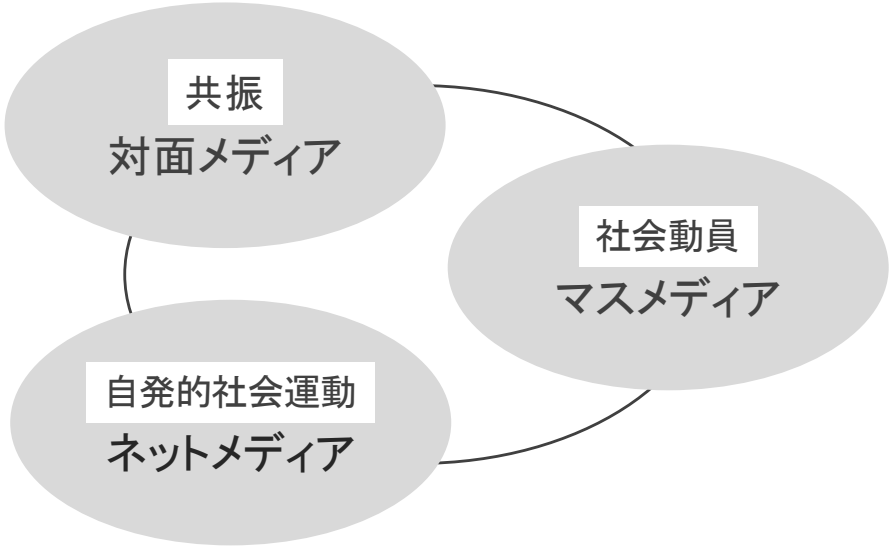
    SNSにおいて，価値観の似た者同士が交流し，共感し合うことで，特定の意見や思想が増幅されて影響力をもつ現象．

  - **フィルターバブル**

    検索エンジンの最適化アルゴリズムがユーザの見たくない情報を遮断することで，好みの情報が集まっている泡の中に閉じ込められ，外の情報から遮断されている様子を表す造語．

研究背景 ： 私たちを取り巻く「間メディア社会」の問題

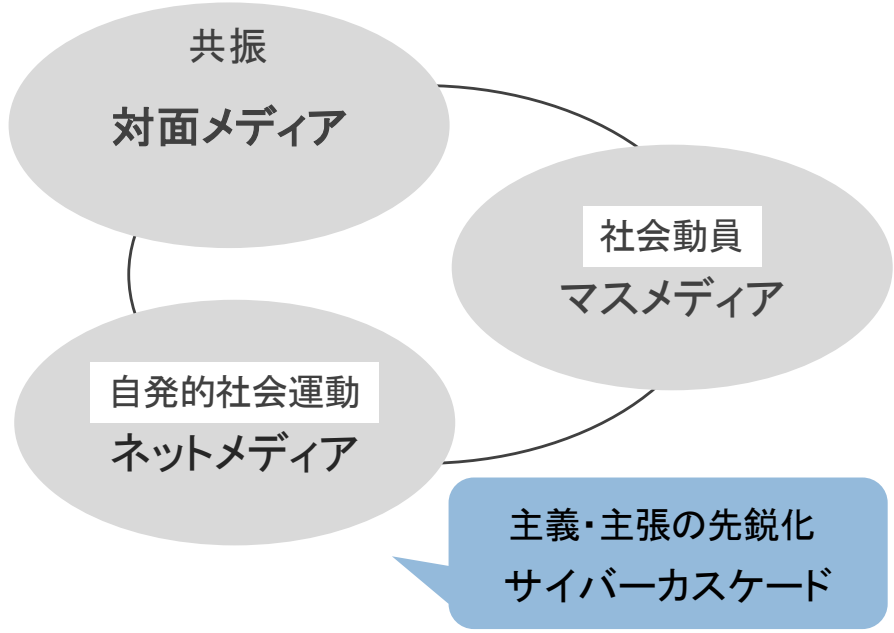
- 「間メディア社会」[*1]とは
  ソーシャルメディアと既存のメディアが
  重層的に相互作用しながら世論を形成する
  現代のメディア環境のこと.
- これまで大きく着目されることのなかった出来事が，メディア間の相互作用が緊密化することにより，社会を揺るがす大きな影響力を持つようになった.
  Ex）新型コロナウイルス流行下のデマ



[*]1：遠藤薫，”間メディア民主主義と<世論>”，社会情報学第5巻1号，2016.

研究背景 ： 私たちを取り巻く「間メディア社会」の問題

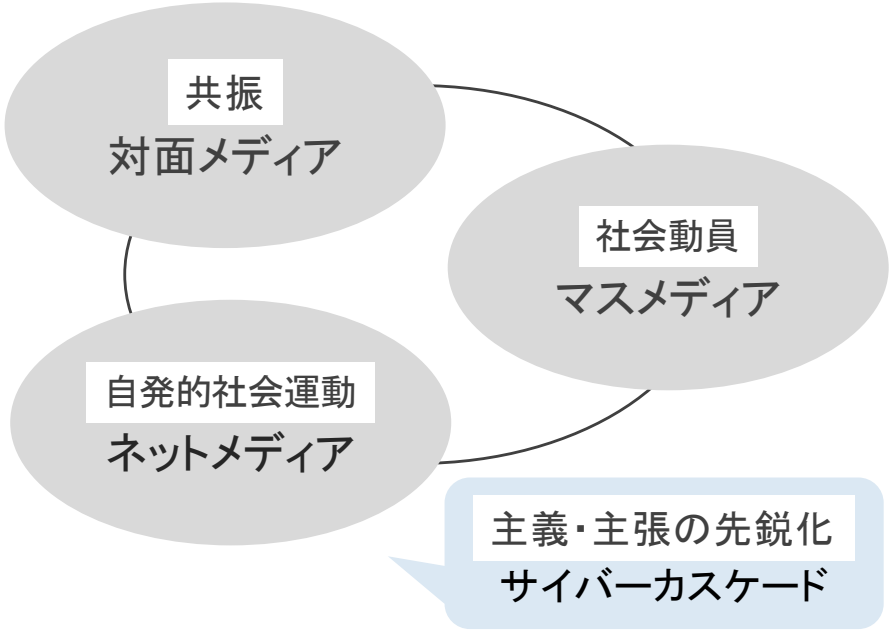
- 「間メディア社会」[*1]とは
  ソーシャルメディアと既存のメディアが
  重層的に相互作用しながら世論を形成する
  現代のメディア環境のこと.
- これまで大きく着目されることのなかった出来事が，メディア間の相互作用が緊密化することにより，社会を揺るがす大きな影響力を持つようになった.
  Ex）新型コロナウイルス流行下のデマ



[*]1：遠藤薫，”間メディア民主主義と〈世論〉”，社会情報学第5巻1号，2016.

研究背景 ： 私たちを取り巻く「間メディア社会」の問題

- 「間メディア社会」[*1]とは
  ソーシャルメディアと既存のメディアが
  重層的に相互作用しながら世論を形成する
  現代のメディア環境のこと.
- これまで大きく着目されることのなかった出来事が，メディア間の相互作用が緊密化することにより，社会を揺るがす大きな影響力を持つようになった.
  Ex）新型コロナウイルス流行下のデマ



[*]1：遠藤薫，”間メディア民主主義と〈世論〉”，社会情報学第5巻1号，2016.

研究目的

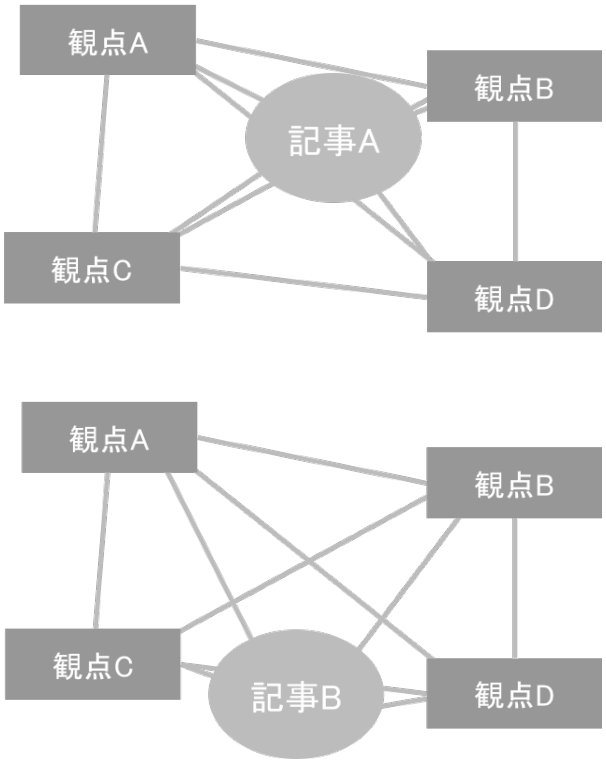
ユーザが入力したニュースの立ち位置を，

関連ニュース記事群が持つ観点(トピック)

との関係と共に可視化・提示することで，

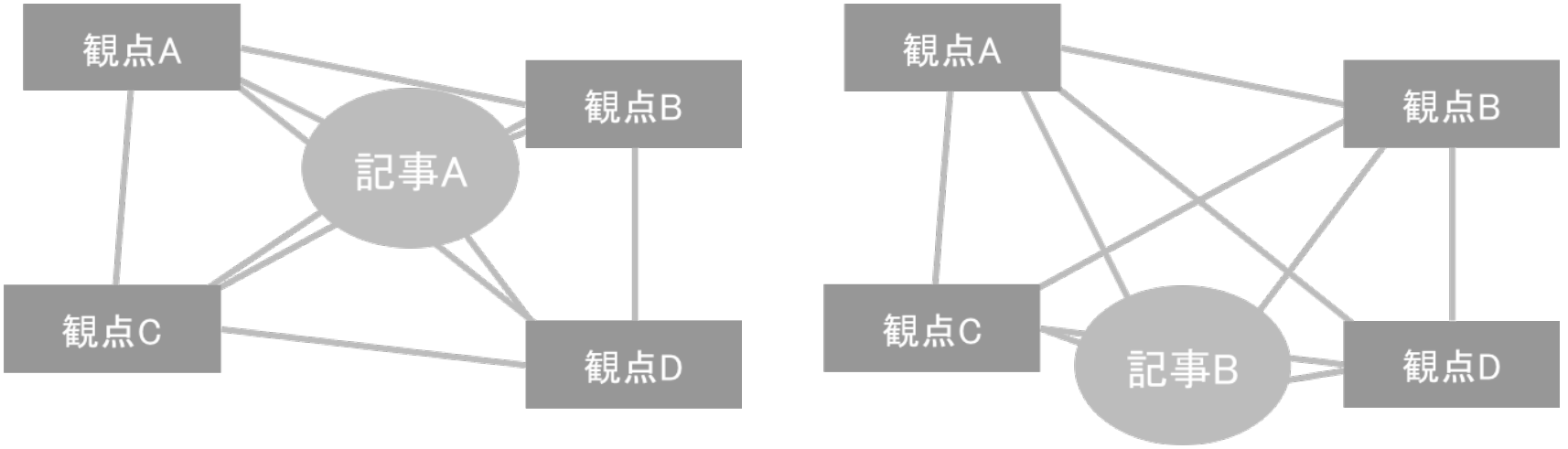
ニュースの情報の偏り(バイアス)の有無を

認識させる手法の提案を行う.



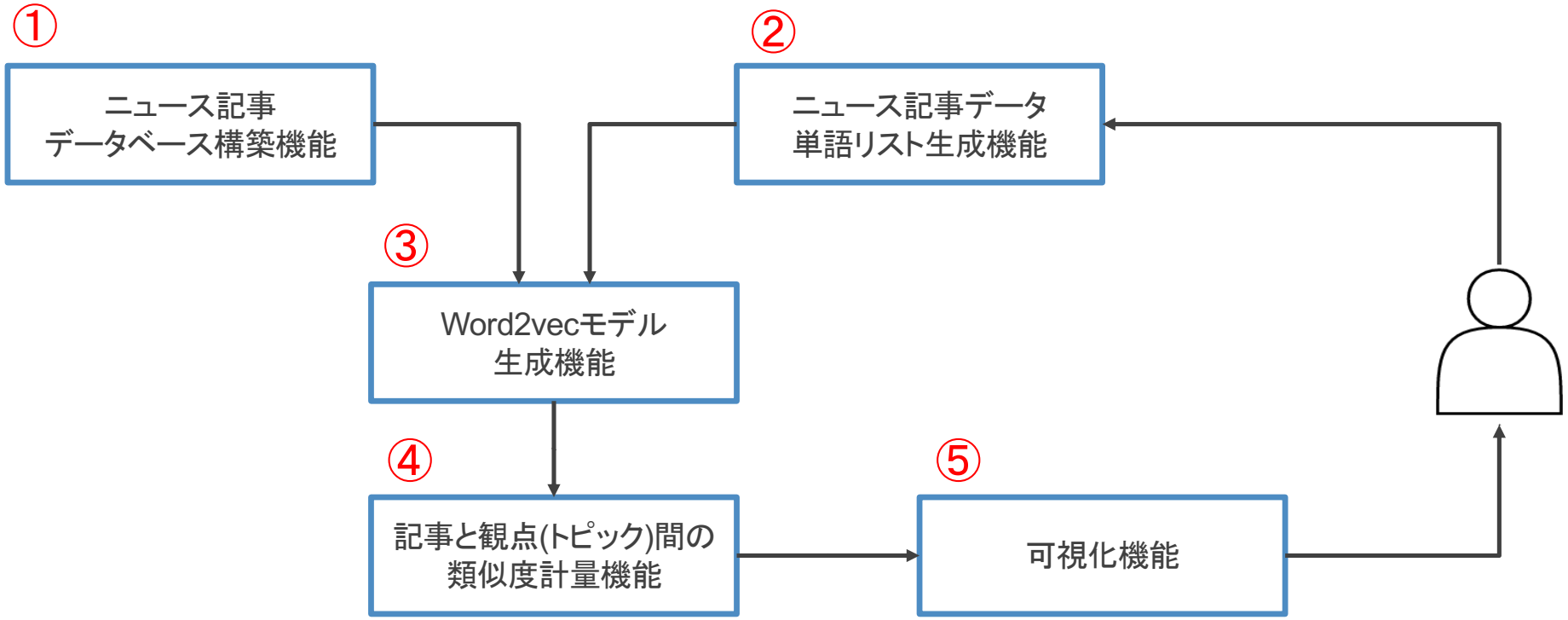
研究目的

ユーザが入力したニュースの立ち位置を，関連ニュース記事群が持つ
観点(トピック)との関係と共に可視化し，提示することにより，
ニュースに情報の偏り(バイアス)があるかどうかを認識させる手法の提案を行う.

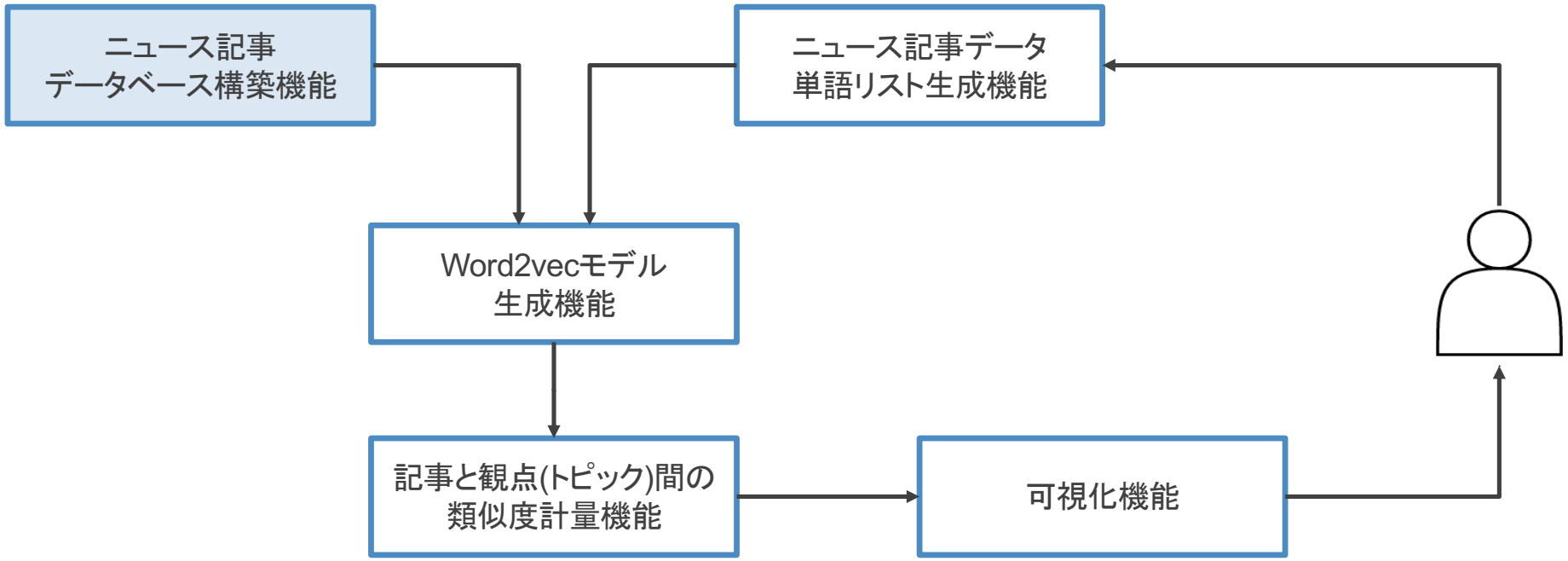


目次

システム構成図

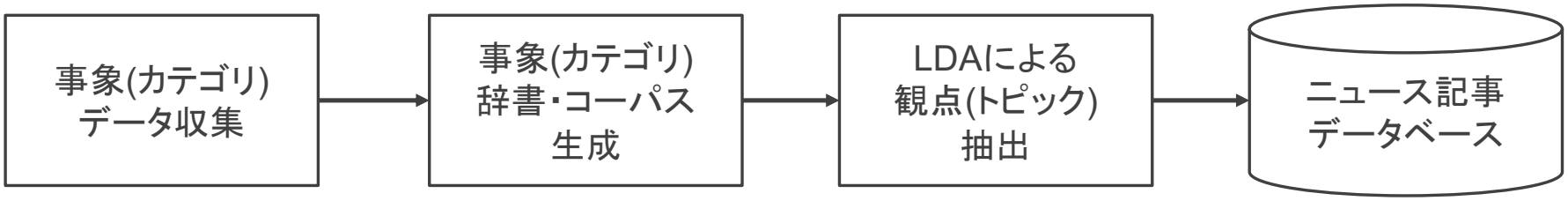

①
ニュース記事
データベース構築機能
②
ニュース記事データ
単語リスト生成機能
③
Word2vecモデル
生成機能
④
記事と観点(トピック)間の
類似度計量機能
⑤
可視化機能

システム構成図


ニュース記事
データベース構築機能
ニュース記事データ
単語リスト生成機能
Word2vecモデル
生成機能
記事と観点(トピック)間の
類似度計量機能
可視化機能

ニュース記事データベース構築機能

様々なカテゴリに属するニュース記事群を収集し，それぞれの観点(トピック)の抽出を行い，ニュース記事データベースを構築する．



*2： Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." the Journal of machine Learning research 3 (2003): 993-1022.

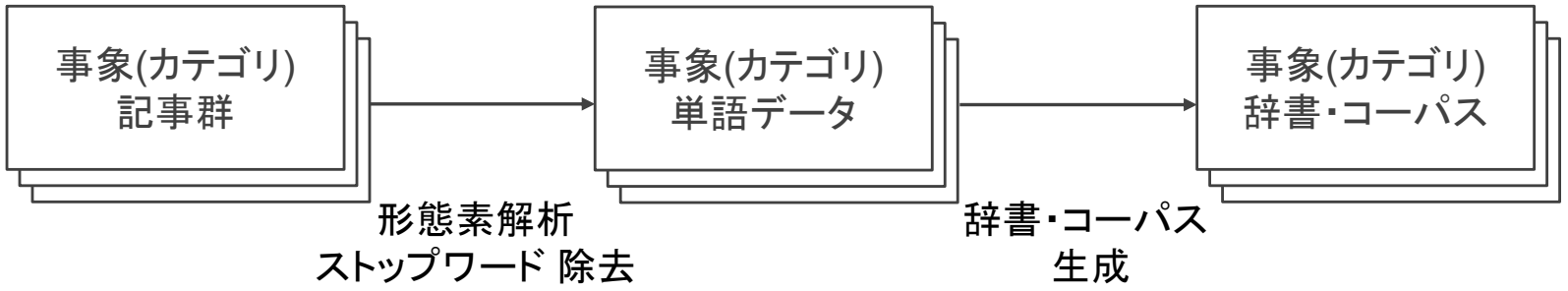
事象(カテゴリ)データ収集部

有用な情報が得られると判断したニュースサイトを選定し，これらに対して，
キーワードと一致する記事のスクレイピングを実行し，事象(カテゴリ)記事群を収集する．

本研究では，キーワードを入力し，特定の事柄について記述された事象(カテゴリ)データ収集する．

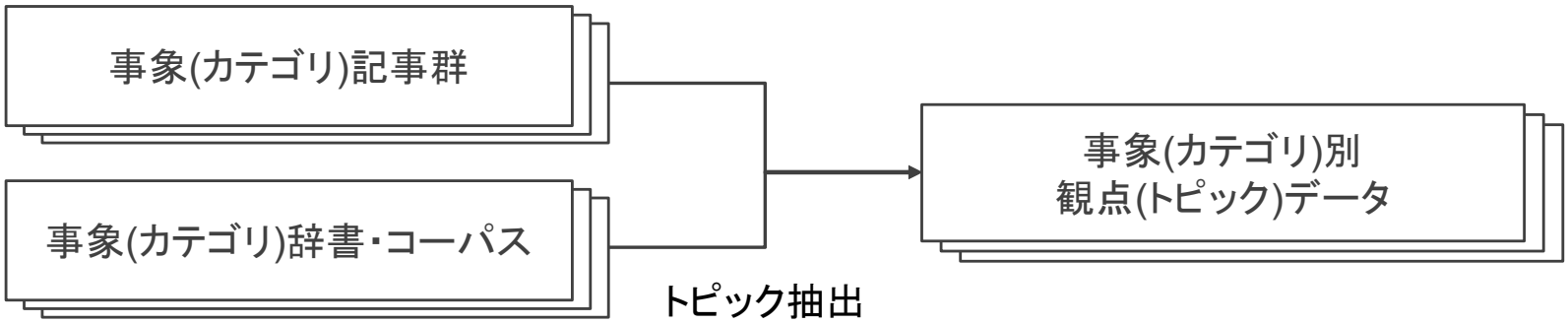
事象(カテゴリ)辞書・コーパス生成部

収集したカテゴリニュースデータ記事群の本文を対象に，形態素解析とストップワード除去を行い，本文中に含まれている単語を抽出し，辞書とコーパスの生成を行う．



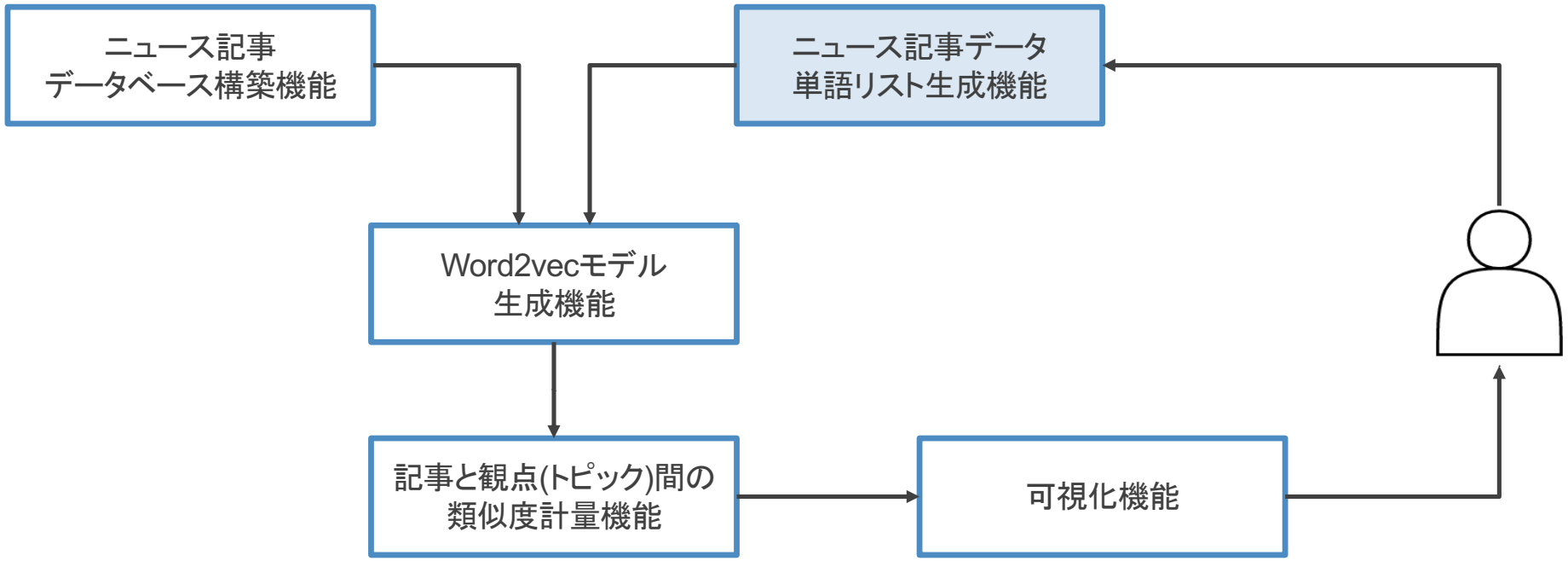
LDAによる観点(トピック)抽出部

事象(カテゴリ)別に格納したニュース記事群がどのような観点から記述されているかを明らかにするために，ニュース記事群に対してLDAを用いたトピック抽出を行う.



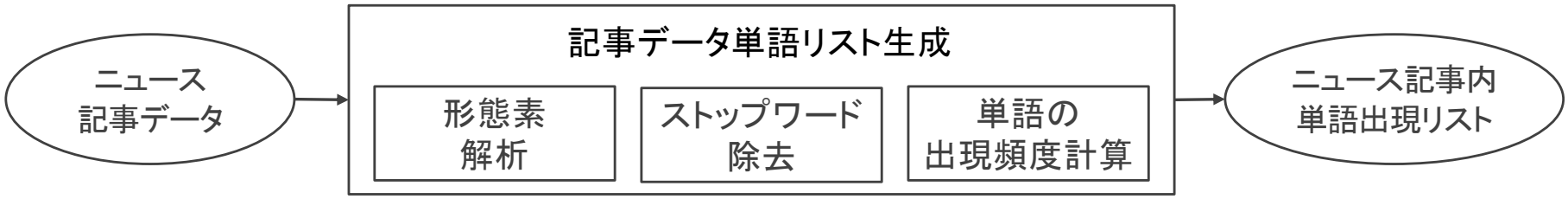
ここで抽出された観点(トピック)は，それ自体がカテゴリで記述されている事象を説明・記述する基準となる重要な観点を示すものである.

システム構成図


ニュース記事
データベース構築機能
ニュース記事データ
単語リスト生成機能
Word2vecモデル
生成機能
記事と観点(トピック)間の
類似度計量機能
可視化機能

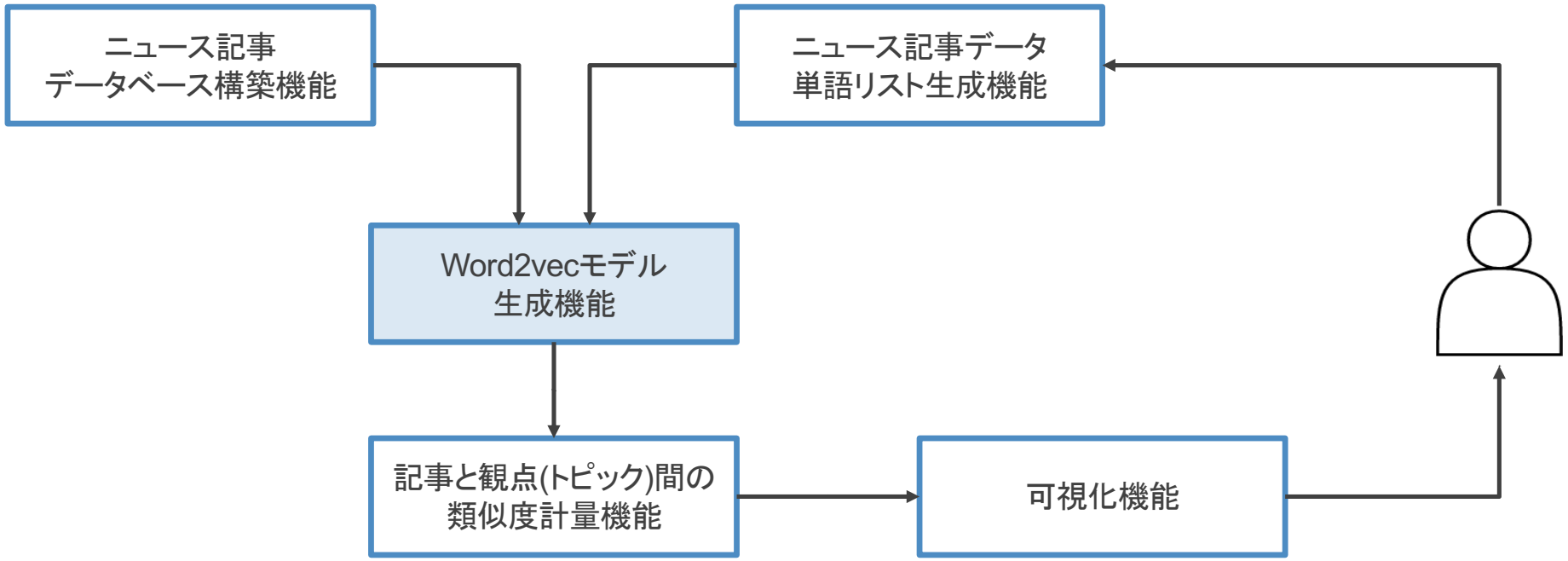
ニュース記事データ単語リスト生成

ユーザが着目するニュース記事データに対して，形態素解析・ストップワード除去・
単語の出現頻度計算を行い，記事に含まれる単語の出現頻度をリストにまとめる.



ここで得られたニュース記事内単語出現頻度リストを用いて，
観点(トピック)と記事間の類似度計量を行う.

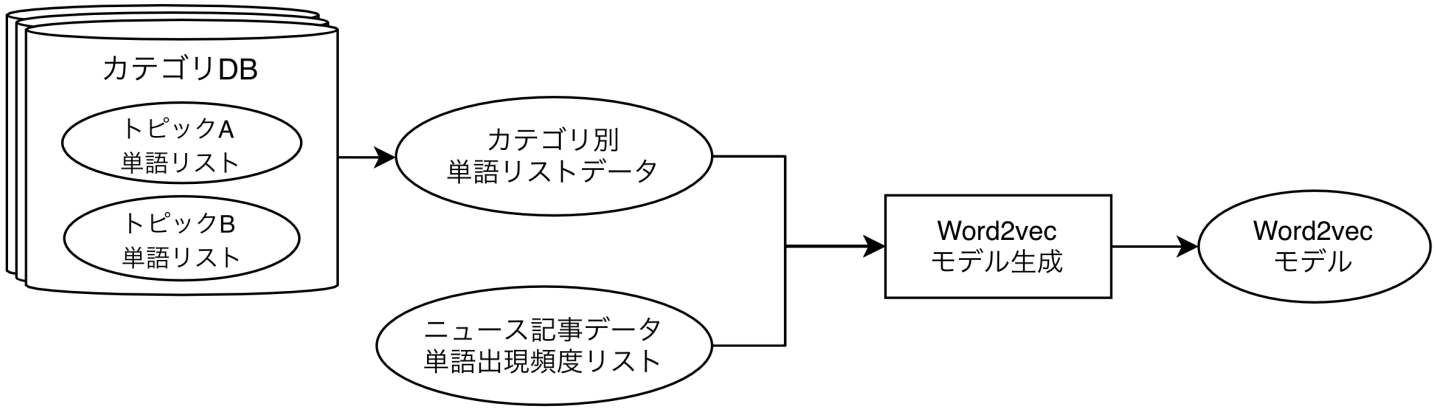
システム構成図


ニュース記事
データベース構築機能
ニュース記事データ
単語リスト生成機能
Word2vecモデル
生成機能
記事と観点(トピック)間の
類似度計量機能
可視化機能

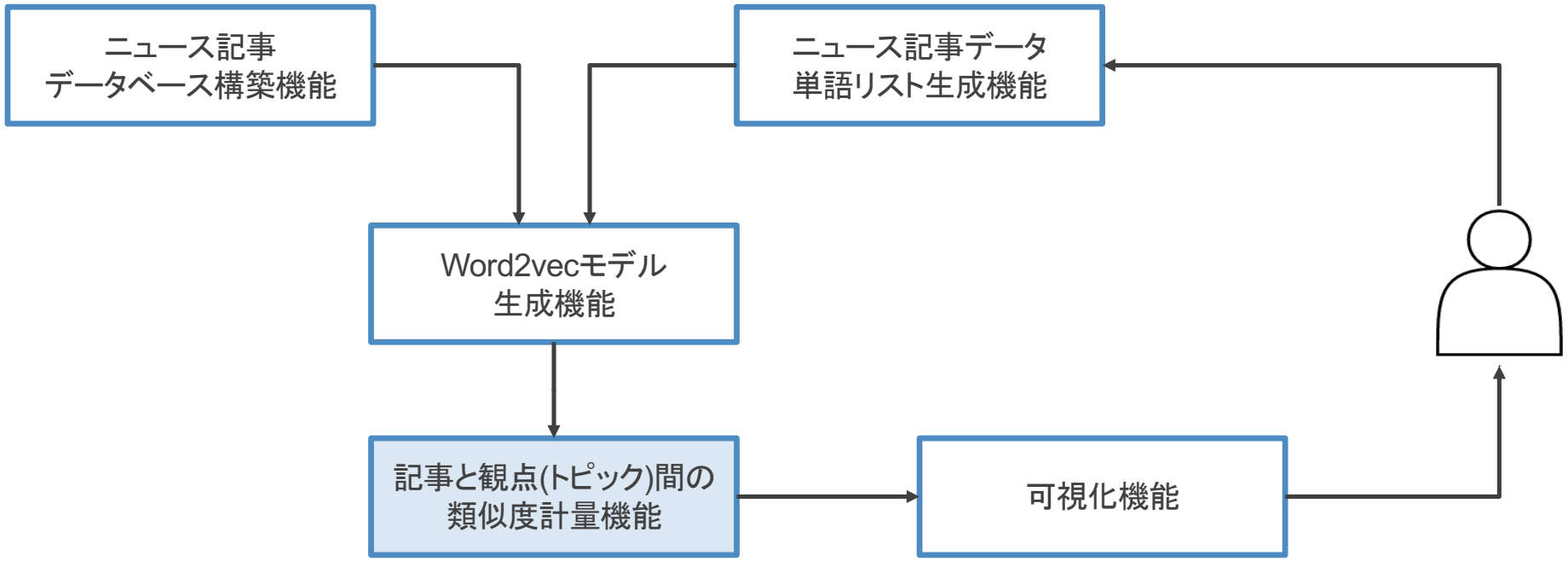
Word2vec モデル生成

あるカテゴリに含まれている単語同士の関係性を明らかにするために，カテゴリごとにWord2vecモデルの生成を行う.

DBに格納されているカテゴリデータとユーザが着目しているニュース記事データから生成されるニュース記事データ単語リストを利用し，Word2vecのモデルを生成する。

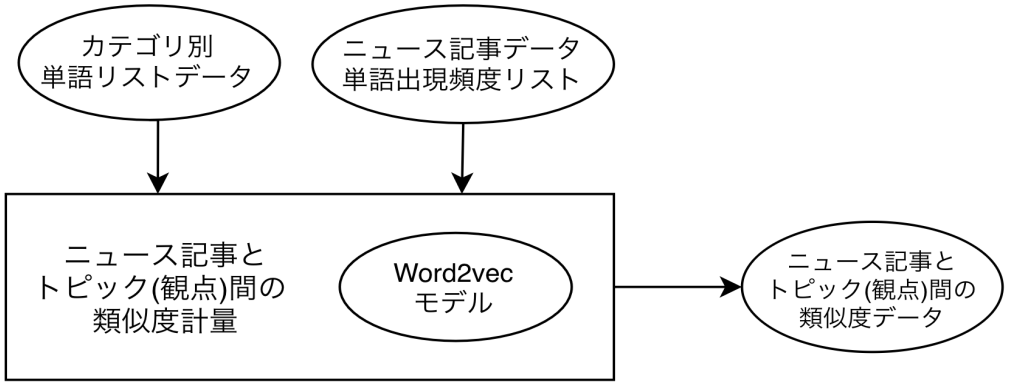


システム構成図


ニュース記事
データベース構築機能
ニュース記事データ
単語リスト生成機能
Word2vecモデル
生成機能
記事と観点(トピック)間の
類似度計量機能
可視化機能

記事と観点(トピック)間の類似度計量

ニュース記事データ単語リストとカテゴリ別単語リストデータをWord2vecモデルで分散表現ベクトルに変換し，これらに対しWord Mover’s distanceで類似度計量を行う.



これらを次に示す可視化機能部に受け渡すことにより，ユーザが支持しているニュース記事の位置づけをニュースカテゴリの観点を用い，可視化することが可能となる.

Word Mover's Distance

文章間の距離を計算するための手法

$$\min_{\mathrm{T}\geq 0} \sum_{i,j=0}^{n} \mathbf{T}_{i,j}\ c(i,j)$$

$$\textbf{subject to}: \min_{\mathrm{T}\geq 0} \sum_{i,j=0}^{n} \mathbf{T}_{i,j} = d_i \quad \forall \in \{1,\ldots,n\}$$

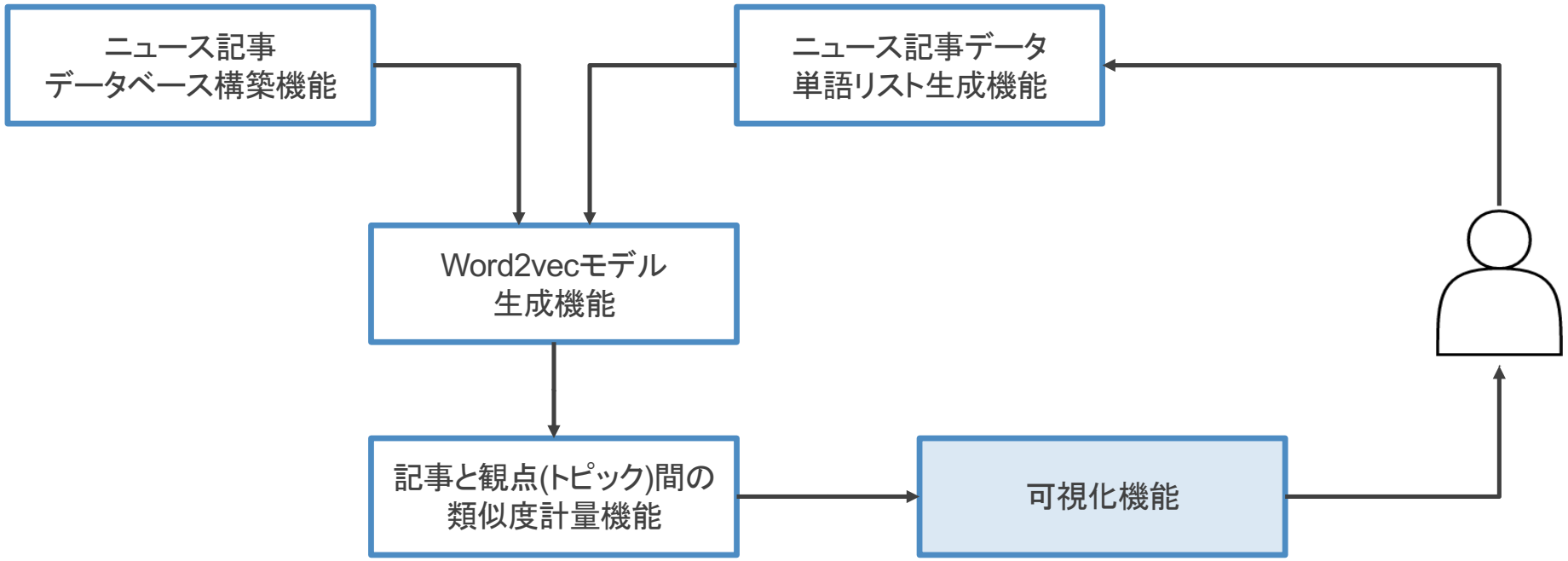
$$\sum_{i,j=0}^{n} \mathbf{T}_{i,j} = d_i \quad \forall \in \{1,\ldots,n\}$$

ある文章Aをある文章Bに変換する際の対応づけ変換コストが最も低い場合の単語間の変換コストの和を文書間距離と考える.

単語間の変換コストは，単語分散表現の差として求められる.

本研究では単語の分散表現をWord2vecを用いて求める.

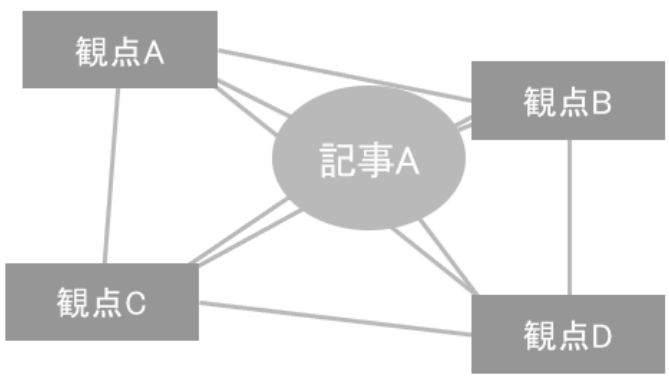
システム構成図

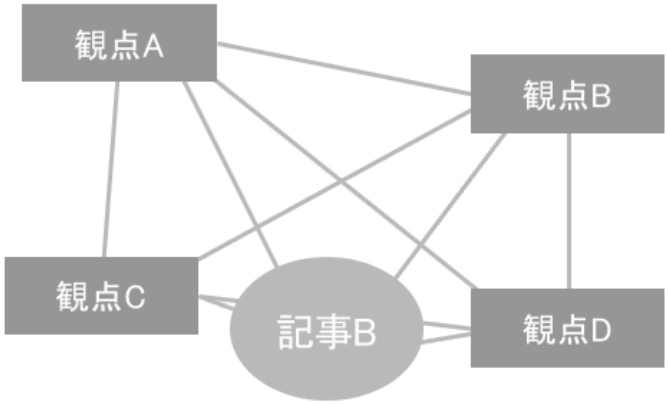

ニュース記事
データベース構築機能
ニュース記事データ
単語リスト生成機能
Word2vecモデル
生成機能
記事と観点(トピック)間の
類似度計量機能
可視化機能

可視化機能

ユーザが入力したニュース記事と導出された観点との関係を提示するために，
得られた情報から導出した結果をネットワーク図で可視化する．

エッジの長さはWord Mover’s distanceから算出された類似度を示している．





目次

1. 研究背景・目的
2. 提案方式
3. 実験
4. まとめ

実験

実験環境：実験に必要なデータの収集/加工方法

実験1：事象(カテゴリ)を説明するために適切な観点(トピック)数の推定

- LDAのトピック数をPerplexity と Coherenceの値を用いて検証

実験2：ユーザが入力したニュース記事に対する結果の検証

- カテゴリ内の観点(トピック)と入力した記事の可視化結果の検証

実験環境：実験に必要なデータの収集/加工方法

ニュース記事データベース構築機能を利用し

40件：地震

42件：大統領選挙

72件：日本学術会議

139件：新型コロナウイルス
（緊急事態宣言 / ワクチン）

上記4つのデータを収集・処理し，実験で用いるニュース記事データベースを構築した．
（ニュースサイトは24件選定）

Perplexity と Coherence

トピックモデルの精度評価指標として以下の2つの値が有用である.

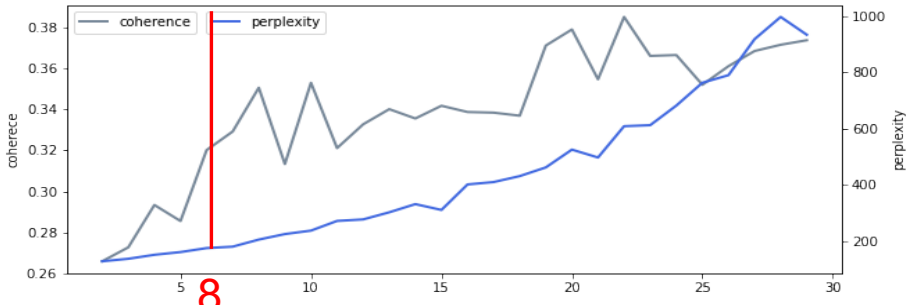
Perplexity
モデルの予測精度を測るための指標. トピックモデルを設定した環境下において,
どれほどの精度で候補となる単語群から, その場に適した単語を選択できるかを示している.

Coherence
単語間同士の類似度を基準に,「人間にとってトピックモデルがどれほどわかりやすいか」という,
曖昧で定義が難しいことを示す指標. 算出方法は複数あるが, 本研究では自然言語処理ライブラリgensimで扱っているMimnoら[*3]が提唱した手法を採用している.

[*3] : Mimno, David, et al. "Optimizing semantic coherence in topic models." *Proceedings of the 2011 conference on empirical methods in natural language processing*. 2011.

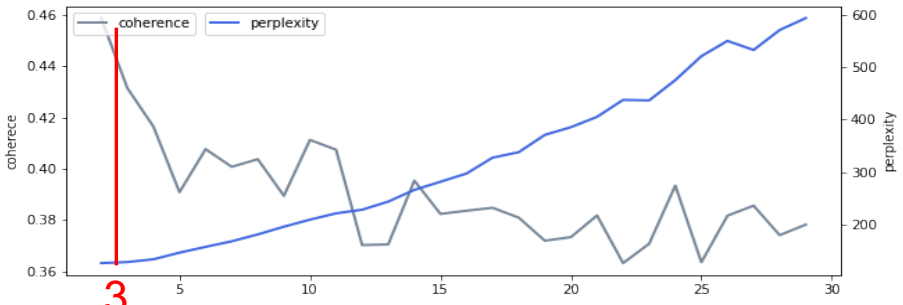
実験1： Perplexity と Coherence の値を用いた検証

横軸：トピック数 / 青線：Perplexity / 黒線：Coherence を示している.



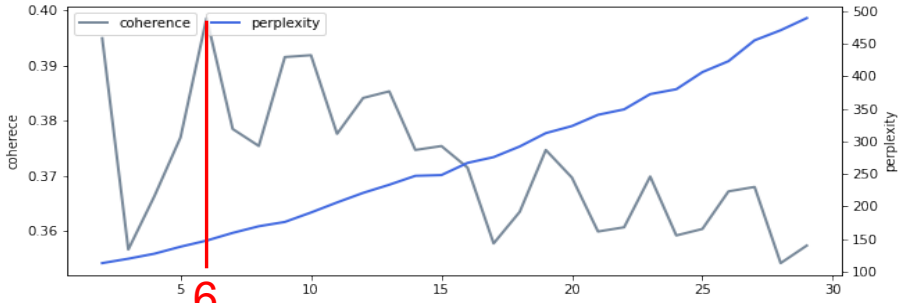
8

地震(記事40件)



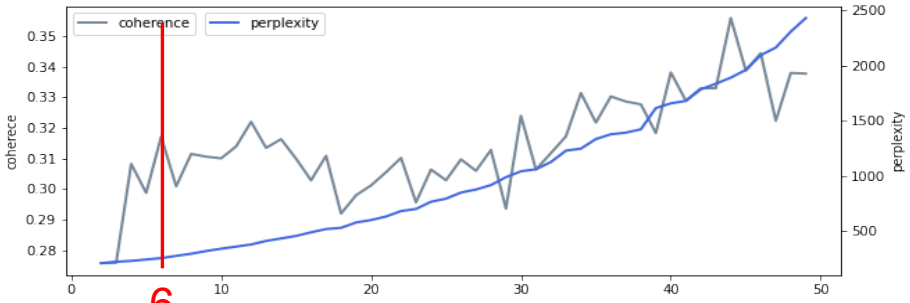
3

大統領選挙(記事42件)



6

日本学術会議(記事72件)



6

新型コロナウイルス(記事139件)

実験1： Perplexity と Coherence の値を用いた検証

実験1より，Perplexity と Coherenceの値を用いることで

ニュース記事群(事象/カテゴリ)を語る上で必要な観点の数(トピック数)の推定が可能

LDAに用いるトピック数の自動設定への適用が可能

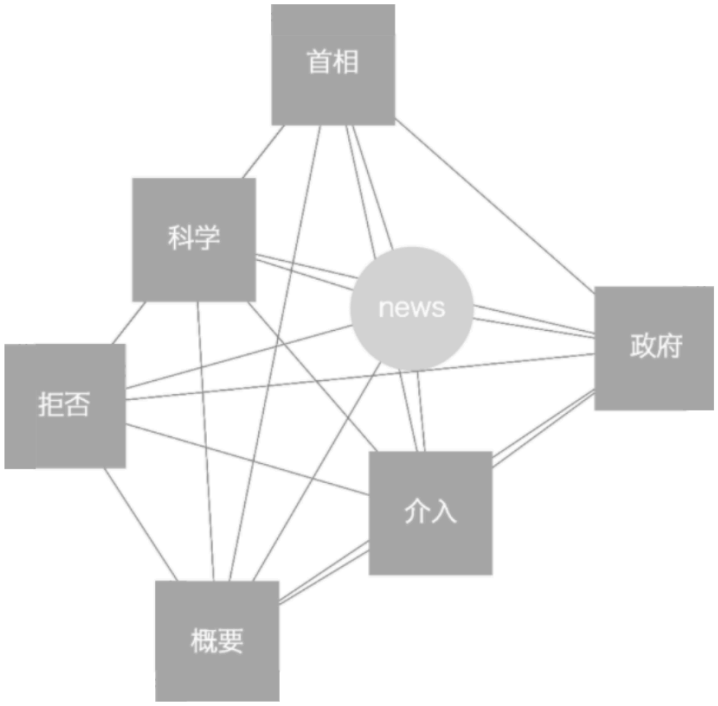
実験2:ユーザが入力したニュース記事に対する結果の検証

以下の記事を用い，観点とユーザが入力したニュース記事の類似度を可視化する.

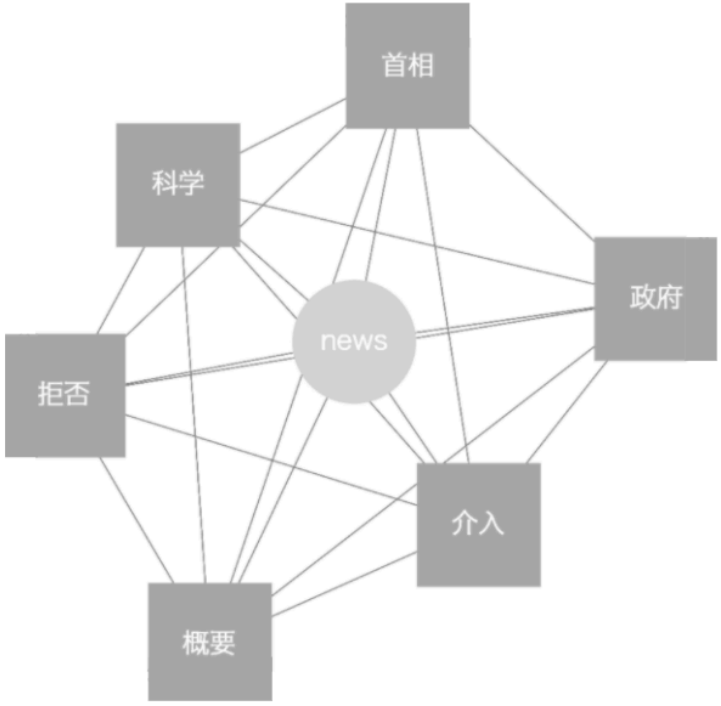
- 『首相の任命拒否「想定にない」学術会議めぐる政府文章』, 朝日新聞, 2020/10/27, (最終閲覧 2021/2/10)
- 『日本型アカデミーとしての「学術会議」に誇りを宇山智彦・北海道大教授』, 東京新聞, 2020/12/22, (最終閲覧 2021/2/10)

1つの観点に含まれる語は複数存在するが，本実験では他の観点と比較し重複のない，その観点を特徴づける語を可視化に適用した.

実験２:ニュース記事に対する可視化結果の検証



朝日新聞『首相の任命拒否「想定にない」
学術会議めぐる政府文章』



東京新聞『日本型アカデミーとしての
「学術会議」に誇りを宇山智彦・北海道大教授』

目次

1. 研究背景・目的
2. 提案方式
3. 実験
4. まとめ

まとめ

- 本研究では，ニュースを語る上で重要な観点とユーザが着目しているニュースの立ち位置の関係を可視化し，ユーザに提示することで情報の偏り(バイアス)を可視化するシステムを構築した.

- 本システムにより，ある事柄について示すニュース記事群を対象として，トピック抽出を行い，各トピックを観点として，ユーザが入力したニュース記事を位置付けることにより，どのような偏りを持っているかを示すことが可能となった.

今後の進展

- ニュース記事群の観点(トピック)数自動設定機能の実装
- 様々な分野のニュース記事に対する本方式の適用
- アンケート調査などを利用した本方式の有効性検証
- 本方式を用いた新たなニュースアプリの開発