トピックへの所属確率分布を考慮した学術論文へのキーワードの割り当て手法に関する研究

1X17C008-9　浅見怜
指導教員　後藤正幸

1. 研究背景と目的

研究者は様々な論文誌から自らの研究テーマに関連した論文の調査を行う．その際，付与されたキーワードによって候補となる論文を特定した後，本文を読むというプロセスをとる場合も多い．キーワードは論文の内容を端的に表しており，重要な要素である．しかし論文の著者は投稿する論文誌の研究分野の学術的背景に沿ってキーワードを付与するため，内容的に類似性が高い論文であっても論文誌によって付与されるキーワードが異なる．例えば，経営情報の学会では「機械学習」がキーワードとして付与された論文が，仮に人工知能の学会に投稿されていたならば「XGBoost」のような具体的手法に関するキーワードが付与される状況も考えられる．そのため，調査したい事柄と内容的に関連がある場合でも，付与されているキーワードが検索者の意図と異なり，検索者にとって必要な情報を発見することが困難になることが考えられる．このような背景から，ある学術的背景でキーワードが付与されている論文に対し，検索者の研究分野に由来する論文誌に対して付与されるキーワード傾向に基づいて，キーワードを再付与する手法が求められている．

ここで，キーワードの付与にあたり，論文のキーワードを文書におけるタグとみなし，タグ推定トピックモデルを活用する方法が考えられる．ある論文誌のデータで学習を行ったモデルを用いてキーワードを付与することで，特定の論文誌に由来したキーワード付与が可能となる．タグ推定トピックモデルを用いたタグを付与する従来手法として，文書とタグの関係性を学習した後，文書内で最も所属確率が高いトピックに基づき，付与するタグを決定する計算方法 [1] が提案されている．しかし，従来手法ではタグはいずれか 1 つのトピックから生起される状況想定している．そのため，研究テーマが複数のトピックから生成される論文に対して，従来手法を適用した場合，文書のトピック分布と異なるトピック分布を持つキーワードが付与されてしまう可能性があり，論文の内容に即したキーワードが付与できないという問題がある．

本研究では，代表的なタグ推定トピックモデルである Correspondence-LDA[2] を用いて，特定の論文誌におけるキーワードの付与の傾向に基づいて，論文に適した新たなキーワードを付与することを目的とする．そしてキーワードを付与する際，文書の内容はトピック分布の形状で表されることに着目し，文書とキーワードのトピック分布の類似性を用いることで，文書内に出現する複数のトピックを考慮した新たなキーワードの付与方法を提案する．最後に実際の論文誌のデータを用いた実験により，提案手法の有効性を示す．

2. 準備

2.1. Correspondence-LDA(Corr-LDA)

Corr-LDA は，タグ推定トピックモデルの 1 つであり，文書とタグの生成過程をモデル化した確率的生成モデルである．Corr-LDA では，単語の背後に単語トピック，タグの背後にタグトピックを仮定し，単語を生成したトピックのみを用いてタグを生成する．ここで，文書数を D，トピック数を K，総単語数を N，総タグ数を M，文書 d 内の単語数を N_d，文書 d 内の単語トピック z_k の単語数を N_{dk}，文書 d 内のタグ数を M_d とする．また，文書 d におけるトピック分布を $\boldsymbol{\theta}_d = (\boldsymbol{\theta}_{d1}, ..., \boldsymbol{\theta}_{dK})$，トピックごとの単語分布を $\boldsymbol{\Phi} = (\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_K)$，トピックごとのタグ分布を $\boldsymbol{\Psi} = (\boldsymbol{\psi}_1, ..., \boldsymbol{\psi}_K)$ とする．このとき，文書 d の単語ベクトル $\boldsymbol{w}_d = (w_{d1}, .., w_{dN})$ とタグベクトル $\boldsymbol{x}_d = (x_{d1}, .., x_{dM})$ の同時分布は式 (1) で表される．

$$p(\boldsymbol{w}_d, \boldsymbol{x}_d|\boldsymbol{\theta}_d, \boldsymbol{\Phi}, \boldsymbol{\Psi}) = \sum_{z_d} \left[\prod_{n=1}^{N_d} p(z_{dn}|\boldsymbol{\theta}_d) p(w_{dn}|\phi_{z_{dn}}) \prod_{m=1}^{M_d} \sum_{k=1}^{K} p(y_{dm} = z_k|\boldsymbol{z}_d) p(x_{dm}|\psi_k) \right] \quad (1)$$

潜在トピックを $\mathcal{Z} = \{z_1, \cdots, z_K\}$，文書 d の単語トピックベクトルを $\boldsymbol{z}_d = (z_{d1}, ..., z_{dN_d})$，文書 d のタグトピックベクトルを $\boldsymbol{y}_d = (y_{d1}, ..., y_{dM_d})$，全ての単語トピックに関する和を $\sum_{z_d}$ と表す．ただし，$z_{dn} \in \mathcal{Z}$　かつ $y_{dm} \in \mathcal{Z}$ である．また，$p(y_{dm} = z_k|\boldsymbol{z}_d) = N_{dk}/N_d$ は単語トピックベクトルが与えられた時のタグトピックが z_k である確率を表す．Corr-LDA のグラフィカルモデルを図 1 に示す．

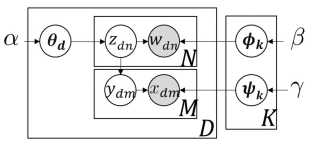

図 1: Corr-LDA のグラフィカルモデル

ここで，$\boldsymbol{\theta}_d \sim \mathrm{Dir}(\alpha)$, $\boldsymbol{\phi}_k \sim \mathrm{Dir}(\beta)$, $\boldsymbol{\psi}_k \sim \mathrm{Dir}(\gamma)$ である．ただし，$\mathrm{Dir}(\cdot)$ はディリクレ分布を表し，α, β, γ はそれぞれのディリクレ分布におけるハイパーパラメータである．

2.2. 従来のタグ付与方法

タグ推定トピックモデルでは，文書中の単語とタグを入力とし，ギブスサンプリング [3] を用いて単語とタグにトピックを割り当てる．タグが付与されていない文書 d(以下，付与対象データ) に対してタグを付与する際，文書 d における単語のトピック分布 $p(z_k|d)$ とタグトピック z_k における，タグ x_i の所属確率 $p(x_i|z_k)$ を用いて，各タグ x_i の出現確率 $f(x_i)$ を式 (2) で算出する．式 (2) により求められる $f(x_i)$ の大きい順にタグを付与する．

$$f(x_i) = \sum_{k=1}^{K} p(x_i|z_k) p(z_k|d) \quad (2)$$

3. 提案手法

3.1. 提案への着想

従来の計算方法では，ある単語トピック z_k において，文書 d の所属確率 $p(z_k|d)$ とキーワード x_i の出現確率 $p(x_i|z_k)$ が共に高い場合，このトピックのみに由来したキーワードが付与されやすく各文書が持つ複数のトピックを同時に考慮したキーワード付与ができない．そのため文書のトピック分布と異なるトピック分布を持つキーワードが付与され，文書の内容に即したキーワードが付与されないという問題点がある．

そこで，提案手法では，トピック分布の形状に着目し，付与対象データ内での単語のトピック所属確率と各キーワードのトピック所属確率の分布の差分を考慮することで，トピック分布の類似性により，キーワードを割り当てる．これにより，文書内に出現している複数のトピックを考慮し，文書の内容に即したキーワードの付与を行うことが可能となる．

3.2. 定式化

文書 d における単語のトピック分布 $p(z_k|d)$ と，全文書内でのタグ x_i に対してタグトピック z_k となる確率 $p(z_k|x_i)$ の二乗誤差をとることで単語とキーワードのトピック所属確率分布の類似性を測り，キーワードを付与する方法を提案する．式 (3) のように二乗誤差をとり，小さい順に並べる．ここでは，$\tilde{f}(x_i)$ の値が小さい程，単語とキーワードのトピック所属確率分布の類似度が高いことを意味している．

$$\tilde{f}(x_i) = \sum_{k=1}^{K} (p(y_i = z_k|x_i) - p(z = z_k|d))^2 \quad (3)$$

4. 実データ実験

論文データに Corr-LDA を適用し，従来手法と提案手法によりキーワードを付与し，割り当て結果の比較を行った．

4.1. 実験条件

評価実験には，2018～2020 年度の人工知能学会全国大会の論文データを用いる．データ数は 2068 件である．加えて，別論文誌として経営情報学会の論文データ 10 件を用いる．Corr-LDA のトピック数 K は 20 とし，ハイパーパラメータ α, β, γ は 0.1 に設定し，エポック数は 10,000 とした．

実験 1．人工知能学会全国大会の論文データのうち，1968 件を学習データ，100 件を付与対象データとする．学習データは概要文とキーワード，付与対象データは概要文のみを取得し，各手法でキーワードを付与し，実際のキーワードと比較する．1 つの付与対象データへ付与するキーワード個数を変化させ，一致していたキーワードの累計正解数を比較する．

実験 2．実験 1 と同様に人工知能学会全国大会の論文データ 1968 件を学習データとし，経営情報学会の論文を付与対象データとする．各手法でキーワードを付与し，人工知能に関連したキーワードが付与されているかの定性評価を行う．

4.2. 実験 1 結果と考察

それぞれの手法ごとのキーワード累計正解数を表 1 に示す．付与対象データ 100 件の正解キーワード総数 (付与対象データのみで出現したキーワードを除く) は 180 個である．

表 1: 手法ごとのキーワード正解数

付与数	5	10	20	30	50
従来	13	27	**47**	55	72
提案	**18**	**28**	**47**	**60**	**73**

いずれのキーワードの付与個数においても，提案手法は従来手法と同等かそれ以上の正解数を示している．そのため，提案手法はキーワードと文書のトピック分布の類似性を考慮したことで，概要文の内容に即して適切なキーワードを付与できていると考えられる．

4.3. 実験 2 結果と考察

経営情報学会の論文データにおいて，抜粋した 1 件の概要を表 2，それぞれの手法で付与された上位 5 件のキーワードの結果を表 3 に示す．表 3 において概要文と内容に関係があると定性評価により判断したキーワードを網掛で示す．

表 2: 付与対象データ概要

タイトル	The Study on Determinant Factors of Non-performing Loan Accumulation
キーワード	Bad Debt, Artificial Intelligence, Determinants
論文概略	不良債権の発生に影響を与えている要因をネットワーク分析により定量的に選択

表 3: 手法ごとの付与されたキーワード

順位	従来	提案
1	**bayesian network**	**big data**
2	data jacket	**data mining**
3	**manifold learning**	**decision tree**
4	**analogy**	**network analysis**
5	co-creation	**complex network**

表 3 より，従来手法に比べて，提案手法は論文内容に関連した人工知能に由来するキーワードが付与されていることがわかる．このことから，提案手法により特定の研究分野に関連するキーワードの付与は可能と考えられる．

5. まとめと今後の課題

本研究では，学術論文に対して，他の論文誌の傾向に基づいて，論文の概要文の内容に即したキーワードを再付与する手法を提案した．また，キーワードの付与に関する実験を行い，提案手法の有効性を示した．今後の課題として，論文データ以外への提案手法の有効性の確認や，他のタグ推定トピックモデルを用いた場合の有効性の検証などが挙げられる．

参考文献

[1] 加藤亮，吉川大弘，古橋武，“潜在的なトピックを仮定した文書への自動タグ付与に関する検討，”*30th Fuzzy System Symposium*, 2014.

[2] Blei M.D.and Jordan I.M.,“Modeling annotated data,” *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp.127–134, 2003.

[3] Griffiths, T. L. and Steyvers, M.,“Finding scientific topics,”*Proceedings of the National Academy of Sciences of the United States of America*, Vol.101, No. Suppl 1, pp. 5228-5235, 2004.