

LDAを用いた 非文字資料検索法

～ 確率変数を非文字として見る ～

工学部 電気電子情報工学科 4年 201603734
木下研究室 五十川 潤一郎

背景

- 近年、インターネット上に大量のテキストデータが日々アップロードされている。
- 正確で早い検索を可能にするために、文書の意味をメタデータに記述する方法や、文書を意味解析し自動で分類する方法が研究されている。
- トピックモデルは文書が複数の潜在的なトピックから確率的に生成される**ベイジアンモデル**である。
- コンピュータで文書のトピックモデルを作成し文書のトピック分布を比較することにより文書間の類似度を測ることができる。

問題

- 文書を分類・整理するとき、ユーザーによって身に着けている知識や考え方が異なる為、適切に分類できなかったり、ユーザーによって分類の方法が異なったりする可能性がある。
- 大量の文書を手動で分類するのは現実的ではない。
- また、従来のカテゴリー分類では、ユーザーが必要としている文書に到達できない可能性がある。

先行研究

- **LDA** トピックモデルを用いた、カウンセリングデータの文書分類は、コンピュータで行うことによるプライバシー保護やコストの低下が可能となる。

(単 壮,加藤 昇平,電気学会論文誌 C 136 巻 2 号 p.226-232 ,Feb.2016)

- 画像から特徴量を抽出して、文書と画像から **BoW** 表現により統合した単語集合を抽出し、これにトピック分布を適用することによって文書と画像間の相互的な検索の有用性を示している。

(矢田 一貴,神奈川大学2017年度修士論文)

目的とオリジナリティ

□目的

- 文書にLDAを用いてトピックモデルを作成し文書と単語の間の類似度を測ることによって、ユーザーが閲覧した文書情報から、次にユーザーが閲覧すべき文書を提示する（非文字）システムの構築。

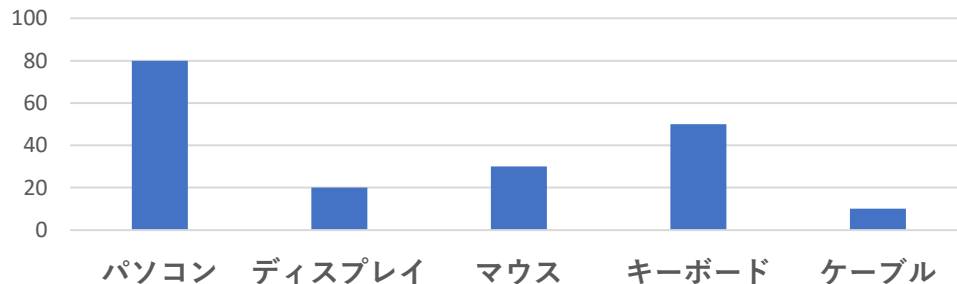
□オリジナリティ

- LDA出力として計算される潜在的確率変数（トピック）を「観測値」と見なし、ベイズの定理を適用して、逆にトピック（単語）の確率分布を算出する方式。

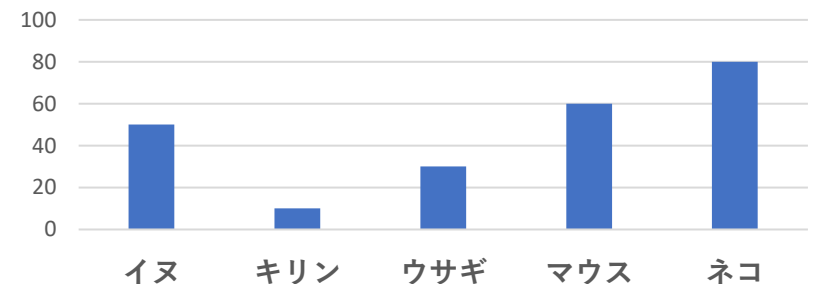
トピックモデルとは

- データが**潜在的な意味のカテゴリ(トピック)**から生成すると仮定したモデル。
- トピックモデルは“**単語の共起性がトピックを作り、トピックから文書が生成される**”という考え方

文書 1 の単語の分布



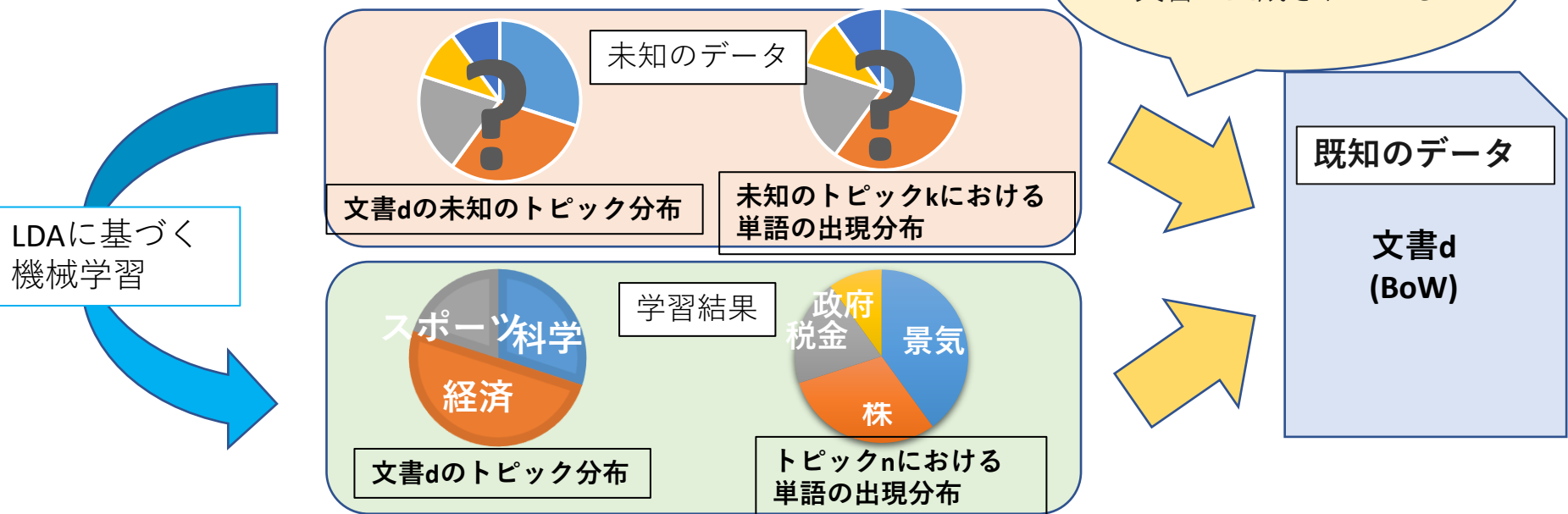
文書 2 の単語の分布



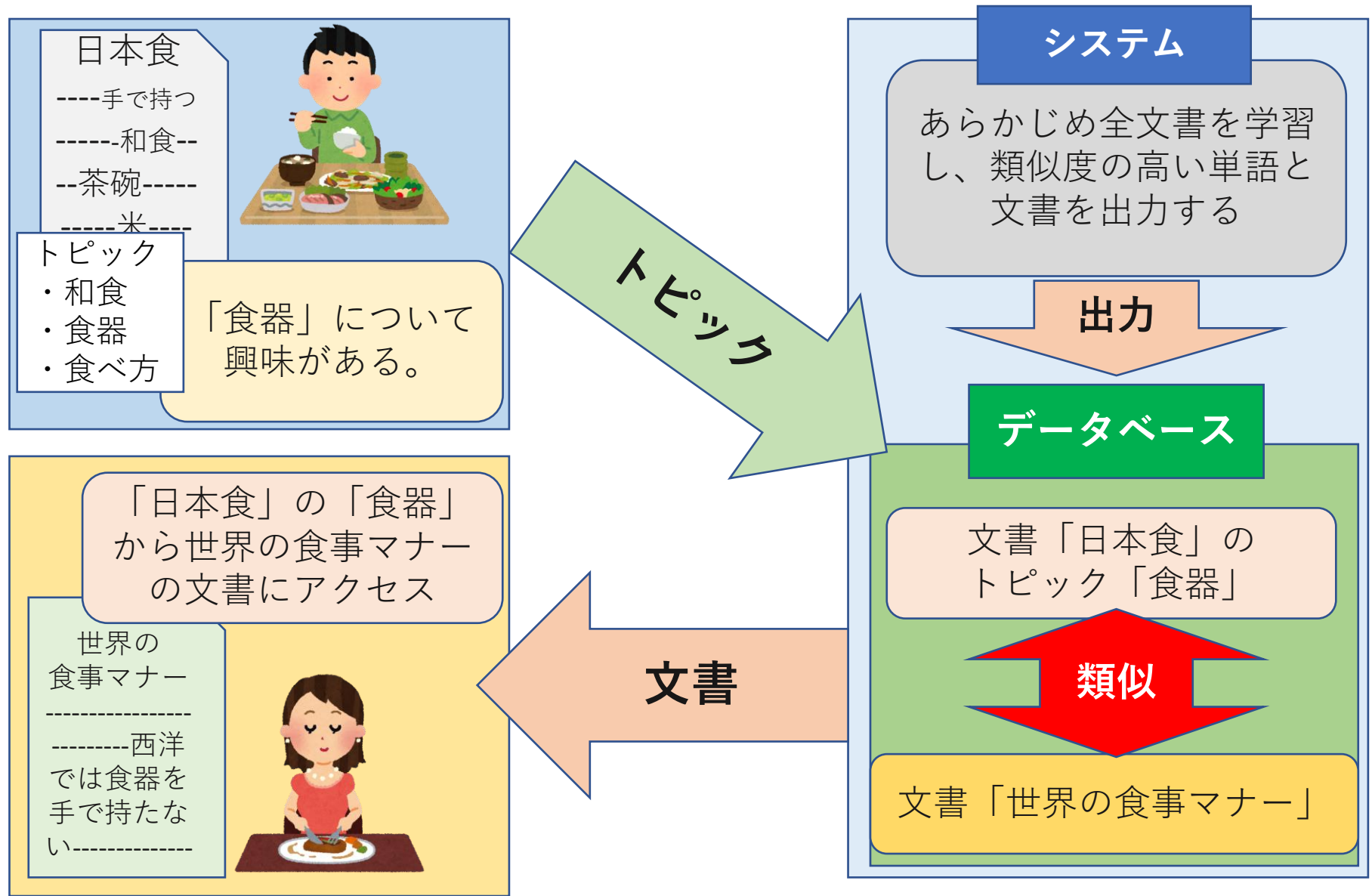
- 文書 1 と 2 では同じ“**マウス**”でも**別の意味**であると想像できる。
- 単語の意味は単語の“**共起性**”によって作られる。

LDAとは

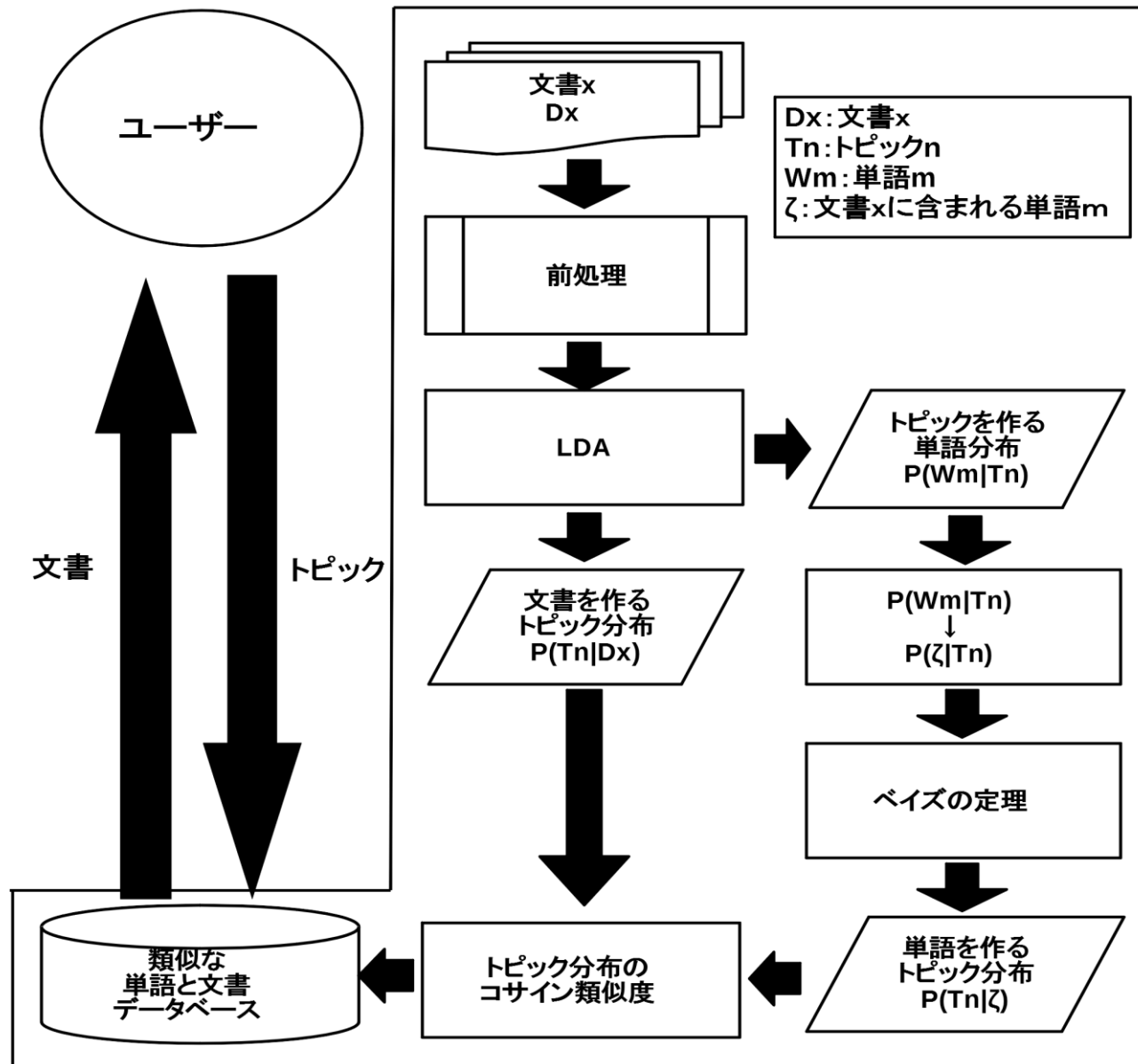
- Latent Dirichlet Allocationの略（潜在的ディリクレ配分法）
- トピックモデルの手法の一つ
- 確率的生成モデル



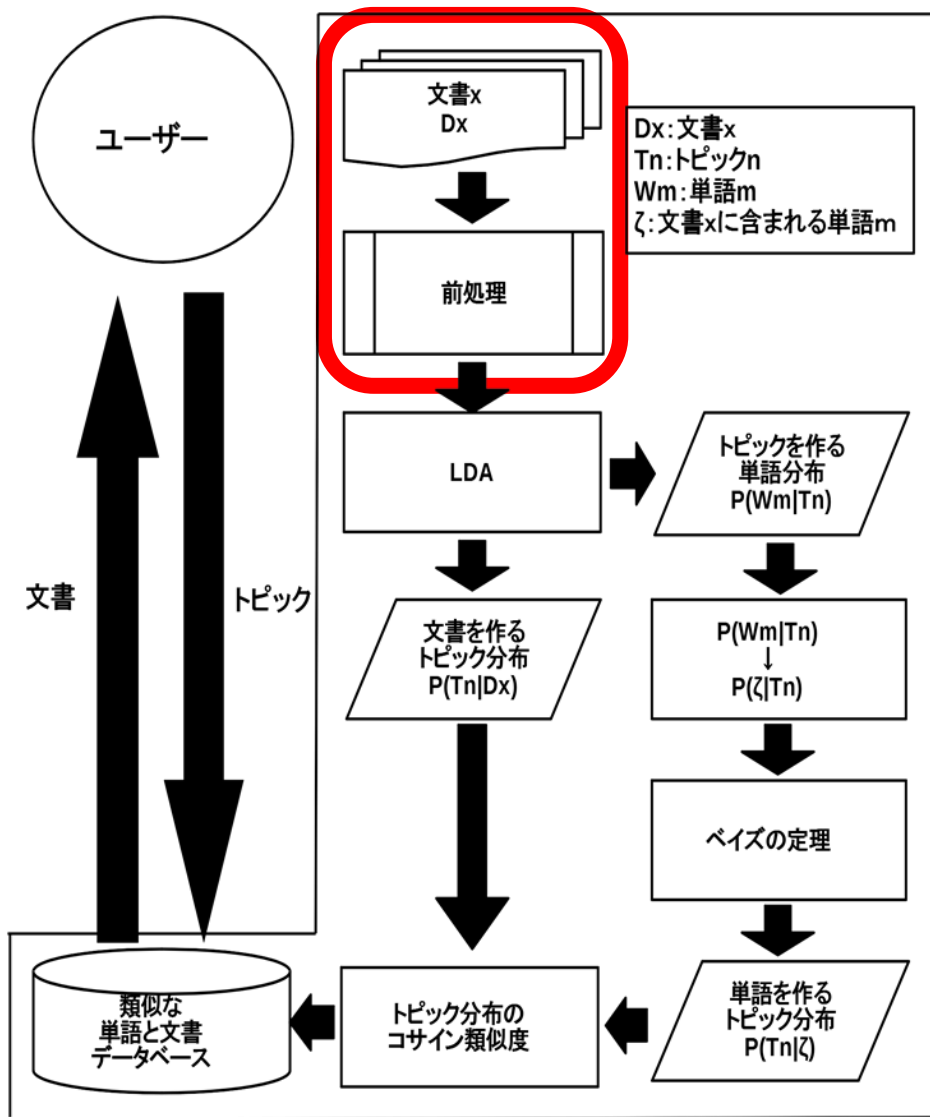
提案システム



提案手法（提案システム図）



提案手法（前処理）



MeCabを用いて文書を品詞分解し、必要な品詞のみ抜き出しDocument-term matrixを作成する。

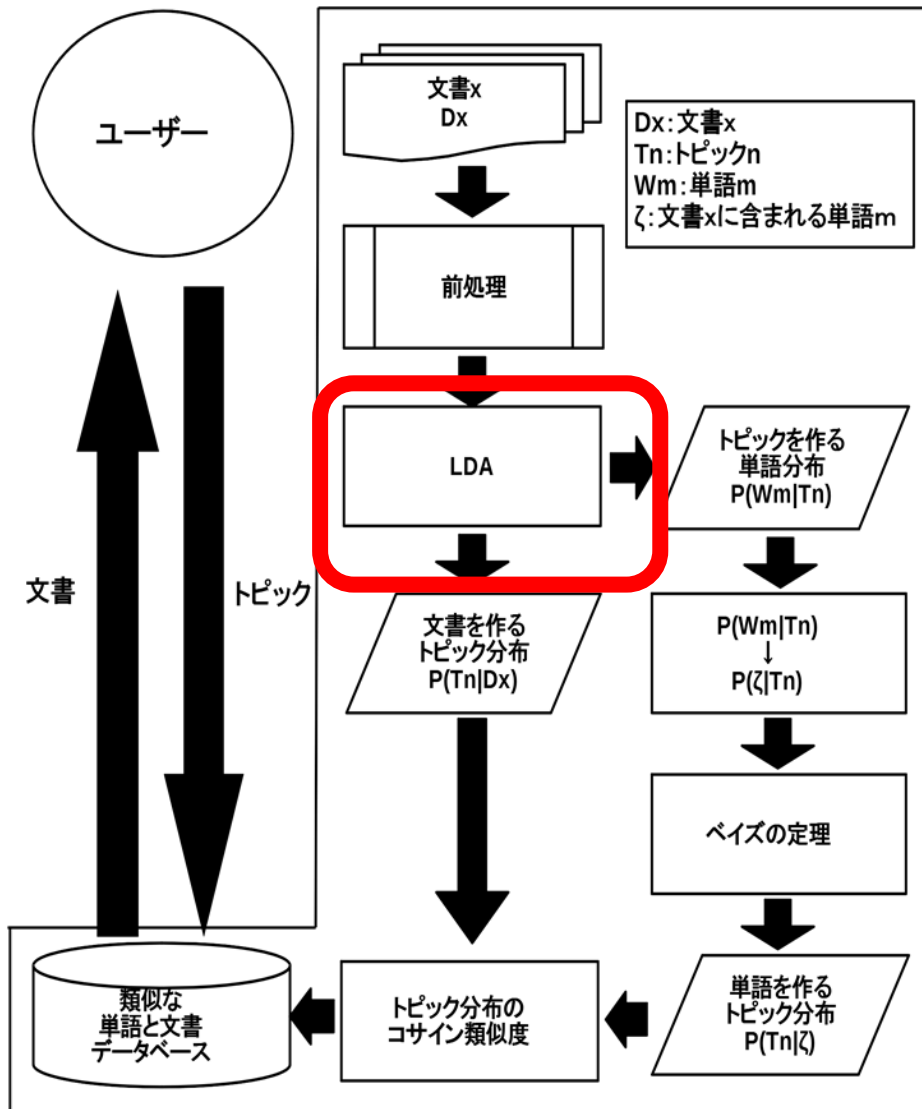
MeCab

日本語の形態素解析ツール
Pythonのパッケージ

Document-term matrix

どのドキュメントにどの単語が含まれるか、それらの単語が何回出現するかを記録したもの

提案手法 (LDA)



作成されたDocument-term matrixに LDA を適用する。

LDA の出力として文書を構成するトピック分布 $P(T_n|D_x)$ と、トピックを構成する単語分布 $P(W_m|T_n)$ が得られる。

提案手法（LDAの出力の一部）

文書を構成するトピック分布 $P(T_n | D_x)$

文書0のトピック分布

```
0: 0.00020505602878984064
1: 0.00020505602878984064
2: 0.056052036583423615
3: 0.06433329731225967
4: 0.00020505602878984064
5: 0.00020505602878984064
6: 0.00020505602878984064
7: 0.00020505602878984064
8: 0.8761287331581116
9: 0.00020505602878984064
10: 0.00020505602878984064
11: 0.00020505602878984064
12: 0.00020505602878984064
13: 0.00020505602878984064
14: 0.00020505602878984064
15: 0.00020505602878984064
16: 0.00020505602878984064
17: 0.00020505602878984064
18: 0.00020505602878984064
19: 0.00020505602878984064
```

トピックを構成する単語分布 $P(W_m | T_n)$

```
罹災: 0.010314014740288258
寄附: 0.008866041898727417
```

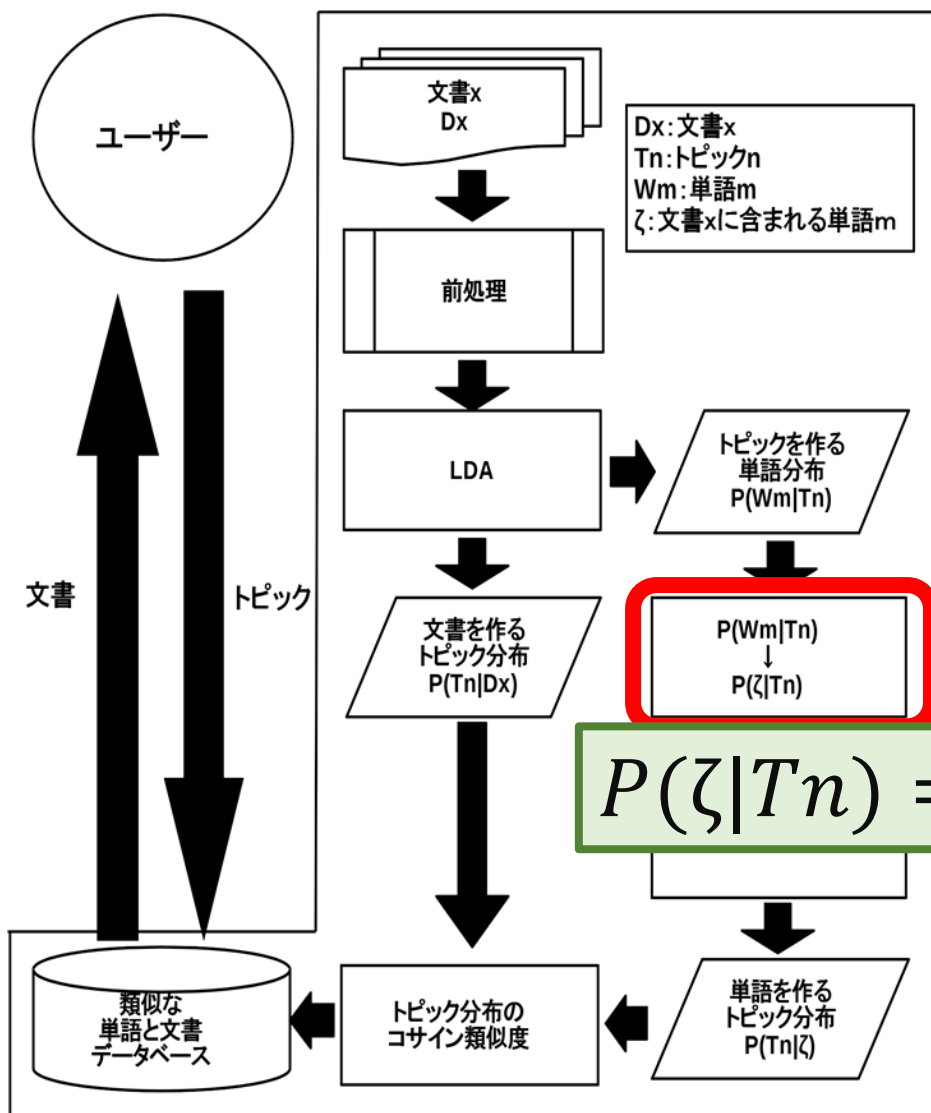
Topic # 2

```
民家: 0.031643446534872055
ポスター: 0.021653857082128525
符: 0.01781388558447361
近藤: 0.01605118438601494
印刷: 0.008671978488564491
図面: 0.008266725577414036
ムラ: 0.00761265866458416
再現: 0.007195750717073679
大里: 0.007159676402807236
上海: 0.006949722301214933
```

Topic # 3

```
民具: 0.015792734920978546
挿絵: 0.008894496597349644
只見町: 0.008153998292982578
南京: 0.0064096045680344105
国立: 0.005486362148076296
ミュージアム: 0.00504695437848568
学芸員: 0.005043998826295137
エコミュージアム: 0.004762511234730482
絵師: 0.004477082286030054
民俗資料: 0.004215035121887922
```

提案手法($P(Wm|Tn) \rightarrow P(\zeta|Tn)$)



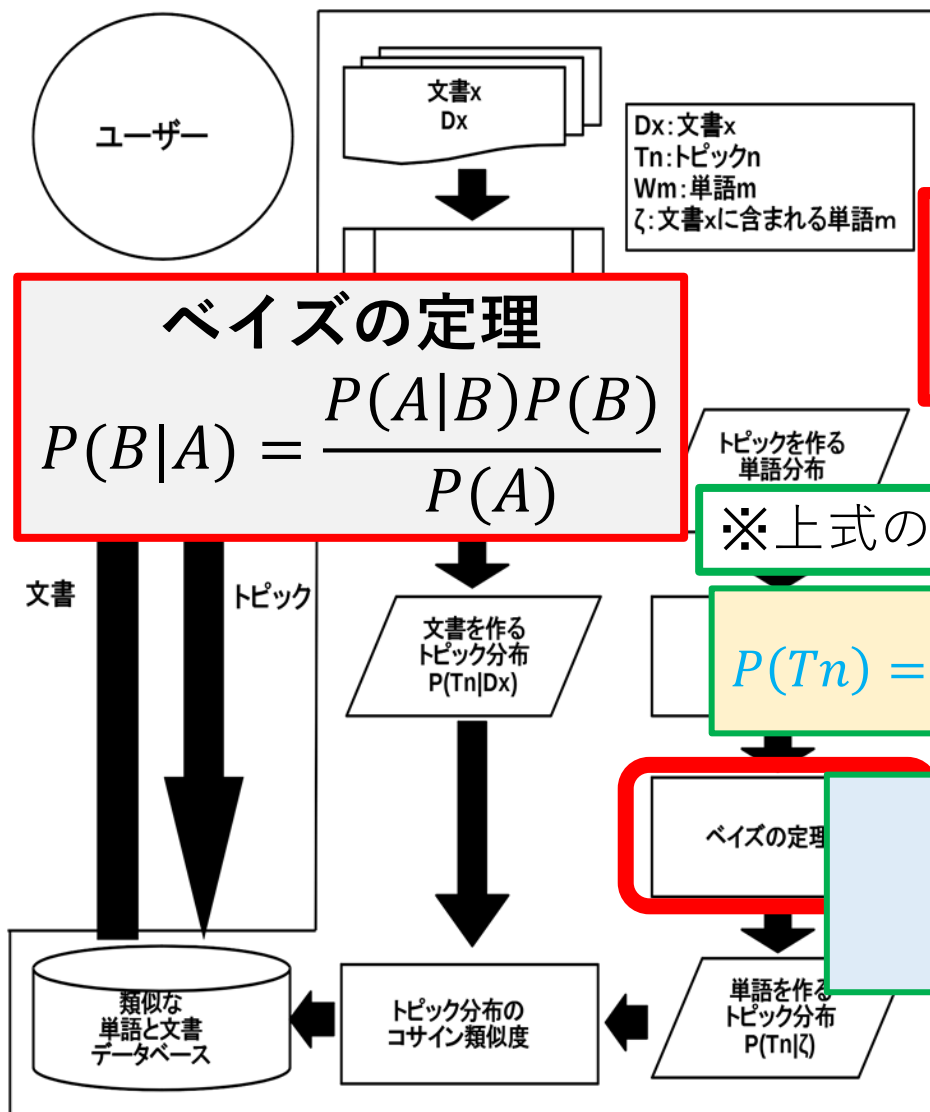
$P(Wm|Tn)$ が示す単語の確率分布は、単語がどの文書に属するか考慮されていない。

本実験では同じ単語でもどの文書に属するか判断する必要がある為”文書xに含まれる単語m”を示す ζ の確率分布を持つ $P(\zeta|Tn)$ を下式より求める。

$$P(\zeta|Tn) = P(Wm|Tn) * P(Tn|Dx)$$

ただし $P(Wm|Tn)$ と $P(Tn|Dx)$ は独立であると考える。

提案手法（ベイズの定理）



$P(\zeta|Tn)$ を $P(Tn|\zeta)$ にするためにベイズの定理を適用する。

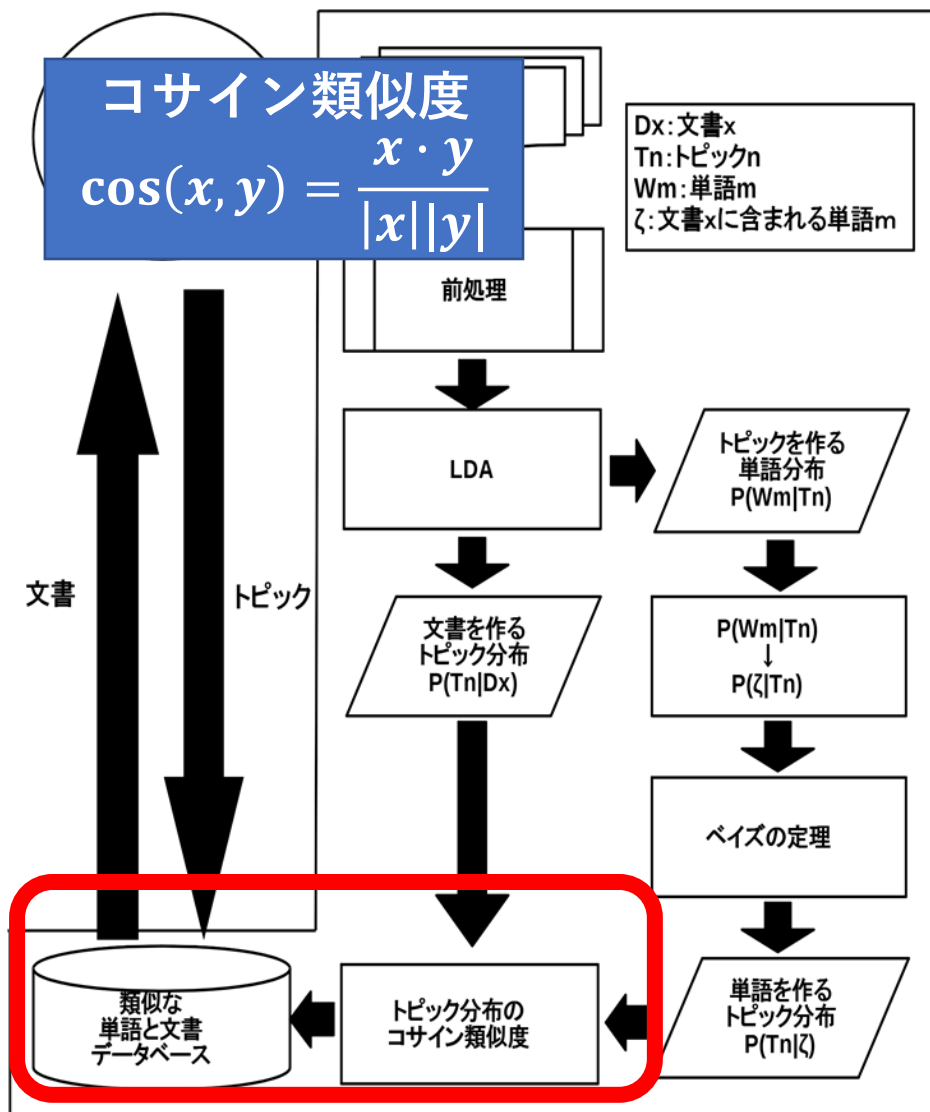
$$P\{Tn|\zeta\} = \frac{P(\zeta|Tn) * P(Tn)}{P(\zeta)}$$

※上式の未知の値は以下のようにして計算する

$$P(Tn) = \frac{\sum_{x=0}^a P(Tn|Dx)}{a} \dots (a: \text{文書の総数})$$

$$P(\zeta) = \sum_{m=0}^b P(\zeta|Tn) * P(Tn) \dots (b: \text{トピックの総数})$$

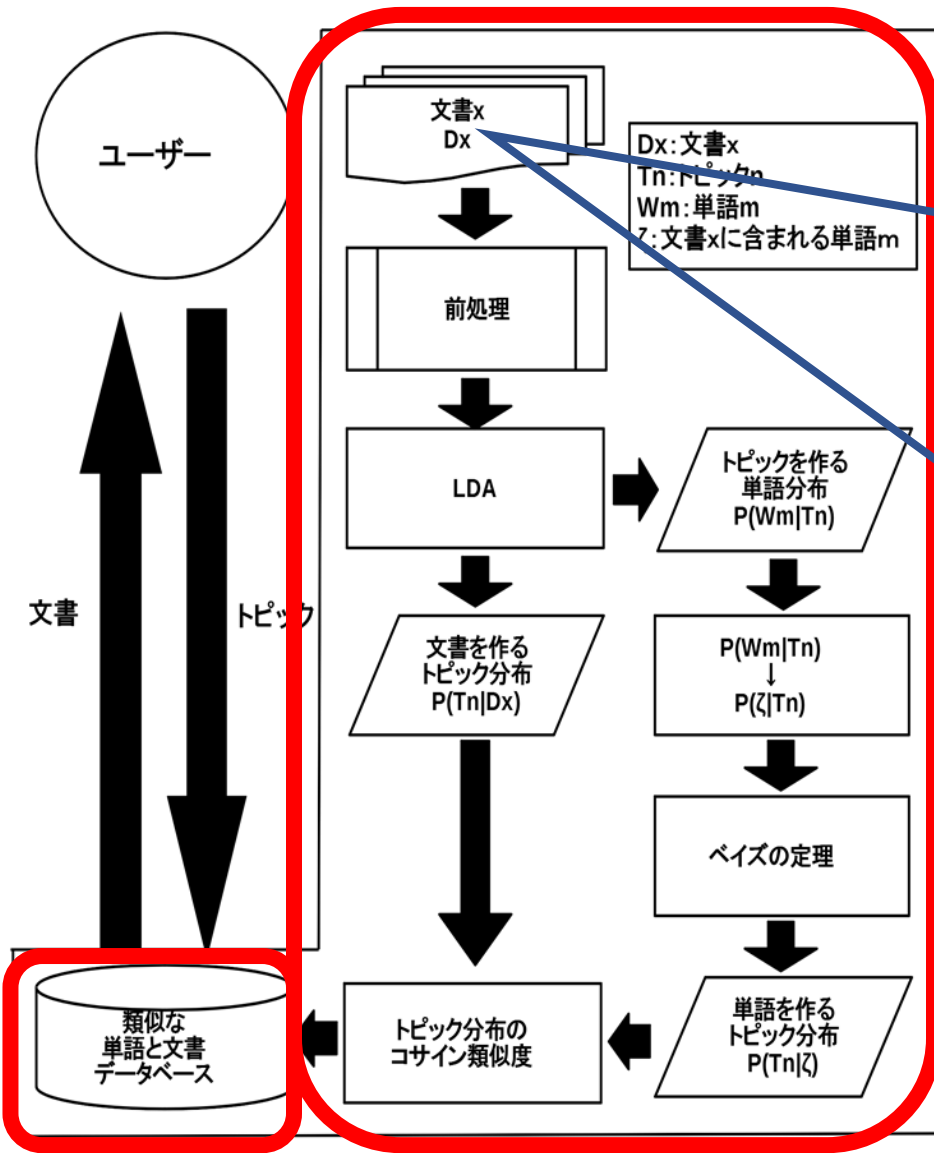
提案手法（データベース作成）



求めた $P(T_n|D_x)$ と $P(T_n|\zeta)$ から文書と単語の類似度をそれぞれのトピック分布のコサイン類似度として求める。

それぞれの単語と文書の類似度をデータベースに出力する。

実験



本実験の解析対象は神奈川大学非文字資料研究センターから刊行されている年度ごとの非文字資料に関する研究成果、共同研究の成果を収録した非文字資料研究の年報のうち日本語のもの(194件)を使用した。

今回の研究では
「中国・韓国の旧日本租界」
のトピック分析を実施する。

結果

文書の題名と含まれる単語

類似度の高い文書とその類似度

	doc_name	ic_pa	word_name	sim_doc1	ic_pa	sim_pro1
10	中国・韓国の旧日本租界…	ne…	シンボル	中国・韓国の旧日本租界 — 研究班…	ne…	0.99931…
11	中国・韓国の旧日本租界…	ne…	シンポジウム	中国・韓国の旧日本租界 — 研究班…	ne…	0.99791…
12	中国・韓国の旧日本租界…	ne…	セッション	租界研究の現状と展望	ne…	0.99954…
13	中国・韓国の旧日本租界…	ne…	タイトル	中国・韓国の旧日本租界 — 研究班…	ne…	0.99793…
14	中国・韓国の旧日本租界…	ne…	デジタル	サハリンにおける「カラフト」期の日本文…	ne…	0.99999…
15	中国・韓国の旧日本租界…	ne…	データベース	ネットワーク形成と中国非物質文化遺…	ne…	0.93692…
16	中国・韓国の旧日本租界…	ne…	ネットワーク	歴史民俗資料としてみる『風俗画報』の…	ne…	0.97077…
17	中国・韓国の旧日本租界…	ne…	プロパガンダ	中国・韓国の旧日本租界 — 研究班…	ne…	0.99777…
18	中国・韓国の旧日本租界…	ne…	ホテル	租界研究の現状と展望	ne…	0.99985…
19	中国・韓国の旧日本租界…	ne…	ポスター	近藤恒弘氏に天津日本租界での体験…	ne…	0.95593…
20	中国・韓国の旧日本租界…	ne…	上海	租界研究の現状と展望	ne…	0.99936…
21	中国・韓国の旧日本租界…	ne…	中国人	租界研究の現状と展望	ne…	0.99836…
22	中国・韓国の旧日本租界…	ne…	中国語	租界研究の現状と展望	ne…	0.99988…
23	中国・韓国の旧日本租界…	ne…	中華民国	租界研究の現状と展望	ne…	0.99843…
24	中国・韓国の旧日本租界…	ne…	仁川	租界研究の現状と展望	ne…	1.0
25	中国・韓国の旧日本租界…	ne…	企業	租界研究の現状と展望	ne…	0.99747…

考察

今回の研究では「中国・韓国の旧日本租界」の文書の分析を実施した。

1. トピック分析と人の解釈（主観評価）について：

- ・「中国・韓国の旧日本租界」の文書に含まれる単語の多くは「租界研究の現状と展望」と類似しているという分析結果となった。
- ・実際に双方の文書の全文を読んで比較すると類似な文書であることが確認できた。
- ・よって文書を構成するトピック分布と人間の解釈が一致したと見なせる。

2. 単語が複数の意味に類別される（非文字としての）可能性について：

- ・同じ単語でも、異なる文書に類別されている。したがって単語の属するトピックを考慮した結果が出力されていると考えられる。

3. 出現数の少ない単語のトピックについて：

- ・全文書で出現頻度の多い「中国」をトピックとする単語は同様の文書を類似なものとして出力しているが、出現数の少ない単語は、主観評価とは異なるトピックに分類されてしまう。
- ・そのため、出現数の少ない単語：「デジタル」「データベース」「ネットワーク」のような同一トピックを構成していると考えられるものでも出力される類似な文書が異なっている。
- ・これは学習に用いた文書の量が不足していることが原因と考えられ、出現数の少ない単語に関わる大量の文書の学習により解決される。

今後の課題

- 提案システムを客観評価するためのデータ及び精度を測る指標
- より良いトピックモデルを作成するためのテキストの前処理の最適化
- 他の類似度の高い文書を提示するシステムとの比較検討