E-015

トピックモデルに基づく協調フィルタリングによる文書推薦手法について

山本　祐生† 三川 健太†† 後藤 正幸†
Yamamoto Yusei Kenta Mikawa Masayuki Goto

1　研究背景・目的

近年，蓄積された膨大な電子文書データから，ユーザの興味に応じて適切な文書を推薦する技術の重要性が高まっている．このような文書データ推薦のための方法として圧縮技術に基づく手法が提案されている [1]．この手法ではまず，ユーザの閲覧した文書を LZ78 符号により圧縮することで辞書を作成し，この辞書を用いて推薦の候補である文書を圧縮する．そして，各推薦候補の文書の圧縮率を閲覧文書との類似度とみなすことで，その中から最も圧縮された文書の推薦を行う．

しかしながら，この手法を用いた推薦ではユーザが予め閲覧した複数の文書に含まれる単語の並びから圧縮に用いる辞書を作成するため，ユーザの閲覧した文書のトピックが異なる場合，安定した推薦が行えないという問題点がある．また，新聞記事等の文書集合ではトピックに応じてその内容の広がりが異なると考えられるが，従来手法ではこうしたトピック毎の性質を考慮することができない．そこで，本研究ではトピックモデルから得られる文書の各トピックへの所属確率を用いて文書間の類似度を直接算出し，文書を推薦する手法を提案する．新聞記事データを用いた実験により本手法の有効性を示す．

2　準備

2.1　LZ78 符号

まず，従来研究でユーザの閲覧文書と推薦候補文書との類似度を算出するのに用いられている LZ78 符号について説明する．LZ78 符号は Ziv と Lempel らによって提案された圧縮アルゴリズムである．いま，文書 $\boldsymbol{d}_j$ は出現してきた記号系列に従って $\boldsymbol{d}_j = (y_{j1}, y_{j2}, \cdots, y_{jR_J})$ と表現されるものとする．ただし，R_J は文書 $\boldsymbol{d}_j$ の系列長である．入力文書は 2 つの値からなる出力 $< i, c >$ に圧縮され，i は最長一致する系列の辞書番号であり，c は最長一致に次ぐ記号を指す．

2.2　問題設定

L 件からなる全ての文書集合 $\mathcal{D} = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_L\}$ が与えらたもとで，そのうち，ユーザが閲覧した N 件からなる文書集合をプロファイルを $\mathcal{D}_G = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_j, \cdots, \boldsymbol{d}_N\}$ とする ($\mathcal{D}_G \subset \mathcal{D}$)．また，$U$ 件からなる推薦の候補である文書集合 $\mathcal{D}_U = \{\boldsymbol{d}_1, \boldsymbol{d}_2, \cdots, \boldsymbol{d}_u, \cdots, \boldsymbol{d}_U\} (\mathcal{D}_U \subset \mathcal{D})$ とし，その中からプロファイル $\mathcal{D}_G$ に最も類似する文書を，推薦文書 $\hat{\boldsymbol{d}}_u$ としてユーザに推薦する．

2.3　PRDC[2]

Pattern Representation Scheme Using Data Compression（以下，PRDC）は LZ78 符号を用いてユーザの閲覧文書集合プロファイル $\mathcal{D}_G$ と推薦の候補である入力文書 $\boldsymbol{d}_u$ との類似度を圧縮率を基に算出し，推薦する文書を決定する手法である．プロファイル $\mathcal{D}_G$ に含まれる各文書 $\boldsymbol{d}_j$ から N 個の辞書 e_j を作成し ($j = 1, 2, \cdots, N$)，推薦の候補である入力文書 $\boldsymbol{d}_u$ とプロファイルに含まれる各文書 $\boldsymbol{d}_j$ との類似度 $sim(\boldsymbol{d}_u, \boldsymbol{d}_j)$ を以下のように算出する．

$$\mathrm{sim}(\boldsymbol{d}_u, \boldsymbol{d}_j) = \frac{l_{com}(\boldsymbol{d}_u, e_j)}{l_{in}(\boldsymbol{d}_u)} \tag{1}$$

ただし，$l_{com}(\boldsymbol{d}_u, e_j)$ は辞書 e_j によって圧縮された系列長であり，$l_{in}(\boldsymbol{d}_u)$ は圧縮前の系列長を表す．この PRDC を文書推薦に用いるため，入力文書 $\boldsymbol{d}_u$ とプロファイル $\mathcal{D}_G$ との類似度を以下のように算出し，その類似度が最も高い文書をユーザに推薦する．

$$\mathrm{sim}(\boldsymbol{d}_u, D_G) = \min_{\boldsymbol{d}_j \in D_G} \left(\frac{l_{com}(\boldsymbol{d}_u, e_j)}{l_{in}(\boldsymbol{d}_u)} \right) \tag{2}$$

3　従来研究

3.1　PRDCUD[1]

PRDC ではユーザが閲覧する文書数が多くなるにつれて，圧縮する計算量が膨大となってしまう．また，各文書それぞれに対して辞書 e_j を作成するため，プロファイル $\mathcal{D}_G$ 全体としての文書の特徴が捉えられていないという問題点がある．そこで，各文書 $\boldsymbol{d}_j$ から複数の辞書を作成する代わりに，プロファイル $\mathcal{D}_G$ に含まれる全ての文書を 1 つの系列 $\bar{\mathcal{D}}_G = (y_{11}, \cdots, y_{R_1}, y_{21}, \cdots, y_{R_2}, \cdots, y_{N1}, \cdots, y_{R_N})$ に統合したもとで 1 つの辞書 E を作成し，圧縮を行う手法 (Pattern Representation Using Data Compression With a United document 以下，PRDCUD) が提案されている [1]．このとき，プロファイル $\mathcal{D}_G$ と入力文書 $\boldsymbol{d}_u$ との類似度は以下のように算出することができる．

$$\mathrm{sim}(\boldsymbol{d}_u, \mathcal{D}_G) = sim(\boldsymbol{d}_u, \bar{\mathcal{D}}_G) = \frac{l_{com}(\boldsymbol{d}_u, E)}{l_{in}(\boldsymbol{d}_u)} \tag{3}$$

ただし，$l_{com}(\boldsymbol{d}_u, E)$ は辞書 E によって圧縮されたときの入力文書 $\boldsymbol{d}_u$ の系列長である．(3) 式を用いて推薦候補である文書集合 $\mathcal{D}_U$ に対して類似度を算出し，圧縮距離が最小となる文書 $\boldsymbol{d}_u$ をユーザに推薦する．

3.2　Combined Method[1]

Combined Method では上記の PRDCUD に加えて，文書を形態素解析することで得られる単語頻度ベクトル $\boldsymbol{d}'_j = (x_{j1}, x_{j2}, \cdots, x_{jV})$ から，入力文書 $\boldsymbol{d}_u$ とプロファイル $\mathcal{D}_G$ との類似度を算出する．ただし，V は語彙 $W = \{w_1, w_2, \cdots, w_V\}$ に含まれる単語の総種類数，x_{jv} は文書 $\boldsymbol{d}_j$ の単語 w_v の出現回数である．いま，$\boldsymbol{v}_{\mathcal{D}_G}$ はプロファイル $\mathcal{D}_G$ に含まれる全ての文書 $\boldsymbol{d}_j$ を統合した $\bar{\mathcal{D}}_G$ から得られる単語頻度ベクトルであり，$\boldsymbol{v}_{\boldsymbol{d}_u}$ は入力文書 $\boldsymbol{d}_u$ の単語頻度ベクトルであるとすると，入力文書 $\boldsymbol{d}_u$ と統合文書 $\bar{\mathcal{D}}_G$ との類似度は以下の (4) 式で算出される．

$$\mathrm{sim}(\boldsymbol{d}_u, \bar{\mathcal{D}}_G) = \frac{\boldsymbol{v}_{\mathcal{D}_G} \cdot \boldsymbol{v}_{\boldsymbol{d}_u}}{|\boldsymbol{v}_{\mathcal{D}_G}||\boldsymbol{v}_{\boldsymbol{d}_u}|} \log \left(\frac{l_{com}(\boldsymbol{d}_u, E)}{l_{in}(\boldsymbol{d}_u)} \right) \tag{4}$$

また，単語頻度ベクトルの各成分は tf-idf 値を用いて以下のように重み付けされている．

$$w_v = tf_v idf_v = \frac{\log(tf_v + 1)}{\log(V_G)} \log \left(\frac{N}{df_v} \right) \tag{5}$$

ただし，tf_v は単語 w_v の出現頻度であり，idf_v は単語 w_v を含む文書数，V_G はプロファイル $\mathcal{D}_G$ に含まれる単語の出現頻度の合計である．

4　提案手法

カテゴリ「スポーツ」における「野球」や「サッカー」等，文書に内在する観測されない潜在的な話題のことを潜在トピックと総称し，それらトピックを確率的に表現するモデルをトピックモデルと呼ぶ．本研究ではこのトピックモデルを扱うことで得られる文書の各トピックへの所属確率を用いて，文書間の類似度を算出し，それを基に文書を推薦する手法を提案する．

†早稲田大学
†† 湘南工科大学

4.1 Latent Dirichlet Allocation

本研究では，トピックモデルとして代表的な Latent Dirichlet Allocation(以下，LDA)[3] を用いて文書の潜在的な話題を表現する．LDA は文書中に含まれる単語の個々に潜在トピックを仮定し，それらのトピックの混合比によって 1 つの文書のトピックを表現するマルチトピックモデルである．LDA は文書 $\boldsymbol{d}_j$ に含まれる各単語がどのトピックから生成されたかを表す潜在変数 $(z_{\boldsymbol{d}_j 1}, z_{\boldsymbol{d}_j 2}, \cdots, z_{\boldsymbol{d}_j V})$ の生成確率 $\boldsymbol{\theta}_{\boldsymbol{d}_j}$ がディリクレ分布 $Dir(\boldsymbol{\theta}_{\boldsymbol{d}_j} \mid \boldsymbol{\alpha})$ に従うと仮定した潜在クラスモデルである．このとき，文書 $\boldsymbol{d}_j$ の生成確率 $p(\boldsymbol{d}_j \mid \boldsymbol{\alpha}, \boldsymbol{\beta})$ は以下の (6) 式のように与えられる．

$$p(\boldsymbol{d}_j \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int Dir(\boldsymbol{\theta}_{\boldsymbol{d}_j} \mid \boldsymbol{\alpha}) \left(\prod_{v=1}^{n_{\boldsymbol{d}_j}} \sum_{m=1}^{M} p(w_v \mid z_{\boldsymbol{d}_j v}, \boldsymbol{\beta}) p(z_{\boldsymbol{d}_j v} \mid \boldsymbol{\theta}_{\boldsymbol{d}_j}) \right) d\boldsymbol{\theta}_{\boldsymbol{d}_j} \quad (6)$$

ただし，$n_{\boldsymbol{d}_j}$ は文書 $\boldsymbol{d}_j$ に含まれる全単語数であり，パラメータ $\boldsymbol{\alpha}, \boldsymbol{\beta}$ は LDA のハイパーパラメータであり，これら 2 つのパラメータを調節することで様々な文書の潜在トピックを柔軟に表現することができる．

4.2 クラスタリングに基づく推薦手法

まず，本研究では全ての文書集合 $\mathcal{D}$ を用いて LDA の学習を行い，各文書 $\boldsymbol{d}_j$ のトピックへの所属確率を算出することで得られた $\boldsymbol{\theta}_{\boldsymbol{d}_j} = (\theta_{\boldsymbol{d}_j 1}, \theta_{\boldsymbol{d}_j 2}, \cdots, \theta_{\boldsymbol{d}_j M})$ を用いて，ユーザの閲覧文書集合プロファイル $\mathcal{D}_G$ と推薦候補の文書集合 $\mathcal{D}_U$ を K-means によってクラスタ集合 $\mathcal{M} = \{\mathcal{M}(\boldsymbol{\mu}_1), \mathcal{M}(\boldsymbol{\mu}_2) \cdots, \mathcal{M}(\boldsymbol{\mu}_K)\}$ へ割り当てる．このとき，$\boldsymbol{\mu}_k$ は各クラスタごとの代表ベクトルであり，以下の (7) 式のように与えられる．

$$\boldsymbol{\mu}_k = \frac{\sum^{\boldsymbol{d} \in \mathcal{D}_U \cup \mathcal{D}_J} q_{jk} \boldsymbol{\theta}_{\boldsymbol{d}_j}}{\sum^{\boldsymbol{d} \in \mathcal{D}_U \cup \mathcal{D}_J} q_{jk}} \quad (7)$$

各クラスタの代表ベクトル $\boldsymbol{\mu}_k$ が与えられたもとで，これらの代表ベクトルと各文書のトピック所属確率 $\boldsymbol{\theta}_{\boldsymbol{d}_j}$ との類似をユークリッド距離によって算出し，それが最小となるクラスタ番号 $\hat{k}$ へ $\boldsymbol{\theta}_{\boldsymbol{d}_j}$ を所属させる．

$$\hat{k} = \arg\min_{\boldsymbol{\theta}_{\boldsymbol{d}_j}} q_{jk} ||\boldsymbol{\mu}_k - \boldsymbol{\theta}_{\boldsymbol{d}_j}||^2 \quad (8)$$

ただし，q_{jk} は j 番目のデータ $\boldsymbol{d}_j$ がどのクラスタに所属するかを示す変数であり，以下のように定義される．

$$q_{jk} = \begin{cases} 1 & (\boldsymbol{d}_j \in \mathcal{M}(\boldsymbol{\mu}_k)) \\ 0 & (otherwise) \end{cases} \quad (9)$$

q_{ik} の状態変化がなくなるまで K-means の更新を行うことで，文書を K 個のクラスタに分割する．最終的に得られた結果より，プロファイル D_G と同一のクラスタに所属する推薦候補文書との類似度を (10) 式によって算出し，そのうち最も類似度が高いものをユーザに推薦する．

$$\mathrm{sim}(\boldsymbol{d}_u, \bar{\mathcal{D}}_G) = \min_{\boldsymbol{d}_j, \boldsymbol{d}_u \in \mathcal{M}(\boldsymbol{\mu}'_k)} \mathrm{sim}(\boldsymbol{d}_u, \bar{d}_j) = \frac{\boldsymbol{\theta}_{\boldsymbol{d}_u} \cdot \boldsymbol{\theta}_{\boldsymbol{d}_j}}{|\boldsymbol{\theta}_{\boldsymbol{d}_u}||\boldsymbol{\theta}_{\boldsymbol{d}_j}|} \quad (10)$$

5 実験・考察

毎日新聞 2010 年版の記事 10000 件を用いて実験を行う．まず，文書集合に LDA の経験的に求めたパラメータ $\boldsymbol{\alpha} = 0.1, \boldsymbol{\beta} = 0.05$ として文書集合に 100 個の潜在トピックをそれぞれ割り当てる．その後，以下の 10 個のトピックに所属する文書のみを抽出する．そして，10 個のトピックの中から T 個のトピックを選択し，そこから計 10 件の記事をランダムに選択することでユーザが閲覧した文書集合であるプロファイル $\mathcal{D}_G$ を作成する．200 件の評価用文書を設定し，うち，20 件をプロファイル $\mathcal{D}_G$ に存在する m 個のトピックから，その他 180 件は関連のないトピックからそれぞれ選択する．

表 2. 抽出したトピックと各トピック毎の特徴語

トピック番号 m'	トピック	所属文書数
m'=1	ゴルフ	40
m'=2	ギリシャ財政危機	30
m'=3	テレビ番組	36
m'=4	陸上選手権	50
m'=5	経済	40
m'=6	観光，ホテル	155
m'=7	相撲賭博	38
m'=8	アジア圏政治	51
m'=9	裁判	31
m'=10	トヨタリコール問題	130

例えば，T=2 のとき，20 件の関連文書は 2 個のトピックから 10 件ずつ抽出し，その他 180 件の非関連文書は 18 個のトピックから 10 件ずつ抽出する．そして従来手法では，プロファイル $\mathcal{D}_G$ と各評価用文書との類似度を算出し，類似度の高い順に文書をソートし，上位 r 件の文書を推薦する．提案手法では，プロファイル $\mathcal{D}_G$ が所属するクラスタに含まれる文書から類似度の高い上位 r 件を推薦する．各手法の評価指標は以下に算出される F 値とする．

$$\text{精度} = \frac{\mathcal{RD}_{\leq r}}{r}, \text{再現率} = \frac{\mathcal{RD}_{\leq r}}{\mathcal{RD}} \quad (11)$$

$$F \text{ 値} = \frac{2 * \text{精度} * \text{再現率}}{\text{精度} + \text{再現率}} \quad (12)$$

ただし，$\mathcal{RD}$ は関連文書総数，$\mathcal{RD}_{\leq r}$ は上位 r 件に存在する関連文書数とする．また，ランク r は 10〜20 で変動させ各トピック毎に 10 回繰り返し実験を行い，得られた F 値の平均を最終的な評価指標とした．

以下，各手法ごとの F 値の結果を示す．

表 3. 実験結果

トピック	PRDUC	Combine	提案手法
T=1	**0.546**	0.559	0.298
T=3	0.369	0.231	**0.475**
T=5	0.304	0.238	**0.395**

表 3 より，トピックが複数存在するような場合では本研究の手法の有効性を確認することができた．ただし，プロファイルに存在するトピック数が少ない場合，従来手法に精度が劣ってしまうことが確認できる．単語の出現する種類に限りがあるような場合では，出現単語を辞書に逐次登録する圧縮技術を用いた文書推薦が適していると考えられる．

6 まとめと今後の課題

本研究ではトピックモデルを用いた文書推薦手法について提案し，実データを用いた実験により本手法の有効性を示した．今後の課題として，トピックモデルを用いて圧縮をより効率良く行う新たな圧縮技術に基づく推薦手法の提案などが考えられる．

参考文献

[1] T. Suzuki, S.Hasegawa, T.Hamamoto, A.Aizawa. “Document Recommendation Using Data Compression,” *Procedia - Social and Behavioral Sciences,* vol.27, pp. 150–159, 2011.

[2] T.Watanabe, K. Sugawara and H.Sugihara,“A new pattern representation scheme using data compression,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.24 pp.579-90 2002.

[3] D.Blei A.Y. Ng and M.Jordan,“Latent Dirichlet allocation,” *Neural Information Proceeding Systems*, Vol.14 2001.