

トピックモデルの考え方について (PLSI・LDA)



「トピックモデルによる統計的潜在意味解析」

奥村 学（監修）

佐藤 一誠（著）



「トピックモデル」

岩田 具治

目次

1. トピックモデル（潜在的意味解析とは）

2. Bug of Words

3. LDAの生成モデル

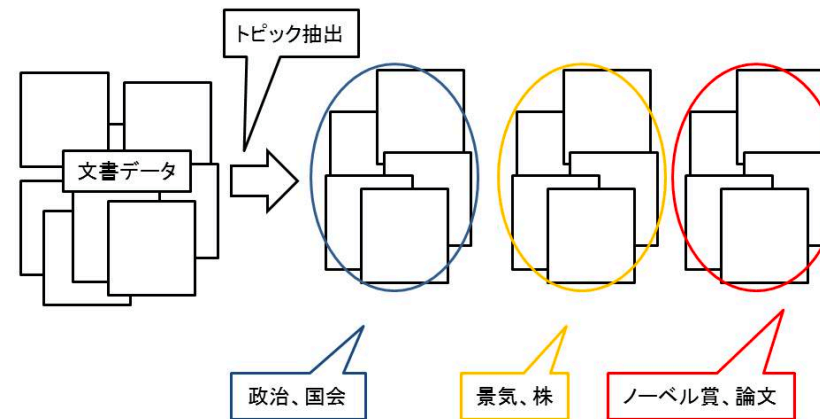
1. 多項分布・Dirichlet分布

2. PLSI

3. LDA

トピックモデルとは

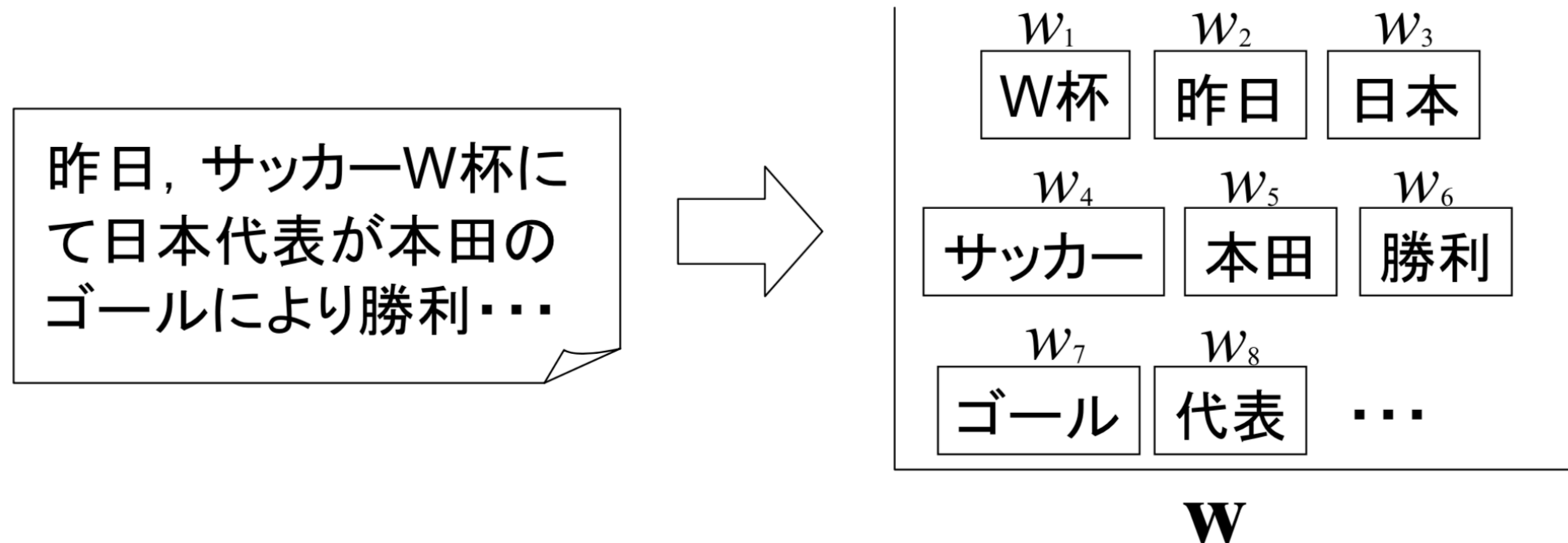
- 様々なデータに隠れた潜在的なトピックを推定するモデル
 - ▶ トピック：話題，分野などの大雑把な意味のようなもの



- このテキストは，だいたいどういう意味か？
- このテキスト群には，どんな「話題」がある？
- この単語は，だいたいどんな意味？
を人が教えることなく，データから自動で学習する

Bug of Words

文章データをどのように扱うか？



- 単語と出現頻度のペアの集合をモデル化する
- 単語が共起している現象を表す
- 単語の順序は無視

多項分布・Dirichlet分布

- サイコロの確率を例にすると

- ▶ 多項分布

- 「サイコロをn回投げてp(1~6)の出る目の分布」

目がどういう順番で出るかのパターンの数n回目で目kができる確率

$$\text{Mult}(\mathbf{p}, n) = \frac{n!}{n_1!n_2!\cdots n_K!} \prod_k p_k^{n_k}$$

- テキストに置き換えると「文章中でどの単語がどのくらい出ているか？」

- ▶ Dirichlet分布

- 「サイコロの目の出やすさの分布」

$$p(\mathbf{x}; \boldsymbol{\alpha}) = \frac{1}{Z} \prod_{i=1}^n x_i^{\alpha_i - 1} \quad \left(Z = \frac{\prod_{i=1}^n \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^n \alpha_i)} \right)$$

αが定まれば定数

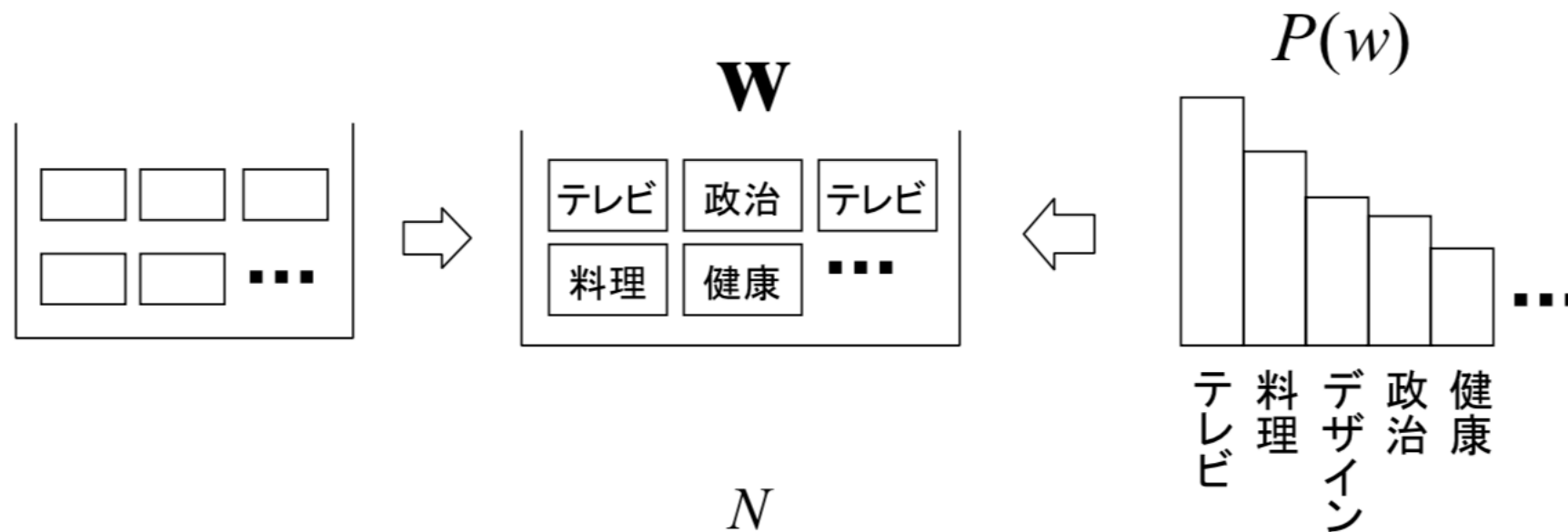
αはパラメータ
xは目の出る確率

- テキストに置き換えると「文章中にどの単語が出現しやすいか？」

生成モデルとは

- 文章データがありトピックを抽出するために、なぜ生成するためのモデルを作る必要があるのか？？
- ▶ その文章データが、ある分布から確率的に生成されたものなんだと考えようというもの

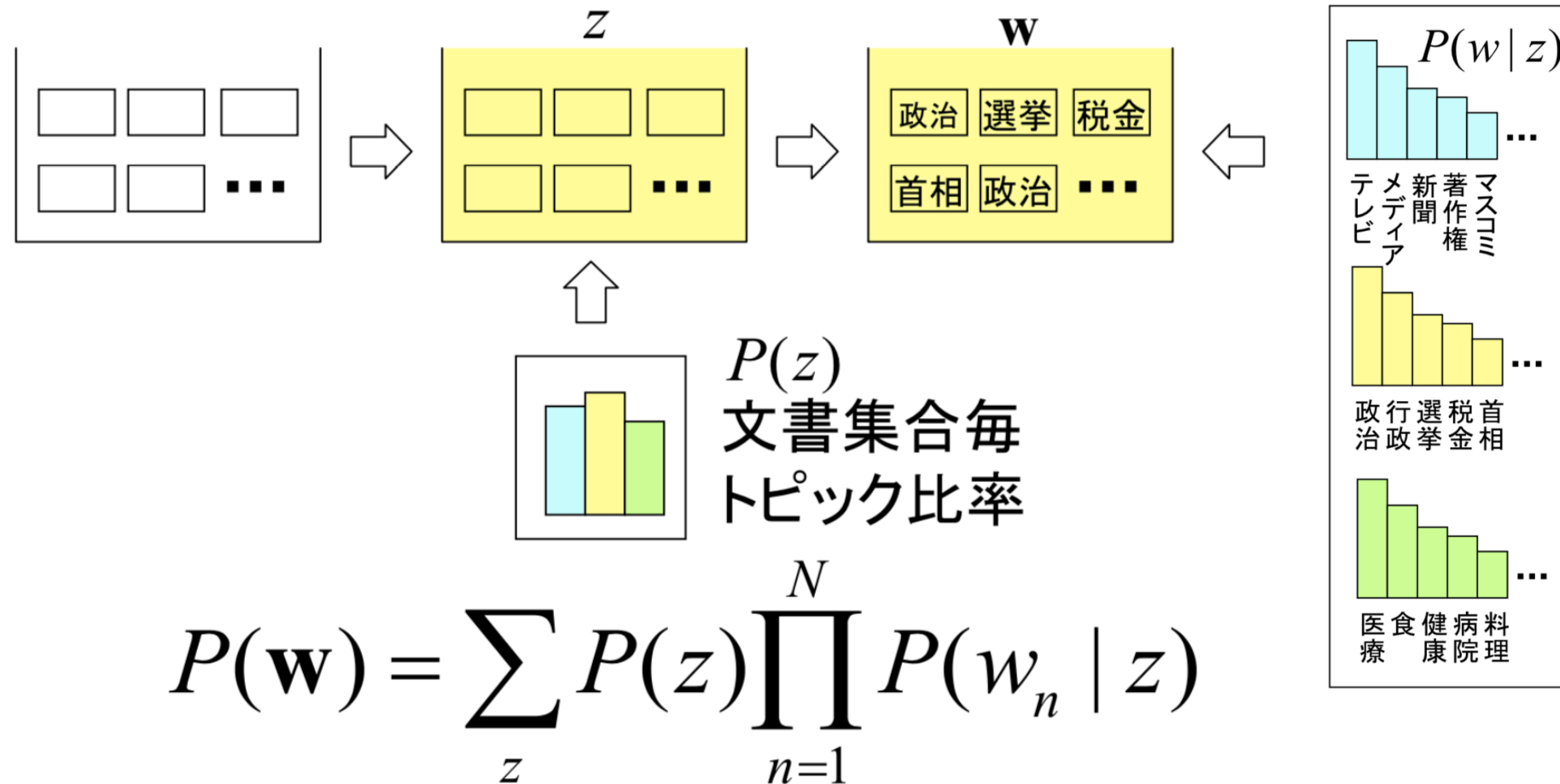
多項分布



$$P(\mathbf{w}) = \prod_{n=1}^N P(w_n)$$

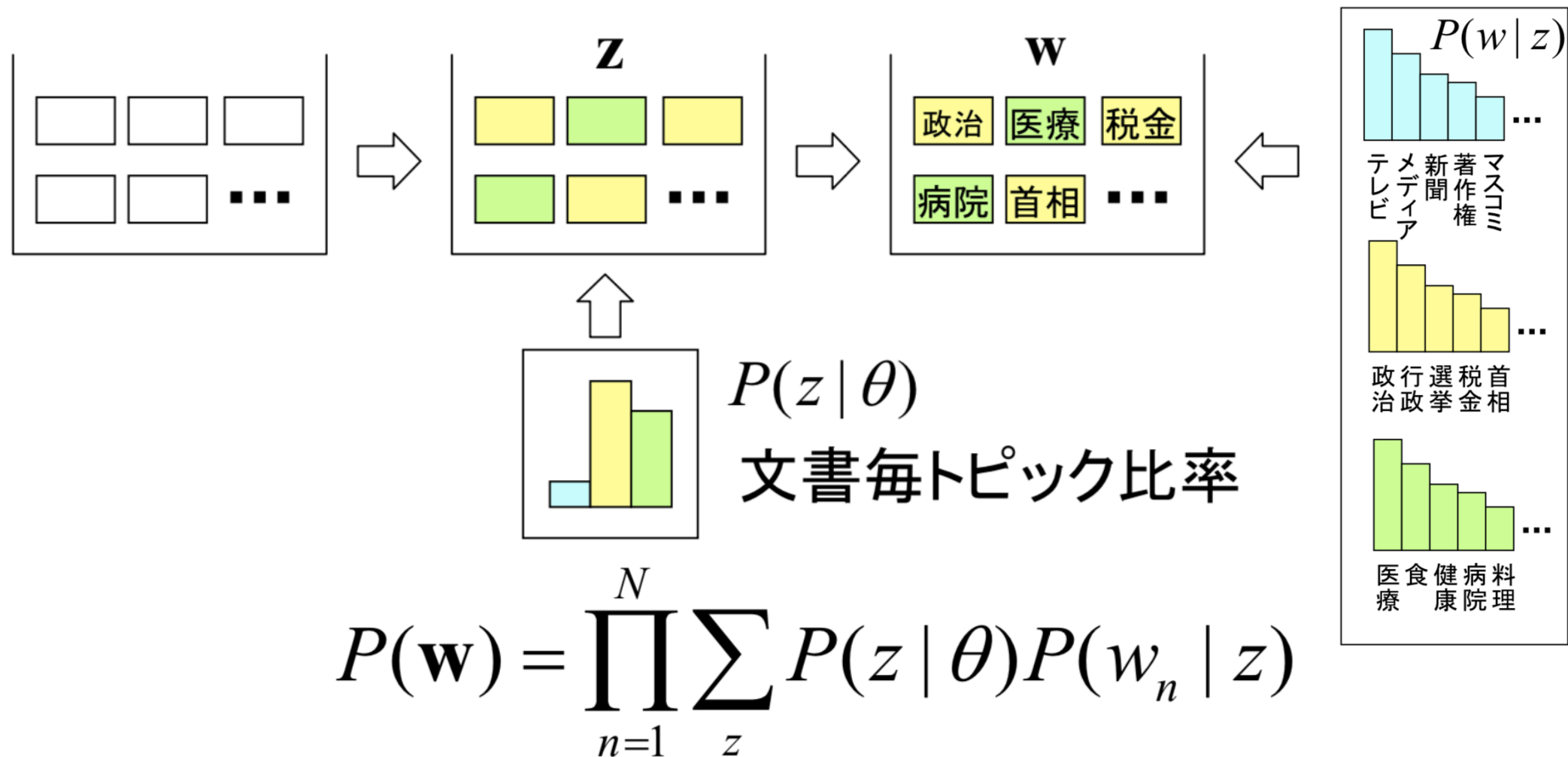
- 全文章の単語が同一の分布から生成

混合多項分布



- 文章毎にトピックを推定
- 1 文章の単語が 1 つの分布からを生成

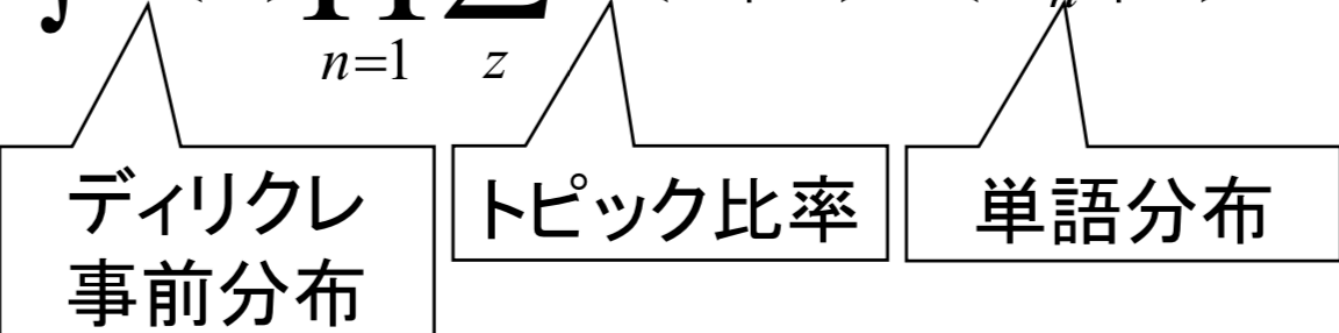
PLSI



- 単語毎にトピックを選択する
- 1 文書の単語が複数の分布から生成される

LDA

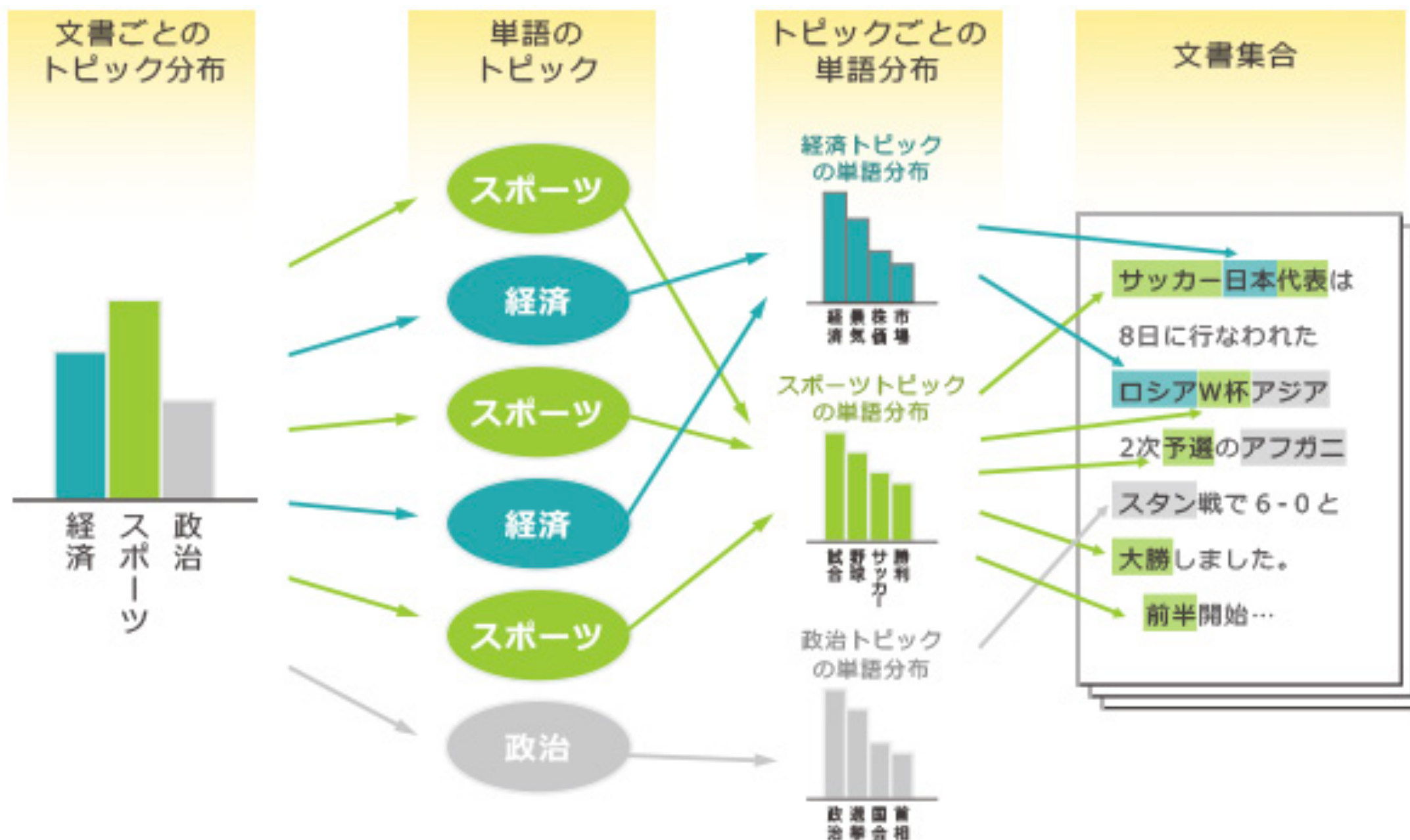
PLSI+トピック比率に対してディリクレ分布を仮定

$$P(\mathbf{w}) = \int P(\theta) \prod_{n=1}^N \sum_z P(z | \theta) P(w_n | z) d\theta$$


ディリクレ事前分布 トピック比率 単語分布

- トピックでの単語分布（多項分布）にディリクレ分布（どの単語が出現しやすいか）を考慮する
- PLSIでは学習データにあった単語の組み合わせしか表現できない
 - ▶ LDAでは新しい単語の組み合わせの文章を生成できる

LDAのイメージ図



学習過程の紹介

- 最尤推定
 - 変分ベイズ推定
 - ギブスサンプリング
- などの手法を使ってモデルを推定するようです.