1D-4

Web ニュースからの LDA に基づく興味学習手法

東原　智幸†　　渥美　雅保†

創価大学　大学院　工学研究科　情報システム工学専攻†

1. はじめに

文書の生成過程を確率的にモデル化したトピックモデルが，文書分類，推薦システムなど，さまざまな分野に適用されている．代表的な手法である LDA[1]では，文書を複数のトピックの混合により表現し，その混合確率を単語頻度データより学習している．本論文では， WEB ニュースに対するユーザの興味応答から，興味有集合，興味無集合に分割し，各集合に対して LDA を適用し，興味有トピック，無トピック間の距離からユーザの興味をモデル化する手法を提案する．そして，LDA に入力される特徴の違いよる興味の学習と推論への影響について評価する．

2. LDA に基づく興味学習

2.1. Latent Dirichlet Allocation(LDA)

LDA は， コーパスの確率的な生成モデルであり，基本的なアイデアは，k 個の潜在トピック z の混合分布として表現される．各トピックは，単語の分布により特徴づけられる．

LDA では，各ドキュメントの単語集合 w に対して，コーパス D でつぎのような生成モデルを仮定する．コーパス D は，M 個のドキュメントの単語集合である．

1) ドキュメントの単語数 Nの選択
2) θ ～ディリクレ分布(α)選択
3) 各単語 w_nについて
   (ア)トピック z_n～多項分布(θ)の選択
   (イ)トピック z_n で条件づけられた多項確率 $p(w_n | z_n, \beta)$により単語 w_n を選択

この生成過程のグラフィカルモデルは図 1 のようになる．

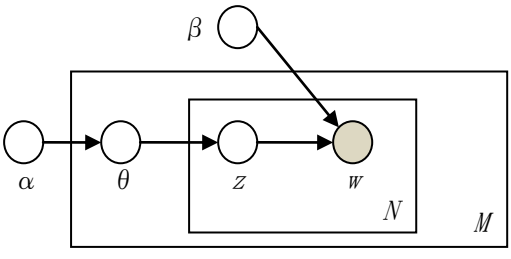

図 1　LDA のグラフィカルモデル

これらの生成過程と変分ベイズ法を用いてトピック混合分布 θ と各トピックにおける単語分布 $\beta = \{\beta^1, ..., \beta^k, ..., \beta^K\}$ が求まる．求めた θ と β を用いて新規文書に対するトピック混合分布 θ^{new} を推論することができる．

2.2. LDA による興味の学習と推論

本研究では興味あり，興味なしニュースそれぞれに対して LDA を適用し，以下に示す手順に従って興味の学習と推論を行う．

●学習部

1) 興味あり文書集合を用いて LDA を適用し，トピック混合分布 θ_{pos} と単語分布 β_{pos} を求める
2) 興味なし文書集合についても同様の操作を行い θ_{neg}, β_{neg} を求める

●推論部

1) 新規文書の単語頻度情報を作成
2) 単語頻度情報と興味ありデータからの学習分布を用いて LDA の推論計算により新規文書の θ_{pos}^{new} を求め，新規文書の平均単語分布 β_{pos}^{new} を以下の式により求める

$$\beta_{pos}^{new} = \sum_K \theta_{pos}^{new,k} \cdot \beta_{pos}^{k}$$

3) 同様に興味なしデータからの学習分布から θ_{neg}^{new} を求め

$$\beta_{neg}^{new} = \sum_K \theta_{neg}^{new,k} \cdot \beta_{neg}^{k}$$

を計算

4) β_{pos}^{new} と各トピックの単語分布 β_{pos}^{k} 間の距離を Hellinger Distance により計算し，最短距離 $dist_{pos}$ を求める
5) 同様に興味なし β_{neg}^{new}, β_{neg}^{k} 用いて最短距離 $dist_{neg}$ を求める
6) 興味の判定

判定法 1 または 2 を用いて興味の判定を行う．

判定法 1：興味あり最短距離 $dist_{pos}$ のみを用いて，$dist_{pos}$ が閾値より小さいの場合に興味ありと判定する．閾値は，0.2，0.4，0.6，0.8，1.0 の５つで判定する．

判定法 2：興味ありトピックに近く，興味なしトピックからは遠いとき，興味ニュースであるという基準を用いて，$dist_{pos} < dist_{neg}$ の場合は興味あり，条件を満たさない場合には，興味なしと判定する．

3. 実験

3.1. ニュース評価システム

ニュース評価システム(図 2)を作成し，2010

LDA-based Method for Learning User's Interests from Web News
†Tomoyuki Higashihara, Masayasu Atsumi, Division of Information Systems Science, Graduate School of Engineering, Soka University

/12/16～2011/01/05/の期間，5 名のユーザにニュース評価を行ってもらった．ユーザは，システムによって提示されたニュースリストから，閲覧したいニュースを選択し，ニュース本文を閲覧する．システムは，ニュースの更新，画面上の表示の切り替えや，評価履歴保存を行う．また，ユーザが閲覧したニュースを興味ありニュース，それ以外を興味なしと判定している．表 1にユーザごとの評価ニュース総数と興味ありニュース数を掲載する．

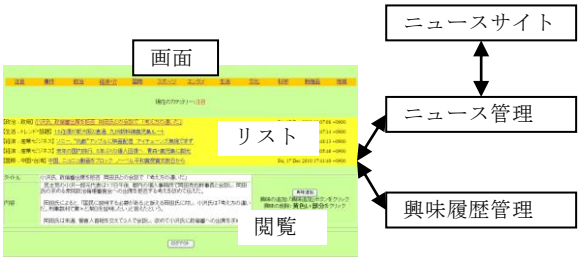

図 2　評価システム概要

表 1　各ユーザの評価総数と興味あり数

ユーザ	A	B	C	D	E
興味あり	8	58	19	53	49
総数	436	233	224	243	410

3.2.　実験データ

ユーザごとに得られた評価データを２つの集合に分割し，１つを学習用，一方を推論用として使用する．分割した集合を，興味ありニュース集合，興味なしニュース集合に分ける．興味なしニュース数は，興味なしデータが増加した場合の振舞について評価を行うため，興味ありニュース数の 1 倍，2 倍の集合を用意する．用意された各集合に含まれるニュースタイトルと本文を形態素解析し，単語 ID と頻度を求め，実験データを用意した．

3.2.1. Supervised-LDA(SLDA)との比較

各ユーザの実験データ対して提案手法と SLDA[2]を適用し，推論精度比較を行った．SLDA は，文章の生成モデルは LDA と同じであり，文章のクラスの経験分布と回帰係数 η の内積を平均とした正規分布から文書に対する応答を生成すると仮定している．

図 3は，SLDA，判定法 1 ，判定法 2 について，ユーザ 5 人の平均を求めたグラフである．横軸は，False Positive(FP)率で興味なしニュースを興味ありと判定した比率，縦軸は，True Positive(TP)率で，興味ありニュースを興味ありと判定した比率を表す．グラフ中のラベル 1，2 は興味なしデータ数の倍率を表し，1.0，0.8 は判定法 1 での閾値を表す．SLDA は，興味なしデータ数を 2 倍にすると，TP，FP 率ともに下がる傾向にある．これは，興味なしデータが増加することで SLDA 内の興味なし経験分布が平均化されたためではないかと考察される．判定法 2 は，興味なしデータ数が増加しても TP 率は，上昇する傾向にある．また，判定法 1 よりも，TP 率が高くなることから，興味なしデータを用いることが興味判定手法に有効であることが分かる．

図 4は，ユーザ C，D の TP，FP 率のグラフである．ユーザ C は，興味なしデータ数を 2 倍にすると，TP 率，FP 率ともに上昇する傾向があり，他 2 名でも同様の傾向が確認された．ユーザ D では，TP 率の上昇，FP 率の下降が確認され，他 1 名のユーザでも同様の傾向が確認された．

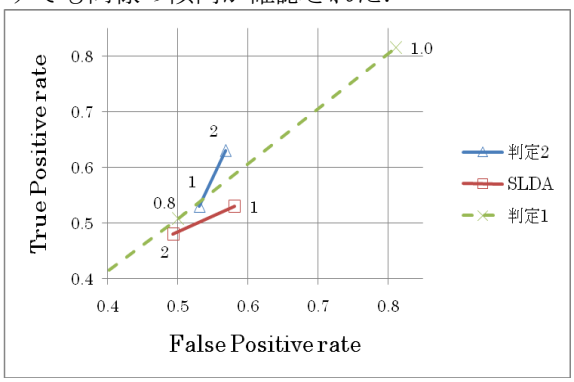

図 3　5 人の平均

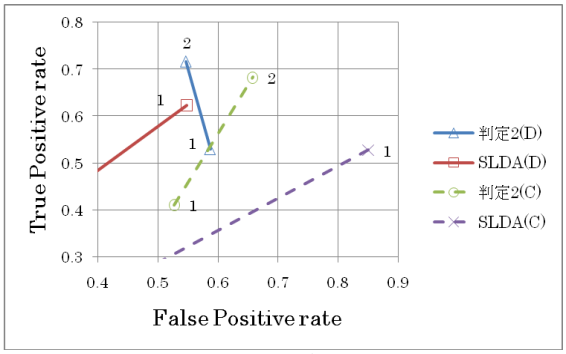

図 4　ユーザ C，D の例

4.　まとめ

LDA を興味ありデータ，興味なしデータそれぞれに適用し，興味の学習と推定を行う手法を提案した．また，実験により興味なしデータを用いることが精度向上に有効であることがわかった．今後の課題としては，長期的な評価データを用いた場合の精度比較，単純な単語の組み合わせだけではなく，格フレームのような意味的組み合わせを考慮した興味学習の手法の開発があげられる．

参考文献

[1] Blei: Latent Dirichlet Allocation, JMLR, 3:993-1022, 2003.
[2] David Blei and Jon McAuliffe, Supervised topic models, *NIPS*, 2007.