2P-3

トピックモデルを用いた中文心理カウンセリングの文書分類

単壮† 渡邊恵太† 加藤昇平†

†名古屋工業大学

1 はじめに

近年，多くの人がストレスによる心理問題を抱えており，社会問題にもなっている．しかし，カウンセリングを受けるには時間的，金銭的負担が少なくない．そこで，誰でも気軽に受けられるカウンセリングシステムが求められている．既存のカウンセリングシステムは人間の言葉でよく用いられる同義語や類義語，婉曲表現といった言語の多様性により，そもそも利用者の発言意図を理解することが難しい．ユーザの質問に対してより高い精度で答えを与えるカウンセリングシステムを構築するために，ユーザがどのような心理的問題を抱えているか分析することが重要だと考えられる．本研究では，ユーザの質問文を入力とし，どのような問題に属するか判別する文書分類手法を提案する．

2 分類手法の概要

本稿で提案する手法ではユーザが入力した質問文を分析するために LDA モデルを使い，LDA の分析結果を用いて未知の文書を分類する．分類の流れを図 1 に示す．

2.1 LDA（Latent Dirichlet Allocation）

LDA[1] は，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の一つであり，一つの文書中に複数のトピックが存在することを表現できる潜在的意味解析手法である．LDA によって，文書を構成するトピックの多項分布及び各トピックを構成する単語の多項分布を表現することができる．LDA において文書は，まずその文書のトピック分布に従いトピックが選択され，そのトピックの単語分布に従って単語が選択される過程で生成されると仮定される．LDA モデルのグラフ表現を図 2 に示す．α はトピックの事前分布の k 次元のハイパパラメーターであり，β はトピックの単語分布のパラメーターである．また，N は文書中の単語数を表す．α と β が与えられたとき，トピックの混合分布 θ，N 個のトピック集合 $\boldsymbol{z}$，N 個の単語集合 $\boldsymbol{w}$ は式 (1) によって与えられる．

$$P(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n\theta)p(w_n|z_n, \beta) \quad (1)$$

ここで，θ と β は潜在的なパラメーターであり，文書を観測したとき，この θ と β を推定することで文書がどのようなトピックで成り立っているか推定できる．

2.2 文書の分類

本研究ではデータベースの問題文数と単語数によって，$\alpha = \frac{50}{\text{トピック数}}$，$\beta = 0.1$，$k = 50$ と設定した．LDA

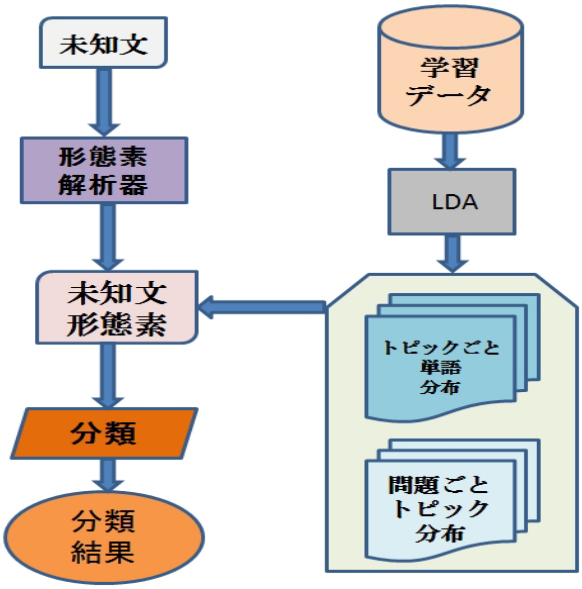

図 1: Flow of the classification method

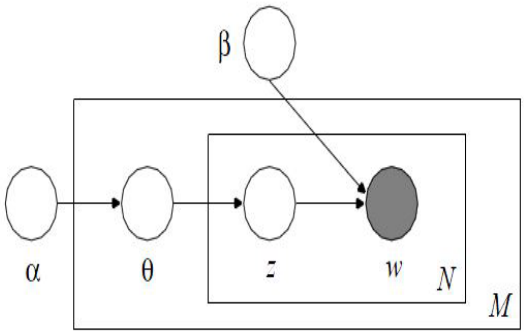

図 2: Graphical model representation of LDA

モデルによって学習データからトピックごとの単語分布を生成し，問題タグごとに各トピック分布を生成する．単語分布とトピック分布の例を表 1 に示す．

未知文は形態素解析器 [2] により，形態素単位に分割した．本稿では形態素単位に分けられた語を単語として扱う．単語ごとにその単語がどのトピックに属するかを判別し，そのトピックがどの問題タグに属するか判別することで，単語ごとに問題タグを判別する．なお，判別には生起確率を用いた．対象の単語の生起確率が最も高いトピックをその単語が属するトピックと判別する．トピックも同様に，対象のトピックの生起確率が最も高いタグをそのトピックが属するタグと判別する．これを未知文中の全ての単語に対して行い，未知文中で最も多く出現した問題タグをその未知文の問題タグとする．

*A Document Classification for Chinese Psychological Counseling Using Topic Model, Zhuang Shan†, Keita Watanabe† and Shohei KATO†
†Nagoya Institute of Technology
Gokiso-cho, Showa-ku, Nagoya 466-8555, Japan
{shan, watanabe, shohey}@katolab.nitech.ac.jp

表 1: The Example of the Distribution of Words and Topics

TOPIC.14		TAG.1	
単語	生起確率	トピック	生起確率
彼女	0.116900	Top.14	0.242518
終了	0.043869	Top.32	0.004342
彼	0.039412	Top.46	0.003237
愛	0.021050	Top.2	0.002921
恋愛	0.020210	Top.18	0.002763
好き	0.020188	Top.3	0.002684
私たち	0.014798	Top.26	0.002658
私	0.014798	Top.11	0.002632
相談	0.013178	Top.38	0.002342

3 実験

提案手法の有効性を確認するため，LDA による文書分類実験を行い，TF・IDF による文書分類法を比較対象として比較実験を行う．Liu ら [4] を参考に実験用のデータベースとして世界最大の中文知識問答プラットホーム BAIDUZHIDAO からよく質問されている典型的な心理問題文の内，3 種類（恋愛，人間関係，自己認識）を収集した．恋愛に関する問題文は 387 問，同僚，クラスメート，友達との人間関係に関する問題文は 395 問，自己認識に関する問題は 215 問，計 995 問の問題文を収集した．サンプリング調査法により，995 問から問題タグ毎にランダムにそれぞれ 100 問ずつ選出し，Leave-one-out 交差検定方法により，提案手法と比較手法の文書分類を各 300 回行った．

3.1 TF・IDF による文書分類法

TF・IDF は，文書中の単語に関する重みの一種であり，主に情報検索や文書要約などの分野で利用されている．TF・IDF は，TF（単語の出現頻度）と IDF（逆文書頻度）の二つの指標に基づいてを計算される. 単語 i における TF・IDF 値は式 (2) によって与えられる．

$$TF \cdot IDF = tf_{ij} \times idf_i = \frac{D_{ij}}{\sum_k D_{kj}} \times \log \frac{N}{n_i} \qquad (2)$$

ここで D_{ij} は単語 i の文書 j における出現回数，$\sum_k D_{kj}$ は文書 D_j の総単語数，N は総文書数，n_i は単語 i を含む文書数である．TF・IDF は単語 i の文書に対する区別性を表している．しかし，一般的な TF・IDF の計算式は文書の類別の違いを考慮しておらず，文書分類における単語の正しい重みを算出できない問題がある．Kuang ら [3] はこの問題に対して，文書の類別の違いを考慮した TF・IDF の式を提案した．式 (3) に Kuang らが提案した式を示す．

$$TF \cdot IDF \cdot C_i = tf_{ij} \times idf_i \times C_i = \frac{D_{ij}}{\sum_k D_{kj}} \times \log \frac{N}{n_i} \times \frac{1}{n_i - m_i + 1} \qquad (3)$$

ここで m_i は単語 i を含む文書の総数である．文書内の全単語に対する TFIDF・C_i の値の集合をその文書の TF・IDF とし，$TFIDF'$ と表す．本稿では $TFIDF'$ を使用した文書分類法を比較対象とする．分類手法の流れとしては，まず未知文の $TFIDF'$ を算出し，データベース内の各文書 $TFIDF'$ とのコサイン類似度を算出する．コサイン類似度が高い文書を上位 10 個選出し，10 個中最も多いタグを未知文のタグとする．

表 2: The Results of the Classification of LDA Method

問題タグ＼分類結果	恋愛	人間関係	自己認識	正解率
恋愛	83	14	4	83.0%
人間関係	12	80	8	80.0%
自己認識	7	15	78	78.0%

総正解率:80.3%

表 3: The Results of the Classification of TF・IDF method

問題タグ＼分類結果	恋愛	人間関係	自己認識	正解率
恋愛	67	25	8	67.0%
人間関係	19	64	17	64.0%
自己認識	16	27	57	57.0%

総正解率:62.6%

3.2 実験結果

実験結果を表 2 と表 3 に示す．提案手法と比較対象の問題タグ毎に 100 個の問題が分類された分布を表す．提案手法は問題タグごとの正解率は全て 75 %を超えており，総正解率も 80.3% の精度で判別できていることが分かる．さらに，提案手法を比較対象と比べ，提案手法の問題タグごとの正解率は全てにおいて比較対象を 16% 以上上回っている．このことから，提案手法は恋愛，人間関係，自己認識などの複雑な心理問題を分類可能な高い分類能力を持っていると考えられる．

4 おわりに

本稿では，カウンセリングシステムにおける LDA を用いた問題分類手法を提案した．判別実験において，提案手法の有効性が示唆された．今後の課題としてはより多種類の問題に対応する実験を行い，回答文を出せる手法を計画することで，カウンセリングシステムを実装する．

参考文献

[1] David M. BleAndrew Y. Ng, Andrew Y. Ng, Michael I. Jordan,“ Latent dirichlet allocation“ , The Journal of Machine Learning Research Volume 3,2003 Pages 993-1022

[2] W. Che, Z. H. Li and T. Liu,“ LTP: A Chinese Language Technology Platform ”, COLING ’10 Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations, 2010, Pages 13-16

[3] Qiaoyan Kuang, Xiaoming Xu“ Improvement and Application of TF・IDF Method Based on Text Classification ”,Internet Technology and Applications, 2010 International Conference on Digital Object Identifier,2010 , Pages 1-4

[4] Yuanchao Liu, Ming Liu, Zhimao Lu, Mingkai Song, ”Extracting Knowledge from On-Line Forums for Non-Obstructive Psychological Counseling Q&A System ”,International Journal of Intelligence Science, 2012, 2, 40-48