LDA を用いた類似項目検索のための前処理法

高木 輝彦[1,2,a)]　高木 正則[3]　勅使河原 可海[4]

概要：本研究ではこれまで，多肢選択式項目を対象とした類似項目の自動検索を目的とし，項目間類似度の算出手法を提案してきた．類似度データや類似項目を用いることで，(1) 類似項目の自動検索，(2) 自動的なアイテム・バンクの構築，(3) 項目間構造の可視化，(4) テスト情報量の向上，(5) 新規項目の難易度の推定，(6) 適応的なテストの出題，(7) 項目の作成，などの支援が可能となる．これまでの手法は，出現する語を特徴量としたベクトルの生成（発見的アプローチ）と，語の共起性から確率的に推定されるトピックを特徴量としたベクトルの生成（確率的アプローチ），の大きく 2 つのアプローチに分類され，確率的アプローチは発見的アプローチの不要語による誤検索や表記ゆれによる検索漏れなどの課題解決が示唆された．トピックの推定では，代表的なトピックモデルである Latent Dirichlet Allocation（LDA）を用いる．本論文では，類似項目の検索精度の向上を目的とし，LDA を用いた類似項目検索のための前処理法を提案する．多肢選択式の 250 項目を対象とした類似項目の検索実験を行った結果，提案手法では，既存手法に比べ類似項目の検索精度が最も高かった．これらの実験・評価結果から，前処理法の有効性が示唆された．

キーワード：e テスティング，アイテム・バンク，類似項目，Latent Dirichlet Allocation（LDA），類似度

A Method of Preprocessing for Retrieving Similar Test Items Using Latent Dirichlet Allocation

Teruhiko Takagi[1,2,a)]　Masanori Takagi[3]　Yoshimi Teshigawara[4]

Abstract: In previous studies, to retrieve similar test items automatically in e-testing, we proposed methods of calculating similarity between test items based on the vector space model, and conducted experiments and evaluations. It is possible to apply similarity data or similar test items to (1) automatically retrieving similar test items, (2) automatically constructing item banks, (3) visualizing structure between test items, (4) optimizing the amount of test information, (5) estimating the difficulty level of unanswered test items, (6) computer adaptive testing (CAT), and (7) supporting the creation of test items. Our methods in previous studies are classified into heuristic approach and probabilistic approach. In the heuristic approach, each test item is represented by a vector using extracted terms from these test items. In the probabilistic approach, each test item is represented by a vector using topics probabilistically estimated by the co-occurrence relation between terms. The probabilistic approach can decrease the number of retrieved dissimilar test items caused by noise such as superfluous terms and spelling variations in the heuristic approach. We have applied Latent Dirichlet Allocation (LDA), a generative probabilistic document model, to topic estimation. In this paper, to improve the accuracy of retrieving similar test items, we propose a method of preprocessing for retrieving similar test items using LDA. We targeted 250 multiple-choice test items tested by Systems Administrator Examination and conducted an experiment where similar items are retrieved. The result of the experiment showed the improvement in accuracy of the proposed method in comparison with existing methods, and it proved the effectiveness of the method of preprocessing.

Keywords: e-testing, item bank, similar test item, Latent Dirichlet Allocation (LDA), similarity

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・バンク内の項目は，項目内で問われている知識に基づいて階層的に分類されることが多い．例えば，Songmuang らが開発した e テスティング・システムでは，分野，大分類，中分類の 3 階層によって分類されている [2]．また，Guzman らが開発した自己評価ツール SIETTE（System of Intelligent Evaluation Using Tests for Teleeducation）では，講義内で取り扱うテーマごとに複数階層により項目が管理されている [3]．これらのシステムでは，適応型テスト（CAT: Computer Adaptive Testing）や自動テスト構成 [4] などの機能が実装されており，項目には解答所要時間や正答率，項目反応理論（IRT: Item Response Theory）[5] のパラメータなどのメタ・データが付与されている．さらに，近年，教育現場で広く普及している moodle [6] や Blackboard [7] などの LMS（Learning Management System）においても同様に，階層的に項目が管理されている．

しかし，項目数が増大すると，項目を適切な階層に手動で分類するには膨大な時間と労力がかかる．また，項目を適切な階層へ分類するためには，対象とする分野や領域の専門知識に精通していなければならない．そこで，本研究ではこれまで，多肢選択式項目を対象とした類似項目の自動検索を目的とし，項目間類似度の算出手法を提案し，実験・評価を行ってきた [8–12]．類似項目とは，「項目内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，分野特有の専門用語であり，大きく単名詞と複合名詞に分類されることが分かっている [10]（以下，対象知識と呼ぶ）．

類似度データを用いることで，以下に示すタスクを支援することができる．

(1) アイテム・バンク内，または，アイテム・バンク間における類似項目の自動検索 [13].
(2) 多次元尺度法（MDS: Multi Dimensional Scaling）[14] などによる項目間構造の可視化.
(3) 類似度の低い項目集合を抽出し，テスト構成することによるテスト情報量 [5] の向上.

[1] 電気通信大学大学院情報システム学研究科
Graduate School of Information Systems, University of Electro-Communications
[2] 日本学術振興会特別研究員 DC2
JSPS Research Fellow
[3] 岩手県立大学ソフトウェア情報学部
Faculty of Software and Information Science, Iwate Prefectural University
[4] 東京電機大学未来科学部
School of Science and Technology for Future Life, Tokyo Denki University
[a)] takagi@tanaka.is.uec.ac.jp

さらに，類似項目の自動検索が可能となり，類似項目を用いることで以下に示すタスクを支援することができる．

(1) クラスタリング手法 [15] などを適用した，自動的なアイテム・バンクの構築.
(2) 類似項目の出題パターンを基に，解答データの無い新規項目の難易度を推定 [16].
(3) 同一知識を問う項目の反復学習における適応型テスト [17].
(4) 類似項目の選択肢を提示することによる項目の作成 [10].

これまでの手法は，ベクトル空間モデル [15] に従い，出現する語を特徴量としたベクトルの生成（発見的アプローチ）[8–10] と，語の共起性から確率的に推定されるトピックを特徴量としたベクトルの生成（確率的アプローチ）[11] [12] の大きく 2 つのアプローチに分類される．実験・評価結果から，確率的アプローチは発見的アプローチに比べ，類似項目の検索精度の向上が示され，発見的アプローチの（1）不要語による誤検索や，(2) 表記ゆれによる検索漏れ，の課題を解決できることが示唆された．

本研究では，トピックの推定において，代表的なトピックモデルである Latent Dirichlet Allocation（LDA）[18] を用いる．LDA は，対象文書や文書集合内から抽出される語の共起性に基づき，対象文書の内容を表すトピックを 1 つ，または，複数推定する．すなわち，トピックの推定精度は，抽出される語の共起性に左右されるため，文書中で抽出する語を決定するための前処理が必要となる．そこで，本論文では，類似項目の検索精度の向上を目的とし，LDA を用いた類似項目検索のための前処理法を検討する．

前処理法が確立されることで，(1) トピック推定精度の向上による類似項目検索精度の向上や，(2) 語の共起性を基にした様々なトピックモデルへの対応，が期待される．

2. Latent Dirichlet Allocation

トピックモデルでは，文書が複数のトピックの混合分布として，各トピックが対象文書集合中に出現する語の分布として表現される．LDA は，各トピックの多項分布 $Mult(\boldsymbol{\theta})$ が多項分布の共役事前分布であるディリクレ分布 $Dir(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$ に従うと仮定した文書の生成モデルである．LDA のグラフィカルモデル表現を図 1 に示す．グラフィカルモデルとは，確率変数またはパラメータを頂点とし，それらの依存関係を有向グラフで表現したものである．黒丸の頂点は観測変数，それ以外の頂点は潜在変数または未知パラメータを示す．矩形部分は，その隅に示された回数だけサンプリングが繰り返されることを意味する．ただし，N は語数，M は文書数を示す．また，図 1 に示された LDA のグラフィカルモデル表現に対応する文書生成過程を書き下すと，以下のようになる．以下の過程を M 回繰り返して文書集合 D が生成される．

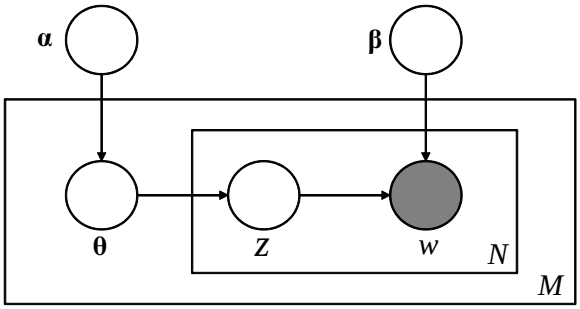

図 1 LDA のグラフィカルモデル表現 [18]
Fig. 1 Graphical model representation of LDA.

(1) 文書中から N をサンプリング.
(2) ディリクレ事前分布 $Dir(\boldsymbol{\theta} \mid \boldsymbol{\alpha})$ 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 $p(w_n \mid z_n, \boldsymbol{\beta})$ から単語 z_n をサンプリングする.

LDA では，語 N を生成する過程で潜在変数 z の値が $\boldsymbol{\theta}$ により確率的に変化し，一つの文書中で複数のトピックが生成されることが分かる．また，モデルパラメータ $\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ が文書の矩形の外にあるため，新たな文書も生成可能であることが分かる．なお，ディリクレ分布の次元 k（トピック変数 z の次元）は既知であり，固定されるものと仮定される．$\boldsymbol{\alpha}$ は要素 $\boldsymbol{\alpha}_i > 0$ をもつ k 次元ベクトルで，ディリクレ分布

$$p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\Sigma_{i=1}^{k} \alpha_i)}{\Pi_{i=1}^{k}(\boldsymbol{\alpha})} \theta_1^{\alpha_1 - 1} \cdots \theta_1^{\alpha_k - 1}, \tag{1}$$

のパラメータである．$\Gamma()$ はガンマ関数である．さらに，トピックに対する語の生成確率は $p(w^j = 1 \mid z^i = 1) = \boldsymbol{\beta}_{ij}$ である $k \times V$ 行列 $\boldsymbol{\beta}$ によって表される．

パラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ が与えられたとき，トピック混合 $\boldsymbol{\theta}$ の同時分布と N 個のトピック $\mathbf{z}$ の集合と N 個の単語 $\mathbf{w}$ の集合は以下のように表される．

$$p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \prod_{n=1}^{N} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta}) \tag{2}$$

$\boldsymbol{\theta}$ に関して積分し，$\mathbf{z}$ に関して和をとることで文書の生成確率を以下のように得ることができる．

$$p(\mathbf{w} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) = \int p(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n \mid \boldsymbol{\theta}) p(w_n \mid z_n, \boldsymbol{\beta}) \right) d\boldsymbol{\theta} \tag{3}$$

本論文では，モデルパラメータ $\boldsymbol{\alpha}$ と $\boldsymbol{\beta}$ の学習には変分ベイズ法 [18] による近似計算を用いる．

近年，LDA をはじめとするトピックモデルの研究が盛んに行われており，文書検索や文書分類に応用されている [19–21]．これらの研究において実験で対象とされている文書は，TREC [22]（文献 [20]）や GENIA [23]（文献 [20]），NIPS [24]（文献 [21]）などがフリーで提供しているコーパスに含まれるニュース記事や学術論文などである．また，Yahoo.com ドメインなどの Web 上で公開されているニュース記事（文献 [19]）も対象とされている．これらの文書集合は，本研究で対象としている項目とは異なり，対象となる文書数や文書中に含まれる語数が多い．文書数は少なくとも 1,000 件以上存在し，語の総数は 10,000 語を超えるものがほとんどである．このような文書集合では，トピック数を設定することが困難であるため，いくつかの場合（50，100 など）で精度が最も良いものを採用している．また，扱う語数や文書数が多いことから，それらを変数とする数式の改良や計算時間を短縮するための推論方法などに焦点化されることが多い．

しかしながら，本研究で対象とする多肢選択式の項目は，近年の e テスティング化に伴い項目数は増えたが，項目内で問われる知識や各フォーム内での言い回しが限られるため，上述したような文書に比べ語の総数は少ない．(5 章の実験で対象とした 250 項目に含まれる語の総数は，最も多くて約 2000 語）このような理由から，上述した研究で提案された手法が項目に対して適応可能かどうかは明らかではない．本論文は，項目に LDA のような隠れ変数を扱うトピックモデルを適用可能にするための一つのアプローチであり，項目の特徴を捉えた適用方法を考案するものである．

3. 先行研究

本研究では，先行研究において，対象知識が出現する個所（問題文，正答，誤答）を自動で決定（以下，知識出現箇所の自動決定）する手順を考案し，それに基づく項目間類似度の算出手順を提案した [10]．本章では，それらの手順について説明する．

3.1 知識出現箇所の自動決定

図 2 に知識出現箇所の自動決定手順を示す．ある項目に対して，問題文と正答を字句解析することによって，対象知識の出現箇所を自動で決定する．図中の 5 種類の単位の語とは，対象知識の語の単位を表す．対象知識は単名詞と複合名詞の大きく 2 つに分けられ，さらに，それらが日本語か英語か，または，日本語と英語の組み合わせによって 5 種類の単位に分類されることが分かっている [10]．表 1 に対象知識の種類と例を示す．なお，文章の形態素解析には「茶筌」[25] を使用し，複合名詞への連結処理は，中川らによって作成された専門用語自動抽出用 Perl モジュール “TermExtract” [26] を使用している（次節や 4.2 節の手順においても同様）．以下，知識出現箇所の自動決定手順

表 1　対象知識の種類と例
Table 1　Targeted Knowledge Type and Examples.

対象知識の種類		例
単名詞	日本語	ルータ
	英語	Telnet, UDP
複合名詞	日本語	電子的コミュニケーション
	英語	TCPIP, BtoB
	日本語と英語	OSI 参照モデル

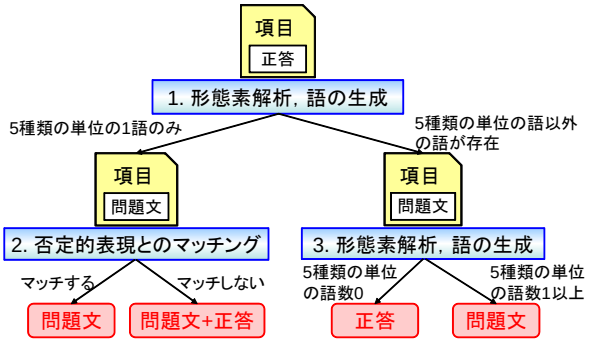

図 2　知識出現箇所の自動決定手順 [10]
Fig. 2　Procedure of automatically identifying the part of targeted knowledge occurrence.

について説明する．

まず，正答を形態素解析し，5 種類の単位の語を抽出する（図 2: 1）．抽出された語数が 1 かつその他の語（助詞や助動詞など）が存在しない場合，問題文と予め登録されている否定的な表現と一致するかどうかを解析する（図 2: 2）．否定的な表現とは，「誤っている〜」や「間違った〜」など，問題文の内容と異なる概念や専門用語などを選択させるときに使用される表現である．これらの否定的な表現と一致する場合，知識出現箇所は問題文に決定される．また，一致しない場合，知識出現箇所は問題文と正答に決定される．一方，正答を形態素解析した結果，5 種類の単位の語以外の語が存在する場合（図 2: 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 2: 3）．抽出された語数が 0 の場合，知識出現箇所は正答に決定される．また，語が抽出された場合，知識出現箇所は問題文に決定される．

3.2　知識出現箇所の自動決定に基づく項目間類似度の算出

図 3 に，前節で説明した知識出現箇所の自動決定に基づく項目間類似度の算出手順を示す．まず，各項目において対象知識の出現箇所を前節の手順により自動で決定する（図 3: 1）．次に，項目ごとに決定された出現箇所を形態素解析し，表 1 の 5 種類の単位で語を抽出する（図 3: 2）．続いて，不要語の削除を行う（図 3: 3）．不要語の対象としては，上位概念を表す語と項目の決まり文句に含まれる語としている．項目内では，対象知識に対して言い換え表現による上位概念を表す語 [27] が存在することが多い．（例

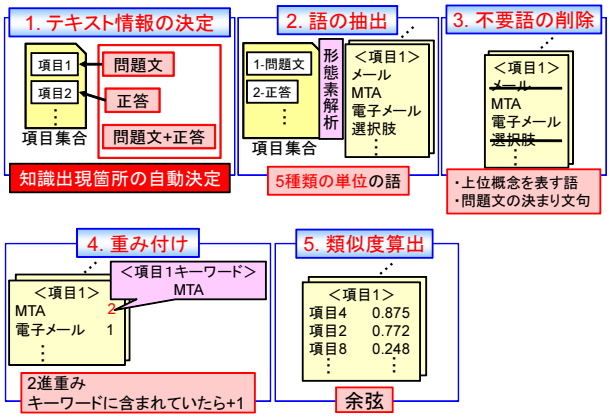

図 3　知識出現箇所の自動決定に基づく類似度算出手順 [10]
Fig. 3　Procedure of calculating similarity between test items based on procedure Fig. 2.

えば，「コンピュータネットワーク」に対して「コンピュータ」，「電子メール」に対して「メール」）これらの語は，広域な意味を持ち，また，類似項目でない項目においても出現すると考えられる．そこで，対象知識のようなより狭義な意味を持つ下位概念の語を抽出するために，ある複合名詞に対して，その複合名詞に含まれる単名詞（あるいは，複合名詞）が同じ項目内に存在する場合，それらの語を削除する．これらの判定は，Perl の正規表現によるマッチングにより行っている．また，項目の決まり文句に含まれる語（次，以下，選択肢など）は項目の内容とは全く関係がない．

さらに，抽出された語に対して重み付けを行う（図 3: 4）．語の出現頻度や出現分布と語の重要度の関連性が明らかになっていないため，それらの情報を使用しない最も単純な手法として 2 進重み [15] を適用している．これは，抽出されたすべての語に対して重み 1 を付与する．また，抽出された語の中でキーワードにも含まれる語が存在する場合，さらに重み 1 を加える．最後に余弦 [15] により類似度を算出する（図 3: 5）．

x_i，y_i をそれぞれ文書 d_x，d_y から抽出された語 i に対する重み，あるいは語の存在を表す 2 値変数，ここで，T を抽出された語の総数とすると，余弦による類似度 (d_x, d_y) は次のような式で表わされる．

$$\sigma(d_x, d_y) = \frac{\Sigma_{i=1}^{T} x_i \cdot y_i}{\sqrt{\Sigma_{i=1}^{T} x_i^2 \times \Sigma_{i=1}^{T} y_i^2}} \quad (4)$$

4.　LDA を用いた項目間類似度の算出

4.1　前処理法の検討

LDA によるトピックの推定では，対象とする項目や項目集合に出現する語の共起行列が必要となり，その推定精度は，それらの語の共起性に左右される．以下に，異なる内容の項目にも関わらず，同じトピックが推定された項目の例を示す．項目 1 の対象知識は「著作物」，項目 2 の対象知識は「スワッピング」である．しかし，項目 1 の問題

＜項目 1 ＞

個人の著作物の保護期間が終了するのは，著作者の死後何年経過したときか．

（1）25 年　（2）50 年　（3）75 年　（4）100 年

＜項目 2 ＞

仮想記憶装置をもつサーバで新しいプログラムを追加して実行したところ，スワッピングが発生し，以前から動作しているプログラムの処理効率が低下した．解決策として，最も適切なものはどれか．

（1）高度な CPU に変更する．（2）高度な主記憶に変更する．（3）磁気ディスク装置を増設し，補助記憶の容量を拡大する．（4） 主記憶を増設する．

文中の「著作物」と項目 2 の問題文中の「プログラム」が，他の複数の項目（誤答）において共に出現していたため，「著作物」と「プログラム」の共起性が高まり，これら二つの項目で同じトピックが推定された．

項目ごとに項目の内容を表した適切なトピックを推定するためには，対象知識，または，これに関連する語の共起性を適切に反映させた共起行列の作成が望まれる．しかし，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，必ずしもすべてのテキスト情報に対象知識やこれに関連する語が出現するとは限らない．そこで，項目ごとに LDA でトピックを推定する際に，(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) 限られた語の共起性を高める，という前処理を行う．前処理 (a) に対して，本研究ではこれまでに，3.1 節で述べた知識出現箇所の自動決定手順を適用した手法を提案した [11] [12]．この処理を行うことで，項目の内容とは関係のない語を大幅に減らすことができ，類似項目検索の精度向上に有効的であることが示唆された．

前処理 (b) では，決定された個所から対象知識とそれらに関連する語を多く抽出することが望まれる．先行研究 [11] では，表 1 に示したように，対象知識は単名詞と複合名詞の大きく 2 つに分類されることから，単名詞と複合名詞を抽出した．先行研究 [12] では，複合名詞では，それらを構成する単名詞間の意味的な関連性が高いことから，複合名詞を構成する単名詞とその他の単名詞を抽出した．その結果，先行研究 [11] の手法と比較し，類似度算出の精度の向上が示された．

しかしながら，複合名詞を構成する単名詞は，それらの複合名詞を包括した意味を表す事が多く（例えば，電子メールと電子，メールや TCP プロトコルと TCP，プロトコルなど），これらの複合名詞と複合名詞を構成する単名詞間の意味的な関連性も高いと考えられる．そこで，前処理 (b) に対しては，決定された出現箇所に出現する単名詞と複合名詞，また，複合名詞を構成する単名詞の共起性を生み出す処理を行う．例えば，項目内で「磁気ディスク」と「コンピュータ」が出現した場合，抽出する語は「磁気ディスク／磁気／ディスク／装置／コンピュータ」とする．この処理を行うことで，類似項目間で同じ意味を表すが表記の異なる語（表記ゆれ）が出現したとしても（例えば，一方の項目では「磁気ディスク装置」，もう一方の項目では「磁気ディスク」），これらの語の共起性を利用すれば同一のトピックが推定されると考えられる．

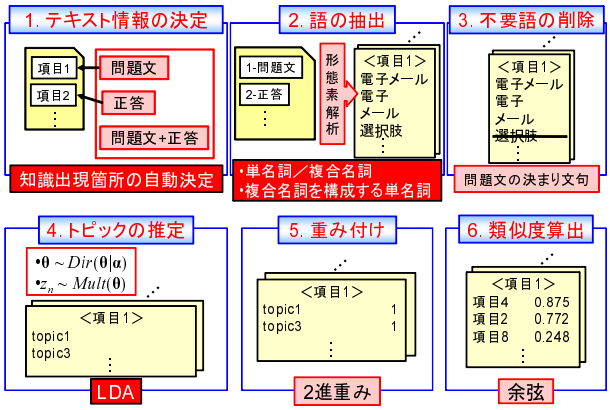

図 4　LDA を用いた類似度算出手順

Fig. 4　Procedure of calculating test item similarity using LDA.

4.2　LDA を用いた項目間類似度の算出

図 4 に LDA を用いた項目間類似度の算出手順を示す．まず，3.1 節の手順に従い対象知識の出現箇所を自動決定する（図 4: 1）．次に，決定された出現箇所を形態素解析し，単名詞と複合名詞，また，複合名詞を構成する単名詞を抽出する（図 4: 2）．そして，項目の決まり文句に含まれる語を不要語として削除する（図 4: 3）．

ここで，語と語の出現頻度を基に，各項目を共起行列で表し，LDA によりトピックを推定する（図 4: 4）．LDA では，ディリクレ分布の次元 k（対象とする項目集合のトピック数）は既知であり，固定されるものと仮定される．現段階では，トピック数の適切な値が分かっていないため，予め，項目を類似項目群へ手動で分類し，分類された類似項目群の数を k として与える．これにより，精度の向上が示されれば，今後，クラスリング手法 [15] などを用いることで，自動的に類似項目群を作成し，その数をトピック数として設定することも可能となる．さらに，推定されたトピックに対して 2 進重みにより重み付けを行なう（図 4: 5）．最後に，トピックの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（式 (4)）により推定する（図 4: 6）．

5.　実験と評価

5.1　実験概要

本実験では，提案手法による類似項目の検索精度の向上を示すと共に，(1) 単名詞／複合名詞／複合名詞を構成す

表 2 実験で利用した LDA を用いた項目間類似度の算出手法
Table 2 Overview of Methods Using LDA in Experiment.

手法	概要
提案手法	前章で提案した手法（図 4）.
LDA 1	提案手法の語の抽出（図 4: 2）において，単名詞／複合名詞を構成する単名詞を抽出する手法.
LDA 2	提案手法の語の抽出（図 4: 2）において，単名詞／複合名詞を抽出する手法.
先行研究による手法	3.2 節で述べた知識出現箇所の自動決定に基づく項目間類似度の算出手法.
termmi による手法	問題文／正答／誤答から語を抽出．語の出現頻度と連接頻度による重み付け．余弦による項目間類似度の算出.
TFIDF による手法	問題文／正答／誤答から語を抽出．TFIDF により重み付け．余弦による項目間類似度の算出.

る単名詞を抽出することの有効性（図 4: 2），(2) LDA を適用することの有効性（図 4: 4）を検証する．これらの有効性を検証するために，提案手法とその他の 5 つの手法により類似項目の検索実験を行った．表 2 に実験で用いた項目間類似度の算出手法とその概要を示す．LDA 1 は，提案手法の語の抽出（図 4: 2）において，単名詞と複合名詞を構成する単名詞を抽出し（先行研究 [11] の手法），LDA 2 は単名詞と複合名詞を抽出する（先行研究 [12] の手法）．その他の条件は提案手法と同じである．

termmi [28] による手法と TFIDF [15] による手法について説明する．両手法とも，問題文，正答，誤答から語を抽出し，単名詞，複合名詞を抽出する．そして，重み付けを行い，最後に，余弦（式 (4)）により類似度を算出する．重み付けでは，termmi による手法では，語の出現頻度と連接頻度に基づく手法 [29]，TFIDF による手法では，TFIDF を用いる．前者の連接頻度とは，ある単名詞が複合名詞を形成するために連接する名詞の頻度である．単名詞 $N_1, N_2, \ldots, N_L$ がこの順で連接した複合名詞を CN とすると，連接する頻度による重み付けの式は以下の $LR(CN)$ で定義される．

$$LR(CN) = (\prod_{i=1}^{L}(LN(N_i)+1)(RN(N_i)+1))^{\frac{1}{2L}} \quad (5)$$

このとき，$LN(N_i)$ は N_i の左方に連接した全単名詞数で，$RN(N_i)$ は N_i の右方に連接した全単名詞数である．CN の出現頻度を $f(CN)$ とすると，最終的な CN の重み $FLR(CN)$ は次式で表わされる．

$$FLR(CN) = f(CN) \times LR(CN) \quad (6)$$

ただし，$f(CN)$ は次に述べる索引語頻度（TF）とは違い，対象とする語が，ある複合名詞に含まれていたとしてもカウントはしない．

後者の TFIDF は，語の頻度（TF（Term Frequency））や IDF（Inverse Document Frequency）を掛け合わせたものであり，ある文書において重要な語とは，その文書中に多く出現し（一般性），かつ，他の文書中にはあまり出現しない（希少性）という考えに基づいている．ある語の頻度（TF）の重みを tf，総文書数を N，その語を含む文書数を df とすると TFIDF による重み w は次の式で表わされる．

$$w = tf \times (\log \frac{N}{df} + 1) \quad (7)$$

実験で対象とした項目は，初級システムアドミニストレータ試験 [30] において，2004 年度から 2008 年度までに出題された 655 項目のうち，類似項目が 2 つ以上存在する 250 項目とした．

検索精度の評価は，再現率と適合率のマイクロ平均と F 尺度 [15] を用いた．再現率は「類似項目が漏れなく検索されているか」，適合率は「検索の結果，類似項目だけを得ているか」を評価する指標となっており，両者はトレードオフの関係にある．F 尺度は再現率と適合率の二つの値の調和平均である．ゆえに，F 尺度により総合的な類似項目の検索精度を評価する．

ここで，再現率と適合率のマイクロ平均と F 尺度の式を示す．項目 i に対して，類似項目の数を A_i，検索された項目の数を B_i，検索された項目のうち類似項目と同じ項目の数を C_i とすると，再現率，適合率のマイクロ平均 $\bar{R}$，$\bar{P}$，また，F 尺度 $\bar{F}$ の式は以下のように定義される．

$$\bar{R} = \frac{\Sigma_{i=1}^{n} C_i}{\Sigma_{i=1}^{n} A_i} \quad (8)$$

$$\bar{P} = \frac{\Sigma_{i=1}^{n} C_i}{\Sigma_{i=1}^{n} B_i} \quad (9)$$

$$\bar{F} = \frac{1}{\frac{1}{2\bar{P}} + \frac{1}{2\bar{R}}} \quad (10)$$

検索の対象となる項目は，実験で対象とする 250 項目とした．さらに，「類似項目が検索結果の上位に位置しているか」を評価するために，検索結果に閾値を設けた場合の評価も行った．

5.2 類似項目検索実験

実験手順を以下に示す．

（1）項目の類似項目群への分類

筆者により，各項目の対象知識を判定し（クライアントサーバシステム，QR コードなど），同じ対象知識の項目をまとめて類似項目群を作成する．筆者は，基本情報処理技術者試験 [31] の資格を保持しており，初級システムアドミニストレータ試験の知識を十分に習得している．今回の実験では，全部で 62 の対象知識

が特定され，250 項目を 62 の類似項目群へ分類することができた．この分類結果を再現率と適合率のマイクロ平均を算出する際の評価指標とする．

（2）項目間類似度の算出

表 2 に示した手法ごとに，全項目間の類似度を算出する．類似度を算出する対象は，実験で対象とした 250 項目とした．手順（1）により，全部で 62 の類似項目群に分類されることが分かった．そこで，LDA を用いた手法では，ディリクレ分布の次元 k の値（250 項目で推定されるトピック数）は 62 とした．各項目に対して，類似度の高い項目順に整列する．

（3）閾値による項目の抽出

手順（2）の結果を基に，各項目に対して上位 7 項目を抽出する．手順（1）の結果から，ある項目に対する類似項目の数は最大で 7 項目であった．そのため閾値を上位 7 項目と定めた．なお，上位から 7 番目と 8 番目の項目が同じ類似度である場合，8 番目の項目も含めて抽出した．

（4）再現率，適合率，F 尺度の算出

検索結果に閾値なしの場合（手順（2）の結果）と閾値ありの場合（手順（3）の結果）で，各手法の再現率と適合率のマイクロ平均，また，F 尺度を式 (8)，(9)，(10) によりそれぞれ算出する．式 (8)，(9) において，類似項目とは手順（1）によって同じ類似項目群に分類された項目とし，検索された項目とは，閾値なしの場合は手順（2）で類似度が 0 より大きい項目，閾値ありの場合は手順（3）によって抽出された項目とする．

上述した実験手順により得られた再現率と適合率のマイクロ平均，また，F 尺度を表 3 に示す．左の値は，閾値なしの場合で，右の値は閾値を上位 7 項目とした場合である．なお，表中の数値は小数点第三位で四捨五入した値である．表から分かるように，閾値なしの場合と閾値ありの場合の両方において，提案手法の F 尺度が最も高い値となった．

表 3 実験結果（閾値なし | 上位 7 項目）
Table 3 Experimental Result (No Threshold | Top 7 items).

類似度算出手法	再現率		適合率		F 尺度	
提案手法	0.71	0.68	0.33	0.37	0.45	0.48
LDA 1	0.65	0.62	0.28	0.35	0.39	0.45
LDA 2	0.39	0.38	0.19	0.22	0.26	0.28
先行研究による手法	0.73	0.65	0.19	0.33	0.31	0.44
termmi による手法	0.89	0.59	0.06	0.30	0.11	0.40
TFIDF による手法	0.89	0.56	0.06	0.28	0.11	0.37

5.3 実験の考察

前節の実験結果から，提案手法では先行研究による手法や既存手法に比べ F 尺度が高い値を示しており，項目間の検索精度の向上を示すことができた．本節では，実験結果（表 3）を基に，5.1 節で述べた 2 つの有効性について考察する．

まず，提案手法と LDA 1，LDA 2 を比較する．提案手法が再現率と適合率のマイクロ平均，また，F 尺度の値が最も高い．ゆえに，項目に LDA を適用する場合，単名詞と複合名詞，さらに，複合名詞を構成する単名詞で語を抽出することが有効的であると考えられる．次に，提案手法を含めた LDA を用いたすべての手法と LDA を用いていない先行研究による手法や既存手法を比較する．LDA を用いた手法では，閾値なしの場合と閾値ありの場合とで再現率と適合率のマイクロ平均，また，F 尺度の値の変化が少ない．再現率のマイクロ平均は約 0.01〜0.03（減），適合率のマイクロ平均は約 0.03〜0.07（増），F 尺度は約 0.02〜0.06（増）の変化量である．一方，LDA を用いていない手法では，これらの値が大幅に変化していることが分かる．つまり，これらの手法では，類似項目を検索した際，類似項目を上位に検索する可能性はあるが，類似項目でない項目が多く検索される傾向があると考えられる．これは，特徴ベクトルを作成する際に，項目内で出現する語を特徴量とすることが主な原因である．例えば，先行研究による手法のように，予め，知識出現箇所の自動決定によって抽出する語を最小限に絞り込んだとしても，抽出される語数は LDA によって推定されるトピック数に比べて非常に多い．今回の実験では，全体で 605 語が抽出された．つまり，各項目の特徴ベクトルの次元数は 605 次元となり（式 (4)），多くの不要な項目を検索してしまった．一方，LDA を適用した手法では，5.2 節の手順（2）で述べたように，全体で推定されるトピック数は 62 と決定した．つまり，各項目の特徴ベクトルの次元数は与えるトピックの次元数で調整することができる．特に，本節の冒頭で述べたように，提案手法は，抽出される特徴量（次元数）が少なくとも，LDA を用いていない手法に比べて適合率の向上が見られた．このように，項目の特徴を捉えた適切な前処理を行うことで，LDA を適用することの有効性が増すことが示唆された．

以上のように，提案手法では検索精度が向上したものの，適合率の値は 5 割にも満たない（0.37）．これは，ある項目に対して検索された項目のうち，類似項目は平均で 3 割〜4 割程度であるということを意味している．ユーザからのフィードバックなどを用いることができれば，検索結果を修正することは可能だが（適合性フィードバック [15]），1 章で述べた，同一知識を問う項目の反復学習のような逐次的な検索には，実用レベルとは言えない．ゆえに，更なる精度の向上のための手法の改善や，類似項目を応用した場合に対応できる処理を行うことが必要である．

6. まとめと今後の課題

本研究ではこれまで，多肢選択式項目を対象とした類似項目の自動検索を目的とし，項目間類似度の算出手法を提

案してきた．本論文では，類似項目の検索精度の向上を目的とし，LDA を用いた類似項目検索のための前処理法を提案した．具体的には，項目の内容を表す語，または，これに関連する語の共起性を高めるために，それらの語を単名詞と複合名詞，また，複合名詞を構成する単名詞として抽出する処理を行った．そして，これらの前処理を基に，LDA を用いた項目間類似度の算出手法を提案した．さらに，初級システムアドミニストレータ試験で出題された 250 項目を対象とし，類似項目の検索実験を行った．提案手法では，LDA を用いた比較手法や既存手法に比べ検索精度が最も高くなっており，(1) 単名詞／複合名詞／複合名詞を構成する単名詞を抽出することの有効性と，(2) LDA を適用することの有効性が示唆された．

今後は，本論文の提案手法に基づく類似項目の検索機能を実装し，以下に述べる検討を行う．まず，トピックに対する重み付け手法の検討を行う．2 章で述べたトピックの生成確率 θ を用いて，ある項目におけるトピックの重要度を確率的な数値により算出可能な手法を検討する．次に，トピック数の自動決定方法の検討を行う．今回の実験では，類似項目群の数をトピックス数として設定し，既存手法に比べ検索精度が向上していた．そこで，クラスタリング手法などを用いて，予め，類似項目群の数を推定する手法を検討する．そして，異なる分野の項目を対象とした実験を行い，提案手法の適応可能範囲や課題などを明確にする．さらに，類似項目や類似度データを用いることで，1 章で述べたタスクを支援し，項目の作成・管理・出題のサイクルを半自動的に行う事が可能な e テスティング・システムの実現を目指す．

謝辞　本研究は JSPS 科研費 24700904，25・8284 の助成を受けたものです．

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.，植野真臣：統合型 e テスティング・システムの開発と実践，日本テスト学会誌，Vol. 4, No. 1, pp. 53–64 (2008).

[3] Guzman, E. and Conejo, R.: Self-Assessment in a Feasible, Adaptive Web-Based Testing system, *IEEE Trans. Education*, Vol. 48, No. 4, pp. 688–695 (2005).

[4] Songmuang, P. and Ueno, M.: Bees Algorithm for Construction of Multiple Test Forms in E-Testing, *IEEE Trans. Learn. Technol.*, Vol. 4, No. 3, pp. 209–221 (2011).

[5] Hambleton, R. K., Swaminathan, H. and Rogers, H. J.: *Fundamentals of item response theory*, Sage Publications (1991).

[6] moodle: `https://moodle.org`.

[7] Blackboard: `http://www.blackboard.jp`.

[8] 高木輝彦，高木正則，勅使河原可海：学生が作成した問題の類似性に基づいた自動分類方式の提案，情報処理学会第 70 回全国大会講演論文集，Vol. 70, No. 4, pp. 687–688 (2008).

[9] 高木輝彦，高木正則，勅使河原可海：学生が作成した問題の出題パターンによる類似度算出手法の提案と評価，情報処理学会情報教育シンポジウム論文集，SSS2008, pp. 95–102 (2008).

[10] 高木輝彦，高木正則，勅使河原可海：学生が作成した問題の類似度算出手法の提案と評価，情報処理学会論文誌，Vol. 50, No. 10, pp. 2426–2439 (2009).

[11] 高木輝彦，高木正則，勅使河原可海，植野真臣：LDA (Latent Dirichlet Allocation) に基づく問題の類似度算出手法の提案と評価，日本教育工学会第 26 回全国大会講演論文集，pp. 275–276 (2010).

[12] 高木輝彦，高木正則，勅使河原可海，植野真臣：e テスティングにおける LDA を用いた項目間類似度の自動推定手法，日本テスト学会第 10 回大会発表論文抄録集，pp. 150–153 (2012).

[13] 高木輝彦，高木正則，勅使河原可海：知識出現箇所の自動決定アルゴリズムに基づく類似問題検索機能の開発，人工知能学会先進的学習科学と工学研究会資料，Vol. 58, pp. 33–38 (2010).

[14] Young, F. and Hamer, R.: *Multidimensional Scaling–history, theory, and applications*, L. Erlbaum Associates (1987).

[15] Manning, C. D. and Schutze, H.: *Fundamentals of Statistical Natural Language Processing*, MIT Press (1999).

[16] 池田信一，高木輝彦，高木正則，勅使河原可海：多肢選択式項目の出題パターンと選択肢の類似性に着目した難易度推定方法の提案と評価，情報処理学会論文誌，Vol. 54, No. 1, pp. 33–44 (2013).

[17] 池田信一，高木輝彦，高木正則，勅使河原可海：類似問題群からの反復学習が可能な適応型テスト出題方式の提案，マルチメディア, 分散, 協調とモバイルシンポジウム論文集，DICOMO2012, pp. 1402–1409 (2012).

[18] Blei, D. M., Ng, A. Y. and Jordan, M. I.: Latent dirichlet allocation, *J. Mach. Learn. Res.*, Vol. 3, pp. 993–1022 (2003).

[19] 上田修功，斉藤和巳：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. 14, pp. 1–8 (2003).

[20] 麻生竜矢，江口浩二：学術文献の潜在トピックに着目したタンパク質相互関係に関する知識の抽出，情報処理学会論文誌. データベース， Vol. 2, No. 2, pp. 86–95 (2009).

[21] 岩田具治，山田武士，上田修功：トピックモデルに基づく文書群の可視化，情報処理学会論文誌，Vol. 50, No. 6, pp. 1649–1659 (2009).

[22] TREC: `http://trec.nist.gov`.

[23] GENIA: `http://www.nactem.ac.uk/aNT/genia.html`.

[24] NIPS: `http://nips.cc`.

[25] 松本裕治：形態素解析システム「茶筌」, 情報処理，Vol. 41, No. 11, pp. 1208–1214 (2000).

[26] 専門用語（キーワード）自動抽出用 Perl モジュール “TermExtract”：`http://gensen.dl.itc.u-tokyo.ac.jp/termextract.html`.

[27] 芳鐘冬樹，井田正明，野澤孝之，宮崎和光，喜多　一：キーワードの関連用語を考慮したシラバス検索システムの構築，日本知能情報ファジィ学会誌，Vol. 18, No. 2, pp. 299–309 (2006).

[28] Windows 用テキストマイニングツール “termmi”：`http://http://gensen.dl.itc.u-tokyo.ac.jp/termmi.html`.

[29] 中川裕志，湯本紘彰，森　辰則：出現頻度と連接頻度に基づく専門用語抽出，自然言語処理，Vol. 10, No. 1, pp. 27–45 (2003).

[30] 初級システムアドミニストレータ試験：`http://www.jitec.ipa.go.jp/1_11seido/h13/ad.html`.

[31] 基本情報処理技術者試験：`http://www.jitec.ipa.go.jp/1_11seido/h13/fe.html`.