適応型テストのためのLDAを用いた項目間類似度の利用可能性

○杉山 剛[*1]，　加藤 嘉浩[*2]，石井 隆稔[*3]
*1 株式会社リクルートキャリア測定技術研究所，
*2 電気通信大学大学院情報システム学研究科，
*3 首都大学東京

1. はじめに

昨今，採用場面で用いられる適性検査は，従来の紙筆型からコンピュータを利用したeテスティングへの移行が進んでいる．eテスティングのなかでも，項目反応理論を用いたコンピュータ適応型テスト(Computerized Adaptive Test，CAT)は，少ない項目数で信頼性の高い測定結果を得ることができるため注目されている．適応型テストの出題項目選択において，出題済みの項目と似た項目を出題することには，局所独立の仮定が崩れることによる特性値推定上の問題や，受検者に出題傾向が偏ったテストであるという印象を与えてしまう問題などがある．これを回避するため，リクルートキャリアが提供している能力適性検査では，似た項目をグルーピングし，一連の出題の中で似た項目が重複して出題されることを回避している．しかし，この重複回避のためのグルーピングは，特に言語能力を主題とした項目においては類似の判定が難しく，現状では新たな項目をデータベースに加えるたびに，既存のすべての項目との類似判定を人手によりおこなっている．

このような問題に対し，項目の問題文へテキストマイニング手法を用いることにより項目間の類似性を算出する研究がなされている．たとえば，Songmuangら(2012年)[1]は複数の等質なテストを構成するにあたり，類似項目が含まれないようにテキストマイニング手法のひとつであるLatent Dirichlet Allocation(LDA)[2]を用いて問題文から機械的に項目の内容を推定し，なるべく類似した内容の項目が同じテストに含まれにくいテスト構成法を提案している．ただし，Songmuangらの研究では，人手による類似判定とLDAによる類似項目判定がどの程度一致しているかについて言及がない．また高木ら(2014年)[3]は項目の類似性を判定するにあたり，項目を解くための知識が問題文のどこに出現するかを出題形式別に明らかにし，その部分へLDAを適用することにより，精度の高い類似度算出を可能としている．ただし，能力適性検査においては，知識を問う問題ではなく，また出題形式や出題内容も多岐にわたるため出題形式と項目を解くための知識の関係を調べることは困難である．

そこで，本研究では能力適性検査において問題文のLDAによる類似度算出が人手による重複回避のためのグルーピングとどの程度一致するかを検証し，これを用いた類似度によるグルーピングが，重複回避のために使用可能かについて検証する．

2. 項目の類似度の数値化

各項目の内容的な類似度の算出にあたっては，トピックモデルのひとつであるLDAを用いる．LDAは文書中に複数個の潜在トピック(話題)が存在すると仮定し，文書中のそれぞれのトピックの割合(以下トピック分布)を単語の出現頻度を元に推定する手法である．能力適性検査における言語能力を主題とした項目の問題文は，多くの場合複数のテーマにわたって記述されていること

から，LDA が今回の目的に対して親和性が高いと考えた．ここで，類似している項目はトピック分布θが類似していると仮定する．トピック分布θの類似度の指標には，確率分布間の疑似距離(非類似度)指標である Jensen-Shannon ダイバージェンス(以下 JS ダイバージェンス)を用いる[4]こととした．なお，JS ダイバージェンスは疑似距離であり，類似度が高いほど値が小さくなり，分布が一致するときに最低値 0 をとる．

ここでまず，LDA で算出したトピック分布θの非類似度が言語能力を主題とした項目の非類似度指標として妥当かどうか確認するための実験(１)をおこなった(表 1)．

表 1．実験(１)概要

対象項目	能力適性検査GATの言語系項目のうち，内容による重複回避をおこなっている626項目（項目管理者によるグルーピング済）
分析前の処理	① 問題文と，誤答を含むすべての選択肢を1つの文章としてまとめる． ② 選択肢番号や設問文など内容に関係ない文字列を除去する．
分析処理	LDAによるトピック分布の推定をおこない，それぞれの項目ごとのトピック分布θと，θ間の疑似距離（JSダイバージェンス）を算出する．
LDAの 潜在トピック数	10,20…,150の15パターン ※150：項目管理者がおこなったグルーピングにおける，複数項目を含むグループ数に近い数字
評価の観点	① θの類似度が高い項目は実際に内容が類似しているか ② 項目管理者が分類して，類似していると判断した項目間のθの類似度は高いか

トピック分布θの類似度が高い(JS ダイバージェンスの低い)組み合わせ上位 100 組について項目内容を実際に読み比べて重複を回避すべきかどうかを調べた．

その結果，上位 100 組のなかには項目管理者が見ても確かに類似していると考えられる項目が多く含まれるほか，使用している単語は重複しているが，内容的には同時に出題しても問題ない組み合わせや，内容が類似していない組み合わせが少なからず含まれていた．内容が類似しているとはいえない組み合わせについて詳しく確認したところ，LDA が推定する潜在トピックのなかには「分類できないトピック」ともいうべき，多くの項目に共通しているトピックが含まれるが，他の項目との共通点が少ない項目では，この「分類できないトピック」の含有確率が目立って高くなっており，共通点が少ない項目同士のトピック分布が似てしまうという現象がおきていることがわかった． このようなトピックを特定して，除外したうえで類似度を再計算し，あらためて上位 100 組について重複を回避すべきかどうかを調べた結果，トピック数 100 近辺で重複を回避すべき組み合わせが最大となっていることがわかった(図 1)．

一方，項目管理者が分類して，類似していると判断した項目間のトピック分布θの類似度については，おおむね類似度が高くなっていることが確認できた．一部の類似度が低い組み合わせについては，「考えようによっては類似していない」と納得できるものと，納得できないものがあった．後者については問題文が短いという特徴がみられ，単語の共起をもとにトピックを推定する LDA の苦手とする対象であることが原因として考えられる．

この実験の結果， 適切な処置を行えば LDA で算出したトピック分布θの非類似度を項目間の非類似度指標として使用しうると判断し，続いてトピック分布θの非類似度をもとにした項目のグルーピングを行うことにした．

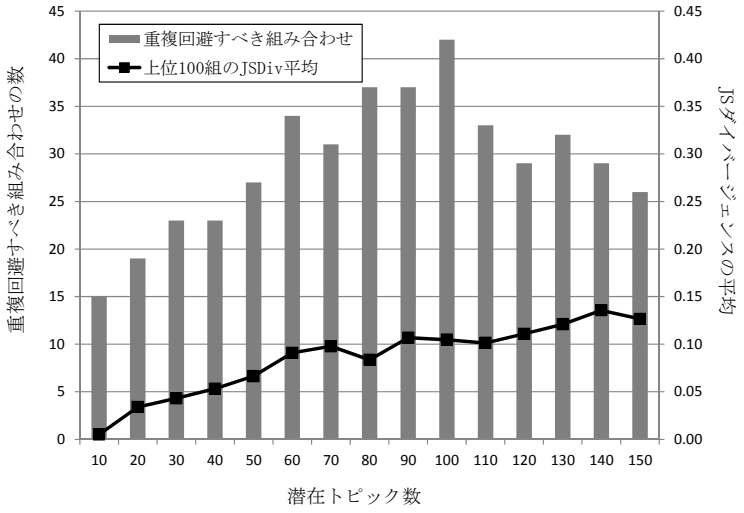

図 1. JS ダイバージェンスと重複回避すべき組み合わせの数（上位 100 組中）

3. 類似度をもとにした項目のグルーピング

前項で算出した各項目間の疑似距離(JS ダイバージェンス)を用い，非階層クラスタリング(k-medoids 法)を試みる．クラスタ数とトピック数の組み合わせをさまざまに変えて実験(2)を行った(表 2).

表 2. 実験(2)概要

対象項目	実験(1)で使用した項目
距離の指標	実験(1)で算出したJSダイバージェンス (トピック数：10，20…，150の15パターン) ※「分類できないトピック」を除外して計算
クラスタリング手法	k-medoids法
クラスタ数	50，100…，450，500の10パターン
評価の観点	F値(適合率と再現率の調和平均)

表 3. F 値とクラスタ数・トピック数 の関係（抜粋）

トピック数 / クラスタ数	50	100	150
100	0.306	0.328	0.321
200	0.402	0.432	0.422
300	0.485	0.507	0.493
400	0.530	0.557	0.539
500	0.575	0.589	0.586

実験結果の一部を表 3 に掲載する．この結果から，クラスタ数を増やすことで F 値は改善するが，F 値そのものはあまり高くないことがわかった．クラスタ数を増やしていくと，項目数が 1 のクラスタが増え，その結果 626 項目中 191 項目ある単独項目(項目管理者が”類似項目がない”と判断した項目)に合致する数が増加するため F 値が改善すると考えられる．また，今回の実験ではトピック数による違いは確認できなかった．

単独項目が多く，またグループ数もかなり多いため項目のグルーピングは通常のクラスタ分析にはなじまない可能性がある．また，項目管理者が実際にグルーピングを行う際には，「似ているかどうか」の判断とともに「テーマ A での類似よりテーマ B での類似の方を回避したい」といった選択もしているため，類似度のみで自動的にグルーピングを行うことには一定の限界がある可能性もある．

4. まとめ

今回の検証を通して， LDA を使った項目間類似度は言語能力を主題とした項目の類似度の指標として，適切な処置をおこなえば使用しうるとの知見を得た．問題文が短い項目については項目管理者の判断と比較してあまり良好な結果が得られなかったが，これについては twitter など短い文章に対応したトピックモデルである Biterm Topic Model[5]の適用可能性を今後検討してみたい．

一方，項目間類似度をもとにしたクラスター分析によるグルーピングについては課題が残った．今後は，教師あり学習をおこなうトピックモデルである Labeled LDA[6]を適用することにより，既存のグルーピングを教師データとして学習し，それをもとに追加項目を分類するような方法について有効性を確認したい．また，項目間の類似関係を可視化することで項目管理者がおこなうグルーピング作業の負荷を軽減するようなアプローチについても今後の検討課題である．

参考文献

[1] Pokpong Songmuang, Maomi Ueno, Keizo Nagaoka : Development of Automated e-Testing Construction System with Redundant Item, The 8th International Conference on eLearning for Knowledge-Based Society (2012)

[2] David M. Blei, Andrew Y. Ng, Michael I. Jordan : Latent Dirichlet Allocation, J. Mach. Learn. Res., Vol. 3, pp. 993-1022 (2003).

[3] 高木輝彦，高木正則，勅使河原可海，田中健次：e テスティングにおける LDA を用いた項目間類似度の算出　情報処理学会論文誌,Vol.55, No.1, pp.91-104 (2014)

[4] Leonhard Hennig,DAI Labor, TU Berlin : Topic-based Multi-Document Summarization with Probabilistic Latent Semantic Analysis(2003)

[5] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng : A Biterm Topic Model for Short Texts(2013)

[6] Daniel Ramage, David Hall, Ramesh Nallapati and Christopher D. Manning : Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora(2009)