e テスティングにおける LDA を用いた項目の類似度算出手法

A Method of Calculating Similarity between Items Using Latent Dirichlet Allocation in E-Testing

高木 輝彦[*1], 高木 正則[*2], 勅使河原 可海[*3], 植野 真臣[*1]
Teruhiko TAKAGI[*1], Masanori TAKAGI[*2], Yoshimi TESHIGAWARA[*3], Maomi UENO[*1]
[*1] 電気通信大学大学院情報システム学研究科
[*1]Graduate School of Information Systems, the University of Electro-Communications
[*2] 岩手県立大学ソフトウェア情報学部
[*2]Faculty of Software and Information Science, Iwate Prefectural University
[*3] 創価大学大学院工学研究科
[*3]Graduate School of Engineering, Soka University
Email: takagi@ai.is.uec.ac.jp

あらまし：本研究では，e テスティングにおいてアイテム・バンク内の項目を有効利用するために，項目の類似度算出手法を提案する．類似度を用いることで，(1) 自動的なアイテム・バンクの構築，(2) 項目間構造の可視化，(3) 項目の難易度の推定などに応用できる．本提案は類似度を算出する過程で，文書の生成過程を確率的にモデル化した LDA を適用し，従来手法に比べ類似度算出の精度の向上が示された．
キーワード：e テスティング，項目の再利用，類似項目，LDA，類似度

1. 研究の背景と目的

近年，e テスティング (e-testing)[(1)]の出現により，大規模なアイテム・バンク (item bank)の構築が必要となっている．アイテム・バンクでは，項目の正答率や難易度などのメタ・データを Web 上で管理できる．これらのデータを利用して項目を管理し，テストの出題や構成などへ再利用する研究が多数行われている[(2)]．しかし，メタ・データの収集には，予め多くの被験者による解答データや人手による労力が必要となる．そこで，本研究では解答データや人手による労力に依存しない項目メタ・データの収集を目的とし，項目の類似度算出手法を提案する．類似度データを用いることで，(1) 自動的なアイテム・バンクの構築，(2) 項目間構造の可視化，(3) 項目の難易度の推定[(3)]などに応用することができる．

本研究では，専門用語の理解度を問う多枝選択式の項目を対象とし，ベクトル空間モデル[(4)]に基づき項目間の類似度を算出する．具体的には，文書の生成過程を確率的にモデル化した Latent Dirichlet Allocation (LDA)[(5)]によって推定されたトピックを特徴量としたベクトルで項目を表現する．LDA では，対象文書中で出現する単語間の共起関係に基づき対象文書のトピックを 1 つ，または，複数推定する．LDA を適用することで，(a) 項目の内容理解に踏み込んだ特徴量の抽出，(b) 不要単語や単語の表記ゆれによって生じるノイズの減少，などの利点が挙げられる．情報検索や文書分類などの分野で LDA を応用した研究が多数存在するが[(6)]，項目への LDA の適用例は全く存在しない．項目のテキスト情報は問題文，正答，誤答から構成されており，既存研究で対象としている文書とはその性質が異なるため，項目の特徴をとらえた適用方法を考案する必要がある．

2. Latent Dirichlet Allocation

LDA は各トピックの多項分布 $Mult(\boldsymbol{\theta})$がその共役事前分布であるディリクレ分布 $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})$に従うと仮定した文書生成モデルである．以下に，LDA による文書生成過程を示す．以下の過程を文書数 M 回繰り返して文書集合 D が生成される．

1. 文書中から単語 N をサンプリング．
2. ディリクレ事前分布 $Dir(\boldsymbol{\theta}|\boldsymbol{\alpha})$から各トピックの生成確率 $\boldsymbol{\theta}$ をサンプリング．
3. 各 N 個の単語 w_n に対して，
   (a) 多項分布 $Mult(\boldsymbol{\theta})$から一つのトピック z_n をサンプリングする．
   (b) トピック z_n で条件付けられた多項確率 $p(w_n|z_n,\boldsymbol{\beta})$から単語 w_n をサンプリングする．

なお，モデルパラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ の学習には変分ベイズ法[(5)]を用いる．

3. 類似項目の定義

本研究では，類似項目を「項目で問われている知識や解決の中心となる知識が一致する項目」と定義する．この知識とは専門用語である（以下，対象知識）. 情報技術に関する項目 1404 問から 1687 個の対象知識を抽出したところ, 単名詞, 複合名詞, また, それらが日本語，英語，日本語+英語（複合名詞のみ）の 5 種類の単位に分類できることが分かった．

4. LDA を用いた類似度算出手法

LDA によるトピックの推定では, 対象とする項目中に出現する単語の共起行列が必要となり，対象知識，または，これに関連する単語の共起関係を反映させた共起行列の作成が望まれる．本研究では，(a) 対象知識の出現箇所（問題文，正答，誤答）を自動で決定し，(b) 限られた単語の共起関係からトピックを推定する，というアプローチをとる．

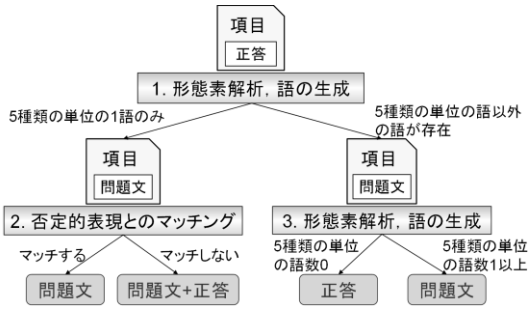

図 1 対象知識出現箇所の自動決定手順

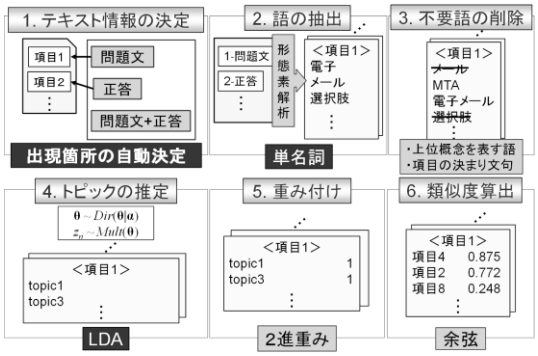

図 2 LDA を用いた類似度算出手順

4.1 対象知識出現箇所の自動決定

アプローチ(a)に対して，問題文の問い方や正答に出現する単語の特徴を字句解析し，対象知識の出現箇所を自動で決定する．図 1 に対象知識出現箇所の自動決定手順を示す. まず, 正答を形態素解析[4]し，3 章で述べた 5 種類の単位の語を抽出する（図 1: 1）．このとき，抽出された語数が 1 の場合，問題文と予め登録されている否定的な表現とのマッチングを調査する（図 1: 2）．マッチする場合は問題文，マッチしない場合は問題文と正答に決定される．一方，抽出された語数が複数存在する場合（図 1: 1），問題文を形態素解析し，5 種類の単位の語を抽出する（図 1: 3）．抽出された語数が 0 の場合は正答，1 以上の場合は問題文に決定される．この手順に基づき 3 章で対象とした項目を分析したところ，99.6%の確率で対象知識の出現箇所を特定することができた．

4.2 LDA を用いた類似度算出手法

図 2 に LDA による類似度算出手順を示す. まず，図 1 の手順に従い対象知識の出現箇所を決定する（図 2: 1）．次に，決定された出現箇所を形態素解析し，単名詞を抽出する（図 2: 2）．ここでは，アプローチ(b)に対して，複合名詞となる単語も全て単名詞に分割する. そして, 不要語を削除した後（図 2: 3），各項目を単語の共起行列で表し, LDA によりトピックを推定する（図 2: 4）．さらに，推定されたトピックに対して 2 進重み[4]により重み付けを行なう（図 2: 5）．最後に，重みを要素とするベクトルで表された項目間の類似度を余弦[4]により算出する（図 2: 6）．

5. 類似項目検索実験

提案手法による類似度算出の精度の向上を検証するために，類似項目の検索実験を行った．情報技術に関する項目 250 問を対象とし，項目間の類似度を算出し，項目ごとに類似度の高い項目を抽出し検索結果とした．この結果から，再現率，適合率，F 尺度[4]を算出した．比較手法としては，専門用語の抽出を目的とした termmi[7]（手法 1）とベクトル空間モデルに TFIDF[4]を適用した手法（手法 2）を用いた．また，提案手法において，対象知識の出現箇所の自動決定を行わない手法（手法 3）を用いた．表 1 に実験結果を示す．提案手法では手法 1~3 に比べ精度が向上していることが分かる. また, 手法 3 では，手法 1，2 と精度が同程度であり，提案手法では，対象知識の出現箇所を自動で決定することで精度が向上したと考えられる．以上のように，情報技術分野の項目において提案手法の有効性が示唆された．

表 1 各手法による再現率，適合率，F 尺度

類似度算出手法	再現率	適合率	F 尺度
提案手法	0.615	0.350	0.446
手法 1	0.593	0.298	0.397
手法 2	0.556	0.281	0.374
手法 3	0.543	0.308	0.394

6. まとめと今後の課題

本稿では，解答データや人手による労力に依存しない項目メタ・データの収集を目的とし，項目間の類似度算出手法を提案した．具体的には，ベクトル空間モデルに LDA で推定されたトピックを適用するために，対象知識出現箇所の自動決定手順を考案した．実験結果から，従来手法と比べ類似度算出の精度の向上と，LDA に対象知識出現箇所の自動決定手順を適用することの有効性が示唆された．

今後は，LDA で推定されたトピックに対する重み付け手法を検討する．また，情報技術分野以外の項目への適用を試みる．さらに，1 章で述べた類似度データの応用例(1)~(3)についての検討を行い，類似項目を利用した適応型テストや作問時に類似項目を提示することによる作問支援などへ発展させる．

参考文献

(1) 植野真臣, 永岡慶三:“e テスティング”，培風館(2009)
(2) Songmuang, P and Ueno, M.:“Bees Algorithm for Construction of Multiple Test Forms in E-Testing”, IEEE Trans. Learning Technologies., vol.4, No.3, pp.209-221, Nov (2011)
(3) Ikeda, S., Takagi, T. and Takagi, M. et al.:“A Study on a Method of Estimating the Difficulty of Quizzes Focused on Quiz Types”, Proceedings of ICCE2011, pp.312 – 316 (2011)
(4) Manning, C. D. and Schutze, H.:“Fundamentals of Statistical Natural Language Processing”, MIT Press, Cambridge, MA (1999)
(5) Blei, D.M., Ng, A.Y. and Jordan, M.I.:“Latent Dirichlet allocation, Journal of Machine Learning Research”, Vol.3, pp.993-1022 (2003)
(6) 上田修功, 斎藤和己:“類似テキスト検索のための多重トピックテキストモデル”，情報処理学会論文誌. 数理モデル化と応用，Vol.44, No.SIG14(TOM9), pp.1-8 (2003)
(7) Text Mining Tool for Windows, “termmi”: http://gensen.dl.itc.u-tokyo.ac.jp/