トピックモデルの妥当性検証

三浦大輝† 白井匡人†† 三浦孝夫†
† 法政大学理工学部創生科学科 †† 法政大学マイクロ・ナノテクノロジー研究センター

１．前書き

トピックモデルとは、どの文書も、複数のトピックがディリクレ確率分布で生成され、さらに出現する語が、トピック・文書に対応する多項分布に従って生成されると仮定したモデルである。各トピックの持つ語分布は多項分布に、トピックの分布はディリクレ分布に、各文書内の語分布はトピックごとに多項分布に従う。

2 トピックモデル

トピックモデルではトピックごとの語分布、文書ごとの語分布が多項分布に従うとする。トピックの分布はカテゴリ分布のパラメータであるとして多項分布の共役事前分布であるディリクレ分布に従うと仮定している。即ち、文書内の単語はその文書に依存しているトピックの多項分布によって生成されると仮定している。本研究ではトピックモデルの仮説をロイターコーパスにより検証し、検証が成功すればトピックモデルの信頼性が確認できる。

3. 検証手法

トピックモデルの前提条件と成立要件を検証するために、コーパスからトピックの分布、トピックごとの語分布、トピック文書ごとの語分布を抽出する。語分布の推定ではトピックごとあるいは文書ごとで異なる語分布を持つため出現頻度 0 となる単語がある。このゼロ頻度問題を防ぐため、本稿では最大事後確率(MAP)推定によって補正しておく。以下では最大事後確率推定によって算出した理論分布と実測値の分布を適合度検定で評価する。

4 実験手順

テストデータにおけるトピックの総頻度と学習データで推定したトピックのディリクレ分布からテストデータにおけるトピックの理論分布を算出する。学習データから推定した各トピックの語分布とテストデータから算出した各トピック内の語の総頻度からテストデータにおける各トピックの語の理論分布を算出する。テストデータで算出した各文書内のトピックの割合と学習データから推定したトピックごとの語分布から理論分布を算出する。理論分布と実測分布で適合度検定を行い、これを評価結果とする。

5 実験結果

表１：トピックごとの検定

トピック	χ^2	有意水準	合格/不合格
資金調達/資本	7.30E-42	10.11	合格
財政	1.32E-40	10.11	合格
国際関係	1.76E-49	10.11	合格

表２：文書ごとの語分布の検定

	χ^2	有意水準	合格/不合格
文書1	317.9	10.11	不合格
文書2	110.42	10.11	不合格
文書3	763.04	10.11	不合格

6. まとめ

トピックの分布とトピックごとの語分布の検定において高精度で合格した。文書ごとの語分布においては精度がかなり低く不合格であった。

参考文献

[1]MLP 機械学習プロフェッショナルシリーズ-トピックモデル-岩田具治