

ノンパラメトリックベイズ言語モデルによる コーパス内トピック抽出

CLML-HDP-LDAパッケージを用いたコーパス分析例

2010年6月22日

知識工学部 阿部裕介

概要

1. トピックとは？
2. ベイズ統計言語モデル
3. コーパスからのトピック抽出例
4. 問題点

トピックとは？

文書集合（コーパス）の単語出現頻度は
時期・分野・地域…etcの影響を受けて変動する

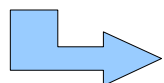
（例）円高…経済記事では出現頻度が高いが
芸能記事では少ない

餅…1月の新聞記事では出現頻度が高いが
8月の記事では少ない

トピックとは？ (2)

トピック = 単語出現頻度を変動させる
(潜在的な) 要因

問題：コーパスのみが与えられている状況で、各文書の単語出現頻度から、そのコーパスに潜在しているトピックにしたがって、コーパス内の各単語を分類・クラスタリングし、その結果からコーパス内のトピックに関する知見を得ることができないか？



近年、画期的な手法（ノンパラメトリックベイズ言語モデル）が提案された

ベイズ統計言語モデル

まずはパラメトリックベイズ言語モデルから

→ LDA = Latent Dirichlet Allocation

コーパス内の文書は、 $\alpha_1, \alpha_2, \dots, \alpha_k$ 個の潜在的トピックからなり、文書内の単語 w_j は各トピック α_i から定まる確率分布にしたがって確率 β_{ij} で出現する

上記の枠組みに基づく生成モデルがLDA

※パラメータ α, β の推定はベイズ推定で行う

どこがパラメトリックか？

LDA : $\alpha_1, \alpha_2, \dots, \alpha_k$ 個のトピック



パラメータ数 k が 事前に固定されている のがパラメトリック

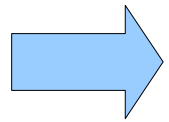
これに対して、

事前にパラメータ数 k を固定しない のがノンパラメトリックベイズ

※機械学習の分野においては、適切な k の値を事前に決めることは、昔からある「非常に」悩ましい問題

HDP-LDA

ところが
適切な k の値(=コーパス内に潜在的に存在するトピック数)を
自動的に決定できる画期的な新手法が提案された

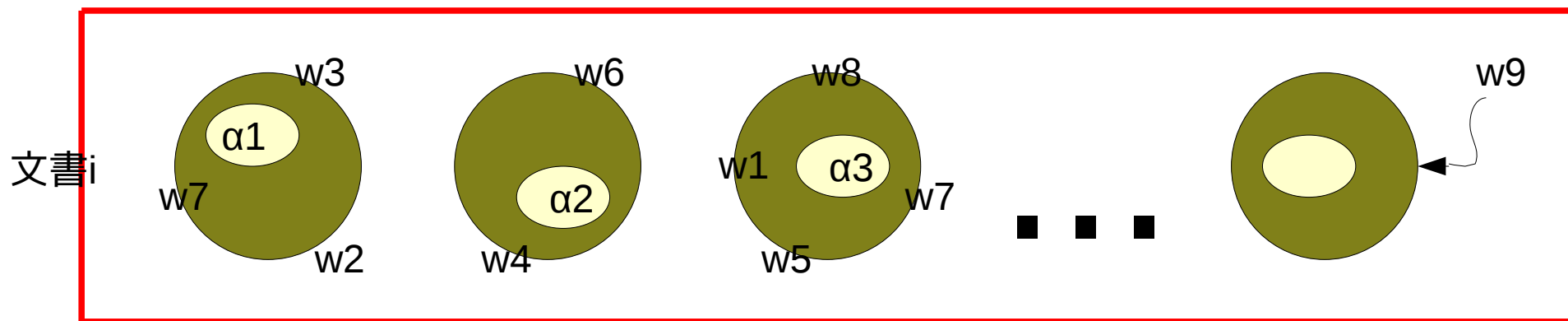


ノンパラメトリックベイズ言語モデル
HDP-LDA

Dirichlet Processで自動的にkを決める仕組み

Chinese Restaurant Process

「中国人は混んでいるテーブルを好む」



コーパス内の各文書が各中華料理フランチャイズ店に相当
各店で、客(単語)は基本的には混んでいるテーブルに着座して
どの店でも同じ料理を食べるが、たまに新しい料理が追加される

HDP-LDAの適用例

(スポーツ・コーパスデータからのトピック抽出事例)

東京読売のスポーツ記事100本を
形態素解析をして1202語を抽出し、
HDP-LDAで各トピックを代表する
上位10単語を選出した

※ トピック数はおおむね40前後

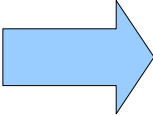
トピック抽出例

("Topic 9"

#("日立" "友" "茨城" "水戸" "鈴木" "茨城大" "クラブ"

"キリスト" "日立建機" "常陸") . 0.024772166440797362)

この数値はコーパスにおけるTopic9の強さ

 「茨城方面の地域」がトピックとして抽出されている、と推測できる

他にも…

("Topic 8"

#("浜松" "静岡" "西" "北" "南" "学園" "中央" "市立" "加藤"
"沼津") . 0.02553611830885679)



トピック:「静岡」

("Topic 10"

#("青森" "八戸" "弘前" "十和田" "黒石" "決定" "山田" "沢田"
"工藤" "三上") . 0.022504066016149956)



トピック:「青森」

「競技種目」としてのトピック例

("Topic 1"

#("少年" "選手" "スピード" "アイスホッケー" "群馬" "国体"
"スケート" "成年" "優勝" "フィギュア").

0.03800044182239797)



トピック：「スケート」

("Topic 38"

#("移籍" "FW" "□" "ブラジル" "サッカー" "リーグ" "発表"
"今季" "千葉" "出場"). 0.022484002817775697)



トピック：「サッカー」

(参考) 政治記事コーパスへの適用例

("Topic 12"

#("ハマス" "両派" "ファタハ" "交渉" "ガザ" "衝突" "議長"
"内閣" "サウジ" "死亡") . 0.023061437678105236)



トピック：「パレスチナ問題」

("Topic 26"

#("大統領" "クリントン" "米" "アイオワ" "ヒラリー" "表明"
"夫" "集会" "民主党" "遊説") . 0.02191042008228612)



トピック：「アメリカ大統領」

他の機械学習手法(NMF)との比較

HDP-LDAによって抽出されたトピックは、単語出現頻度行列をそのまま用いたNMF（非負行列因子分解）による特徴抽出結果ときわめてよく一致する

NMFによる分析例

Feature 3

浜松	0.027365546299457543
静岡	0.025582627071738286
北	0.013076228573034934
西	0.010612858027338355
学園	0.008911088806579738
南	0.00877518063678776
中央	0.0077830897196918235
沼津	0.007061895907934063
加藤	0.007033139427334862
磐田	0.00616205374976534

Feature 14

青森	0.030434342146863597
弘前	0.015826703453690052
八戸	0.013605316409593613
十和田	0.010551135635793364
黒石	0.008811963467451505
安田	0.008139753787799497
むつ	0.007034090423862245
三上	0.00702951087208542
決勝	0.006970824753584393
準決勝	0.000616657209283125

HDP-LDAの問題点 (1)

HDP-LDAはトピック数 k を事前に決めずとも自動的に最適な k を推定してくれるのが特長だが、実行毎に k の値が実はけっこうバラつく

先のスポーツコーパスデータで $k = 34 \sim 49$

※それでもコーパス内の妥当なトピック数の大きさが自動的におおよそわかるのは大きい

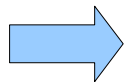
HDP-LDAの問題点 (2)

HDP-LDAならびにNMF双方についていえることだが、
得られたトピックや特徴については、人間側が適切に解釈
する必要がある

うまく意味づけられないトピックや特徴も出現する

("Topic 37"

#("山本" "戦い" "井上" "土田" "仲間" "自衛隊" "会場" "試合" "姿" "相手").
0.021383541306727056)



トピック：「???」

※全トピックを解釈するのではなく、コーパス内の幾つかの
特徴的なトピックの発見を目的とするような用途が現実的か