6P-01

多段型トピックモデルを用いたアンケート自由記述からの情報抽出

徐　闖†　　湯川　高志‡

長岡技術科学大学情報・経営システム工学専攻†

s121045@stn.nagaokaut.ac.jp　yukawa@vos.nagaokaut.ac.jp

1. はじめに

アンケートは選択肢設問と自由記述設問があり、前者は容易に集計できるが、後者は人間が文章を読むことが必要で分析するコストが高い。分析者(ユーザ)は自由記述設問においても回答者が書く話題をある程度想定している。実際の記述は、ユーザが想定した話題の場合も多いはずだが、想定外の話題も含まれている。これらを分けてそれぞれ分析すれば、よりわかりやすく話題を抽出できると考える。本研究では自動的に自由記述から話題を抽出することを目的とし、トピックモデルを多段階に用いて自由記述を想定内話題と想定外話題に分けてそれぞれ分析するシステムを提案する。想定した話題とはユーザが関心を持つ話題であり、想定外話題とはユーザが事前に考えていない話題である。

2. 関連研究と要素技術

細井ら[1]は、解析ソフトにより頻出語の共起ネットワークを描き、全体的傾向を分析したが、ユーザが自分で話題を抽出する必要がある。渡辺ら[2]は、文章間の類似度計算により、文書を分類したが、文章の意味解析が行われていない。以上の背景に踏まえて本研究はLatent Dirichlet Allocation(LDA)[3]を用いて自由記述から精度よく話題を抽出するシステムを提案する。

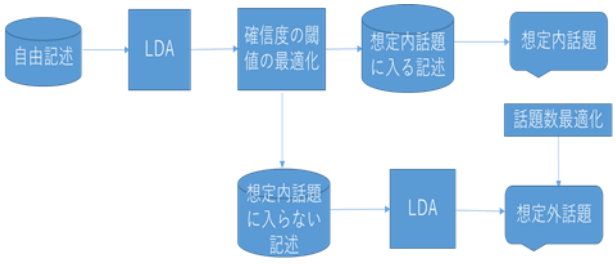

図１　提案システムの流れ

3. 多段型話題抽出システムの提案

本研究は、図１で示したようにシステムの全体構成としてまず、LDA は抽出する話題数は制御（外部から指定）できるが話題そのものを制御することはできない、それを入力の文書ベクトルの一部を協調することでユーザが指定した話題が抽出されるようにする。ただし，ユーザ指定のどの話題に対しても確信度が低い文章は想定外の話題とみなして、次の段に送る。次段では通常の LDA を用いて、指定された数の話題を特に事前に話題指定することなく抽出する。ここで確信度の閾値を決める必要がある。閾値以上の確信度を持つのは想定内話題の文章だとし、それ以下の確信度を持つのは想定外話題の文章だと考える。

3.1. 想定内話題の抽出

ユーザがあらかじめ指定した話題を抽出するために、システムでは話題の特徴語の重みを３倍にし、ベクトルを強調することで話題を抽出する。LDA は文章内の単語の出現確率を推定するトピックモデルなので、語の重みを大きくすると、トピックとして生成されやすい。倍数を 4,5 でも試したが、3 倍で十分に良い効果が出るため、3 倍にしたからである。

3.2. 想定外話題の抽出

想定外話題は話題数が事前に決められないという問題がある。そこで coherence という指標に基づいて半自動的に決定する手法を採用する。

4. パラメータの最適化

上述したように１段目の LDA において想定内話題と想定外話題を分離するために，確信度の閾値を決定する必要があり、その決定は実験的に行った。実験に用いたデータは大学の授業アンケートの講義に関する 2737 件の自由記述である。文章から名詞だけを抽出し、コーパスとして利用する。一定の確信度以上の想定内文書から、想定内話題に対し、それぞれ 20 件の文書を取り出し、目視でトピックに一致した件数を確認した。ここで一致率と言う指標を利用する。

一致率＝ 一致した文章数 / 文章総数(60)

図２と図３で示した閾値別の文章数と一致率を利用して計算すると、確信度が 0.85 の際に、正しく話題が抽出された文章数が最も多いことがわかるため、確信度の閾値は 0.85 である。

Information Extraction from questionnaire free description using multistage topic model

†Xu Chuang ‡Yukawa Takashi

Nagaoka University of Technology of Information & Management Systems Engineering

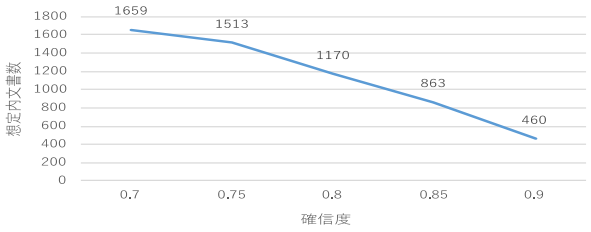

図 2 確信度別の文章数

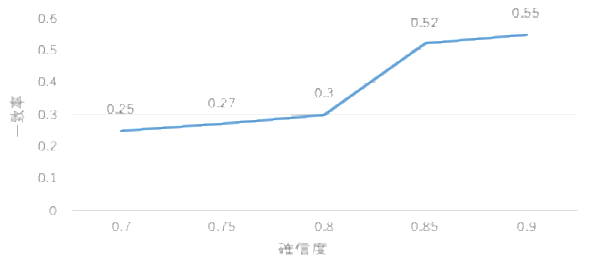

図 3 確信度別の一致率

本研究では、想定外話題数を決定するために話題の一貫性を表す Coherence 値を指標として話題数の最適化を図った。Coherence 値が高いほど単語同士の関連性が高いが、評価する際に Coherence 値は明確な範囲がなく、Coherence 値が高くてもそのトピック数が適切とは言えない。そのため、Coherence 値の正規化を行った。正規化した Coherence 最大値は 1 である。図 4 で示したようにトピック数を２から２０まで試し、それぞれの Coherence 正規値を算出した。トピック数が 3 の時に、文章がよくまとめられ、その後は Coherence 正規値が上がっていくが、傾向が変わらないのでトピック数 3 が良いことが分かる。

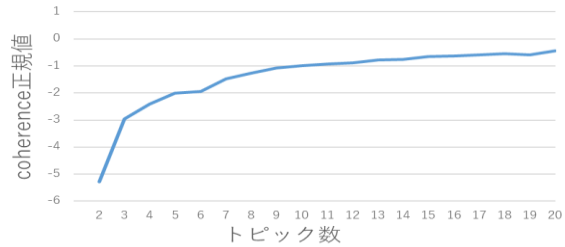

図 4 トピック数別の正規値

5. 評価実験と考察

提案したシステムの性能を実験によって評価した。評価実験においてデータセットはパラメータ最適化に用いたのと同じ授業アンケートであり、想定内話題としては板書、資料、テストの 3 つを設定した。想定した話題の抽出結果はユーザが関心を持つ話題をシステムに入力し抽出した話題である。この結果から，ユーザが与えた話題が確かに抽出されていることがわかる。

Topic1：板書　文字　説明　質問　スライド
Topic2：資料　内容　プリント　配布　教科書
Topic3：テスト　大変　答え　復習　時間

想定外トピックの抽出結果:

Topic1：課題　解説　レポート　理解　大変
Topic2：教室　人数　部屋　黒板　設備
Topic3：授業　内容　理解　話　先生

システムが出力した各トピックに分類された文章から人間がそれぞれ 20 件をランダムに抽出し、個々の文書はどのトピックに意味が一致するかまたはそれ以外かを実験者 3 人に評価してもらった。3 人の評価結果が違った場合、多数決で決める。評価結果は表 1 と表 2 で示す。

予想　/ 正解	テスト	資料	板書	それ以外
テスト	15	0	1	4
資料	3	9	1	7
板書	0	3	16	1

表 1　想定内話題の評価

予想　/ 正解	課題	教室	授業内容	それ以外
課題	11	0	2	7
教室	0	7	3	10
授業内容	3	0	14	3

表 2　想定外話題の評価

表から想定内話題の全体適合率を計算すると 0.67 で、想定外話題の全体適合率は 0.53 である。実際のトピックとその文書を見ると、「グラフについて分かりやすい解説のついた資料などがほしかった」という文書が「資料」の話題に分類されると考えられるが、「資料」と「解説」が共起したため、「課題」のトピックに分類された可能性がある。誤判定のなかで，ここで挙げたようなものが多くを占めた。

6. まとめ

本研究では、自由記述をユーザの想定内話題と想定外話題に分けてそれぞれの分析するシステムを提案した。想定内話題に関して、話題の特徴語の重みを強調することで明確に抽出できた。想定外話題に関して、確信度閾値と話題数の最適化により、分かりやすくトピックを抽出できた。今後の課題として、単語同士の共起により文章は違うトピックに分類された問題を解決する予定である。

参　考　文　献

[1] 細井　亮佑, 寺田　充伸,”テキストマイニングを用いたアンケート自由記述欄の分析による生活環境評価”, 日本建築学会九州支部研究報告第 50 号.

[2] 渡辺　智幸,”情報検索技術を用いたアンケートデータの分析手法に関する研究”, 2005 年度卒業研究概要.

[3] Blei, D.M., Ng, A.Y. and Jordan, M.I.: Latent Dirichlet Allocation, Journal of Machine Learning Research 3 (2003) 993-1022.