Wikipedia 構造分析による科目シラバスと専門分野知識の関連付け

戴　憶菱† 　浅野　泰仁‡ 　吉川　正俊‡

†京都大学大学院情報学研究科 〒606-8501 京都市左京区吉田本町

E-mail: †daiyiling@db.soc.i.kyoto-u.ac.jp, ‡{asano, yoshikawa}@i.kyoto-u.ac.jp

あらまし　近年，MOOC（大規模オンライン公開講座）のような教育資源のグローバル規模での流通が盛んになっている．現状の環境では，様々な教育者が個別に科目を提供していて，大学のような従来の教育機関による統一的カリキュラムの設定がないため，学習者は科目の選択が困難であり，教育者は自身の提供すべき科目のデザインが困難である．この問題を解決するためには，様々な科目に含まれる知識を体系化することで，学習者の科目間の比較や教育者の科目デザインを支援することが望ましい．そこで本研究では，各科目を従来の教育機関がカリキュラム整備のために作成された専門分野知識と関連付けることを考える．具体的には，計算機科学分野を対象とし，カリキュラム標準 ACM/IEEE-CS Computer Science Curricula 2013（以下 CS2013）に対する，各科目の知識カバレッジを求める手法を提案する．CS2013 の各知識カテゴリを説明する文書と科目シラバスの関連の強度を，それらと関連する Wikipedia 記事のカテゴリ構造から計算するというアイデアによって，これを実現している．実験の結果，より細かい知識カテゴリで知識カバレッジを求める場合に，語彙の差が大きい知識体系文書と科目シラバスに関しては，従来の bag-of-words を用いた手法より Wikipedia 構造を用いた手法の方が結果の類似性が示された．さらに，実験結果の分析を踏まえて今後の課題を議論する．

キーワード　科目シラバス，知識体系，Wikipedia 構造，知識カバレッジ

1. はじめに

ここ数年，情報通信技術と教育の公開化の推進によりウェブ上に誰でもアクセスできる教育資源が増加しつつある．その一例として，MOOC（Massive Open Online Courses，大規模公開オンライン講座）の提供プラットフォーム Coursera が 1,600 科目を提供し，145 教育機関，22,000,000 学習者を参加させる規模に達している[1]. このような教育資源のグローバル規模での流通が学習者と教育者と共に大きな転機をもたらす．

学習者にとって，従来の教育システムの下で大学から卒業した人でも，経済，政治，宗教などの外因で従来の大学教育を受けられない人でも，自由に科目を選択し自身のための教育プログラムを作る機会が増える．MOOC の利用現状に関する調査[1]によると，現在の MOOC の受講者は 25-34 歳の年齢層で高度教育学位を持つ人が中心になっていることが分かる．これは、MOOC が生涯学習や職業発展の一手段としての役割を果たし始めていると示しているだろう。また，[1]の統計分析によると、伝統的な教育にアクセスできない MOOC 受講者の優秀修了証明書の獲得率はそれ以外の受講者より高いことが覗える。このような高いポテンシャルを持つ学習者を支援するために，現存の科目資源を最大限に活用せねばならない．

教育者も，自身の科目をオンラインで公開することでより多くの学習者に触れたり，他教育機関が提供する科目を自身のカリキュラムに導入したりして MOOC の流れに進取的に動き出している[2].

しかし，大量の科目資源をグローバルで流通させ，より多くの人に利用してもらうことは容易ではない．例えば，同じ「データベース」を題目とする科目でも工学部と経営学部といった違う学部から提供されるものはそれぞれの重点が異なる. 初めて「データベース」を耳にする学習者にとっては，それらの科目が設立された経緯の違いをシラバスなどの限られた情報から判断することが難しい．科目の内容に基づく科目選択の判断ができないと，科目の人気や提供側の知名度などの要素に引き付けられ衝動的な意思決定を下す恐れがある．一方，教育機関が自ら必要だと思う科目を黙々と提供し続けると，資源の全体が人気のある科目に偏ってしまう可能性がある．あるいは逆の状況として，教育市場のトレンドや進展を見誤り時代遅れな科目をデザインしてしまうケースもありうる．

以上のような問題が発生する一つの要因は，多種多様な科目がその専門分野における知識的な位置付けが明確ではないことが考えられる．ある専門分野に対して，もし一つ共通な知識体系が備えられたら，提供者を問わずに科目を同様な基準で観察することが可能になる．専門分野知識体系として活用できるものにはカリキュラム基準が考えられる．計算機科学分野を例にして，Computer Science Curricula 2013（CS2013）[3]というカリキュラム基準が存在する．CS2013 は ACM と IEEE との連携特別委員会により編集され，大学学部計算機科学プログラムで扱うべき知識をまとめた指

[1] Coursera ホームページ, https://about.coursera.org/, 2017

導要領である.このカリキュラム基準が 40 年以上の歴史を経て，計算機科学の進展と細分化に伴って最新バージョンの CS2013 で計算機科学の主な知識を二階層の知識カテゴリにまとめる.

このようなカリキュラム基準もよく教育機関のカリキュラム分析に用いられてきた．Sekiya ら[4]は CS2013 の知識体系を参照して違う大学の計算機科学学科のカリキュラムを比較した.彼らの分析によると，違う大学が重視する知識が異なることが分かる.また，日本国内の情報系学科の授業内容の傾向をみるため、計算機科学カリキュラム基準の知識体系が用いられる[5]。CS2103 以外に，似たような役割を担う中小学校の指導要領で教育資源をラベル付ける研究も行われた[6]．以上の例から，異なる教育機関の科目を比較する際に，共通な専門分野知識を参照基準として活用することが有効であるとうかがえる.

ゆえに，本研究は CS2013 を専門分野知識体系として用い，科目シラバスを科目内容の要約とし，二つの手法を提案し，科目内容の CS2013 に定義された知識カテゴリにおける知識カバレッジを推定する．まず，CS2013 とシラバスの文章を bag-of-words の視点から捉え，機械学習モデル――Labeled Latent Dirichlet Allocation によりシラバスと各 CS2013 の知識カテゴリとの関連度を求める．また，CS2013 の情報不足と bag-of-words 手法の文脈情報の損失を克服するため，CS2013 とシラバスとの対応関係をそれらの文書が外部知識資源 Wikipedia の記事とカテゴリ構造における距離で推定する.

以上の研究目的と研究内容によって，本研究の貢献は以下のとおりになる：

- 潜在的トピックモデルと違って，専門分野知識の利用によって，抽出された科目の知識カバレッジが人間で解釈でき，教育的に示唆的である.
- Wikipedia 構造の手法で,従来の bag-of-words の手法が対応できない専門的かつ簡潔な文書の処理を可能にする.

本文は以下のように構成される．2 節では関連研究を紹介する．次に，3 節にて本研究の問題設定を明確にする.4 節では bag-of-words 手法と Wikipedia 構造手法を提出する経緯とそれらの内容を述べる．また，5 節では実験の概要と結果及び考察を記述する.最後に，6 節ではまとめと今後の課題について討論する.

2. 関連研究

2.1 Learning Object Metadata

学習資料を学習者，教育者，学習システムから便利に検索され,利用され,評価され,獲得されるために，LOM（Learning Object Metadata）基準が IEEE により発表された[7]．この基準には，学習資料の ID，タイトル，言語，キーワード，利用できるシーンなどの基本情報についての項目以外に，教育面に関する項目も含める．しかし，こちらの項目には学習資料のタイプ,インタラクティヴ性,知識の密度,想定した利用者像,使われる文脈，難易度，学習時間など学習のスタイルと関わる要素しか挙げられず，学習資料作成者に学習資料の内容自体についてメタデータを付与させる規定がない.

このような LOM によりあるキーワードと関連する学習資料に絞ることができるが，それらの学習資料が内容面でどのような同異があるかを識別することには不十分である．また，もし内容についての詳細な項目が定められたとしても，学習資料の作成者が項目の説明を照らしながらメタデータを作ることは多くの時間と人力を要する．カリキュラム基準の策定に携わった専門家すら全エリアに精通しているとは限らないからである[5]。この点から，本研究は学習資料，すなわち科目資源の内容面に着目し，自動的に専門分野知識と対応するラベルを付与することを目標とする.

2.2 学習資料と専門分野知識との対応

Contractor ら[6]は中小学校教育に着目し，指定された教科書以外の教育コンテンツを活用する際に，学習指導要領などとの対応付けがあったら便利であろうというモチベーションから，教育コンテンツの学習指導要領におけるラベリングの自動化を取り上げた．彼らは Wikipedia, WordNet などの外部資源で情報量に乏しい学習指導要領を拡張し，bag-of-words 手法で教育資源に適切なラベルを付与する．しかし，以下のような問題設定の違いによって彼らの手法が本研究に適応しきれないと考える：

- 対象科目の違い

[6]では中小学校の数学と科学を研究対象にしているため，大学の計算機科学で扱う知識と比べると，“processor to processer communication”の“communication”のような一般的な意味で使われているが特定の分野で専門的な意味を持つ多義語が少ないことから，bag-of-words 手法で十分な特徴語を作ることができたと考えられる.

- 研究目標の違い

[6]は実際の教育資源を直接にラベリングするため，計算量が膨大になる．本研究では科目のシラバスといった科目内容の要約文を専門分野知識と対応付けるため，計算量が減る分，問題の困難さが増える．また，彼らの研究ではラベル付与の有無で結果を評価するため，精度に対する要求がそれほど高くない．それに対して，本研究ではラベルと文書との関連する度合いを

評価するため，より正確な推定が必要となる．

3. 問題設定

前述のように，本研究は CS2013 を専門分野の知識体系として活用する．CS2013 では，18 個の Knowledge Area（下記 KA）から成り立つ．各 KA がさらに幾つの Knowledge Unit（下記 KU）を含み，各 KU は複数の Topic を含む．表 1 が示すように，「Information Management」という KA が「Database Systems」などの KU を含み，「Components of database systems」などの Topic を含む．このように，KA と KU は専門分野の重要かつ広範なカテゴリで単語やフレーズで表現されている．一方，Topic はそれぞれのカテゴリを構成する詳細な概念で簡潔な短文で表現されている．

表 1　CS2013 知識体系の例

KA	KU	Topic
Information Management	Database Systems	・ Approaches to and evolution of database systems
		・ Components of database systems
	Data Modeling	・ Data Modeling
		・ …
..	..	・ …

一方，科目の内容や内容を学生に伝える媒介（テキストブック，スライド，ビデオ，音声など）は多様であるが，科目の内容を纏める文章――シラバスがどのような教育機関でも作成されている．ここで，本研究は科目のシラバスを科目内容の集約と見なして科目内容の解析に用いる．

違う教育機関が提供する科目を一つの空間で比較するため，科目の内容と専門分野の知識体系と対応付けることが必要となる．本研究では CS2013 の KA と KU を専門分野知識のカテゴリとして扱い，同じ KU に属する Topic の集合を一つの文書とみなし，与えられた科目のシラバスがそれぞれの知識カテゴリにおけるカバレッジを求めることを目的とする．

4. 提案手法

上記の問題を解く方法として，まず，自然言語処理の最もシンプルかつ有効的なアプローチ――bag-of-words 手法で CS2013 とシラバスの文書を特徴語でベクトル化し，要素は各特徴語が各文書の重要度を表し，文書のベクトル間の類似度で CS2013 知識カテゴリとシラバスの関連度を求めることが考えられる．そこで，本研究では 4.1 節に述べる bag-of-words 手法を提案する．

しかし，本研究が想定した問題設定では，専門分野知識を表すカリキュラム基準 CS2013 の文書と科目内容を代表するシラバスの文書には二つの特徴がある．その一，両方とも簡潔で不完全な文章表現が見られる．その二，両者の語彙の選択には差がある．CS2013 ではより専門的かつ抽象的な単語を使いがちで，シラバスの方が学生に理解しやすい単語を選択する傾向が見られる．

以上の問題点を解決するため，カリキュラム基準 CS2013 の文章を外部の文章で補足することが考えられる．例えば，CS2013 に挙げられた概念を説明するようなウェブページなどを導入することで，語彙空間を拡張することができる．しかし，従来の bag-of-words 手法には限界がある．それは，bag-of-words 手法は文書が使う単語を独立したものとして扱うため，文書の前後の文脈情報が見落とされている．言い換えると，意味的に関連度の高い文書でも語彙の差が著しい場合，単語の使う頻度で文書間の類似度を推定する手法は対応できない．

これを克服するため，本研究では上記の手法に加えて，Wikipedia を利用して，Wikipedia の記事とカテゴリ構造における距離を用いて文書間の関連度を求める手法を提案する．図 1 は本研究の提案手法をまとめた概念図である．次にそれぞれの手法について詳しい説明を展開する．

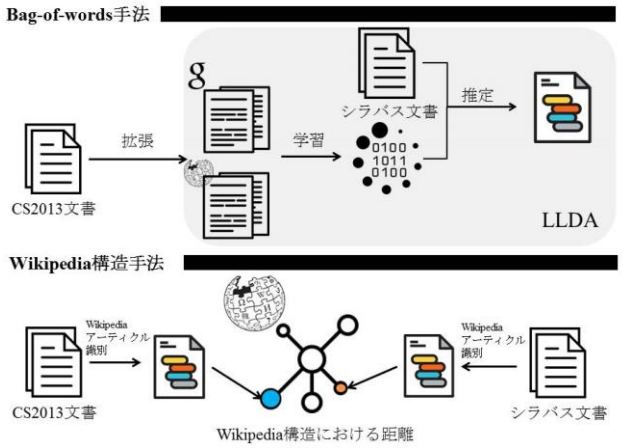

図 1　提案手法の概念図

4.1　Bag-of-words 手法

Labeled Latent Dirichlet Allocation（LLDA）[8]では，語彙集合Vの空間で表現された文書群$d(d \in D_t)$（D_t学習用文書の集合）のベクトル$w_d = (w_1, w_2, \dots, w_{N_d})$（$w_i \in \{1, \dots, V\}$）とトピックの有無を意味する二値ベクトル$\Lambda_d = (l_1, l_2, \dots, l_k)$（$l_k \in \{0,1\}$）を入力することで，新しい文書群$d(d \in D_i)$（$D_i$は推定用文書の集合）のトピック空間における確率分布$P(z|d)$を推定する．

本研究では, CS2013 の知識体系の各 KU を説明する文章を LLDA での一つの文書と見なし，その文書が属する KU と KA を LLDA でのラベルと見なし，学習用文書の集合に投入する．一方，各科目のシラバスを LLDA での一つの文書と見なし，推定用文書の集合に入れる．よって，各シラバスが KU あるいは KA に属する確率が推定される．KA 空間における確率の分布を推定する際, 学習用データに KA のラベルを付ける．同様に KU 空間における確率の分布を推定する際，学習用データに KU のラベルを付ける．

ここで一つ注意すべきことは CS2013 文章の簡潔さがもたらす問題である．表 1 の例が示すように，「Components of database systems」という概念が書かれて具体的なデータベースシステムの構成要素が明記されていない．しかし，実際の科目のシラバスに「software」,「procedures」,「data access language」といったデータベースシステムの構成要素が書かれる可能性がある. CS2013 とシラバスの知識を記述する粒度の非対称を解消するため，外部知識への拡張が考えられる．我々の先行研究[9]では CS2013 の知識と関連するウェブページのスニペット情報を導入して，CS2013 を拡張した．より正確に関連するウェブページを抽出するため，以下の三つのタイプのクエリを試みた：

- KU のタイトル
- 当該 KU が属する KA のタイトル＋当該 KU のタイトル
- 当該 KU のタイトル＋当該 KU を代表する特徴単語

これらのクエリそれぞれによって得られたスニペットを用いた結果，外部文章の導入でモデルの推定精度が上がった．しかし，導入されたウェブページにはソフトウェア会社のホームページなどの雑音が含まれているため，本研究ではより専門性の高いウェブページに限定する手法を提案する. CS2013 は教育コンテンツであり，それと補うような資源と言えば，ウェブ百科事典 Wikipedia が挙げられる．

具体的に，各 KU の説明文に対して，DBpedia Spotlight[10]というツールを利用して，文章の中に Wikipedia 記事との対応がある単語やフレーズを検出し，それと対応する Wikipedia 記事の概要文を導入する．比較のため，先行研究で一番精度の高いクエリタイプ――当該 KU が属する KA のタイトル＋当該 KU のタイトル――から抽出されたウェブページのスニペットも今回の実験に入れる．

4.2 Wikipedia 構造手法

Wikipedia は利用者が自由に編集できるウェブ上の百科事典で，大勢の利用者の知識を集めており，専門用語を中心とするオントロジー（WordNet など）と比べるとより広範な知識を網羅している．また，一般的なウェブサイトと比べると相当専門的かつ学術的な情報を含む．このように, Wikipedia は情報源として量と質のバランスに優れている．

人間の自然言語処理には，単なる文面の意味での読み取りではなく，その人の常識や背景知識も関与している．これらの点から見ると，機械で自然言語を処理する際に, Wikipedia のような人間に編集された知識倉庫を利用することは有用であることが多い[11]．これによって，Bag-of-words 手法より，語彙の曖昧さ，多義語などがもたらす弊害を排除できると期待される．

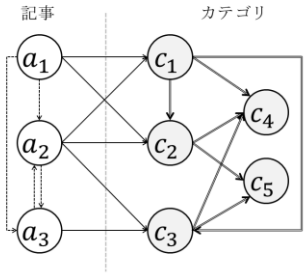

図 2　Wikipedia の記事とカテゴリの構造図

Wikipedia には実体や概念を説明する記事があり, 各記事には関連のある他の記事へのハイパーリンクが付けられている．また，膨大な記事を索引するためのカテゴリが構築されている．具体的には，一つの記事が複数のカテゴリに属する可能性もあるし，一つのカテゴリが複数の親カテゴリに属したり，複数の子カテゴリを持ったりすることも可能である．図 2 が記事とカテゴリとの関連関係を表す概念図である．ここに注意すべきことは, Wikipedia 記事やカテゴリの編集にあたって従属関係の制限がないため，実際の構造は単純な木ではないということである．こういったような関連関係や従属関係には概念間の意味的な関連性が埋め込まれている．文章の表現には共通点が稀な記事であっても，概念的に共通点があればカテゴリの繋がりで見つけられることが多い．ゆえに, Wikipedia の記事間のリンク関係あるいはカテゴリの従属関係を利用して，文書，特に短いかつ非構造的な文書の注釈，分類や概念の抽出に関する研究が数多く行われてきた[11]–[15].

本研究は Wikipedia 記事とカテゴリ構造の関連付けを手段として用い，文書間の関連度を推定する．以下では具体的な手順を説明する．

(1) Wikipedia 記事と対応する単語の検出とその Wikipedia 記事の抽出

ここでは，4.1 節で言及した DBpedia Spotlight とい

うツールを利用して，各 KU の説明文kuと各シラバスsそれぞれについて，Wikipedia 記事と対応する単語を検出する．ku，sと関連する Wikipedia 記事の集合をそれぞれA_{ku}とA_sで記す．記事aと文書kuまたはsとの関連を表す辺を(ku,a)とし，その強度をそれぞれ$w(ku,a)$，$w(s,a)$と表記する．グラフ上には，CS2013 やシラバスから構成される文書頂点をV_Dで，Wikipedia 記事頂点の集合をV_Aで，文書頂点と記事頂点との間の辺の集合をE_0で記す．ここで，図 3 の示すように，一つの文書の複数の単語あるいはフレーズが同じ Wikipedia 記事と関連する可能性があり，本研究ではその数を上記の重みとする．

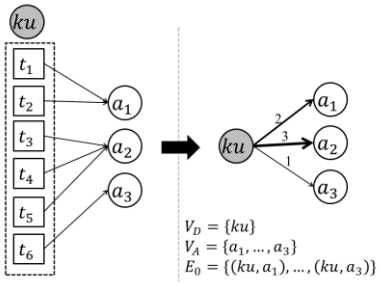

図 3　Wikipedia 記事の検出

(2)　Wikipedia 記事が属するカテゴリの抽出

(1)で得た Wikipedia 記事が属するカテゴリを SPARQL 言語で DBpedia データベースから抽出する．それぞれの記事aが属するカテゴリの集合をC_aで記す．記事aとカテゴリc（$c \in C_a$）との従属関係を示す辺を(a,c)で表し，その強度を$w(a,c)$で表す．

(3)　Wikipedia 構造における距離の算出

Wikipedia 記事とカテゴリ構造は複雑になりがちで処理しやすい階層構造ではない．そのため，不必要なカテゴリ間の辺を削除したり必要な辺だけに着目したりするなどの前処理が必要となる[13], [15]. あるいは，[12]のようにカテゴリを 5 階まで考慮して記事間の最短距離を求める方法がある．しかし，それらの先行研究は異なる研究目標を持ち，本研究の文脈と違うため，彼らの距離の計算方法を直接に応用することは相応しくないと考える．

例え記事とカテゴリの整った階層グラフが作成されたとしても，その階層を何階まで遡ったら計算された距離が文書間の関連度を適切に表せるかはまた一つの課題である．極端に言うと，木構造の根まで用いるとすべての頂点が関連してしまい雑音だらけとなってしまう．一方，木構造の葉しか用いないと情報が損失し頂点間の関連性が得られない．本研究では，その中間となる三つのパターンを試みる．

a　　記事層だけを用いる方法

図 4 が示すように文書ku，sを表す頂点と，それらと関連する記事を用いる．そして，式(1)のように文書間の関連度$rel(ku,s)$を計算する．二つの文書が持つ共通な Wikipedia 記事が多ければ多いほど両者の関連性が高いと考えるからである

$$rel(ku,s) = \sum_{a_k \in A_{ku} \cap A_s, (ku,a_k),(s,a_k) \in E_0} w(ku,a_k) + w(s,a_k) \quad (1)$$

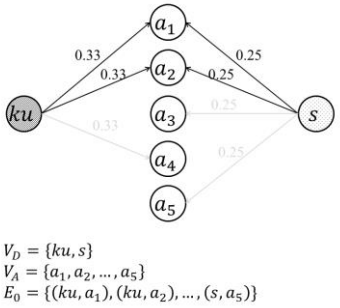

図 4　記事層だけのグラフ例

b　　記事層と第一層目のカテゴリ層を用いる方法

このパターンでは記事と直結するカテゴリを計算の範囲に入れる．図 5 が示すように，記事aとカテゴリc（$c \in C_a$）を連結する辺の集合をE_1で記し，式(2)でその強度$w(a,c)$を求める．

$$rel(ku,s) = \sum_{c_l \in \left(\bigcup_{a_k \in A_{ku}} C_{a_k}\right) \cap \left(\bigcup_{a_k \in A_S} C_{a_k}\right), (a_k,c_l) \in E_1} w(a_k,c_l)$$

$$w(a_k,c_l) = \begin{cases} \frac{1}{|C_{a_k}| \times |A_{ku}|}, & a_k \in A_{ku} \\ \frac{1}{|C_{a_k}| \times |A_s|}, & a_k \in A_s \end{cases} \quad (2)$$

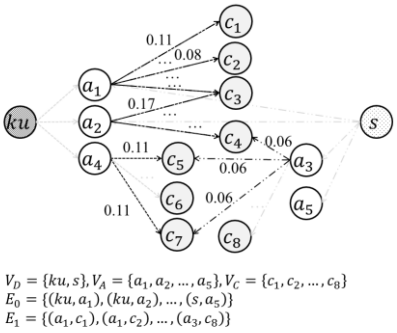

図 5　記事層とカテゴリ一層のグラフ例

c　　第一層目のカテゴリ層を用いる方法

記事とカテゴリとの従属関係の強度がそれぞれ違うため，TFIDF の考え方から多数の文書と関連するカテゴリからの辺の重みを下げることで普遍的なカテゴリとの関連を抑える．具体的には，式(3)が示すように，

Wikipedia 記事を経由して文書と間接的に関連するカテゴリの集合を$C_d = \{c_1, c_2, \ldots, c_m\}$で表記し，各カテゴリが当該文書への貢献度を$Cont(c, d)$で表記し，TFIDF[16]の算式で算出する（ここでは省略）．図 6 はこの手法のグラフ例である．

$$rel(ku, s) = \sum_{c_l \in C_{ku} \cap C_s} Cont(c_l, ku) + Cont(c_l, s) \quad (3)$$

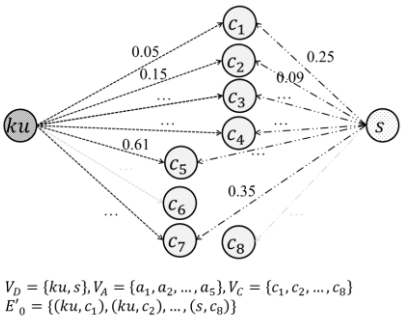

図 6 カテゴリ一層のグラフ例

以上，KU とシラバスとの関連度を求める手法を述べてきたが，KA は KU の集合であるため，ここは式(4)の示すように，あるシラバスに対して当該 KA に属する KU との関連度の合計値で KA とシラバスの関連度を求める．

$$rel(ka, s) = \sum_{ku_k \in KU_{ka}} rel(ku_k, s) \quad (4)$$

なお，KU_{ka}は当該 KA に属する KU の集合である．

ここで注意すべき点は，以上の手法で求められたシラバスと各 KU との関連度の数値への解釈である．この数値はシラバスと KU と両方に共通して関連する Wikipedia 記事の数から換算されるもので，シラバスや KU の説明文の長さにより変化するため，値自体が知識の分量を意味しない．しかし，あるシラバスとそれぞれの KU との関連度を比較することで，シラバスが含める知識の各 KU における分布が分かる．

5. 実験概要と結果

5.1 データセット

本研究では CS2013 の知識体系に記述された 163 個の KU の説明文を一つの文書として収集して，その説明文が属する KU と KA をラベルとして付与する．そして，CS2013 に掲載されている実際の大学の科目が当該カリキュラム基準との対応状況をテストデータとして使う．CS2013 サンプル科目には指導者による科目内容の記述と各 KA と KU へどれほど対応しているか表す講義時間数が記載されている．本研究は科目内容についての記述をシラバスとして活用し，上記の KA と KU への対応情報を正解として利用した．また，提案手法の MOOC 教育コンテンツへの適応性を試すため，幾つかの MOOC 科目を収集し，二人の教員に依頼して自ら KA と KU への対応情報を作成し正解とした．ここで，合計 143 科目を分析に用いた．

5.2 評価基準

二つの手法で推定されたシラバスが各 KA と KU におけるカバレッジの正確さを評価するため，以下の指標を用いる．

(1) nDCG

nDCG（Normalized Discounted Cumulative Gain）は Information Retrieval で検索結果のリストの良さを評価するによく使われる指標である．式(5)が示すように，評価される item の順位が理想的な順位で再現できる度合を 0~1 の値で表す．本研究では，シラバスの各 KA と KU におけるカバレッジによって上位k個の KA/KU だけに着目する際に nDCG の採用が望ましいと考える．ここで，kの値はサンプル科目がカバーする KA/KU の数の平均値とする．

$$\begin{cases} G_c[i] = rel_c[i] \\ DCG_c[k] = \sum_{i=1}^{k} \frac{G_c[i]}{\log_2 i+1} \\ nDCG_c[k] = \frac{DCG_c[k]}{IDCG_c[k]} \end{cases} \quad (5)$$

ここで：

c:科目

$rel_c[i]$：ある科目に対し，指導者が付与したi番目の KA の知識を扱う講義時間数

$G_c[i]$：ゲイン値

$DCG_c[k]$：上位k個に着目する際のディスカウント累積ゲイン値

$IDCG_c[k]$：上位k個の KA に着目する際の理想的な順位で得られるディスカウント累積ゲイン

(2) Cosine similarity

Cosine similarity（下記 cossim）はベクトルの類似度を測るためによく使われる（式(6)）．本研究では，シラバスが KA と KU の全般における分布を評価する際に cossim を利用する．

$$\text{sim}(v_1, v_2) = \frac{v_1 \times v_2}{\|v_1\| \times \|v_2\|} \quad (6)$$

5.3 結果

提案した手法を比較するため，実験の ID と対応する手法を表 2 にまとめる．4 節の提案手法で言及したツールはそれぞれの言語で，それ以外の実装は全て Python で行い，以下の結果を得た．

表 2　実験 ID と実験内容

実験 ID	実験内容	対応
LLDA	CS2013 の文書を学習用データに，シラバスの文書を推定用データに用いる.	4.1
LLDA_snippet	外部ウェブページのスニペット文章を加えた CS2013 の文書を学習用データに，シラバスの文書を推定用データに用いる.	4.1
LLDA_wikiabs	関連する Wikipedia 記事概要文を加えた CS2013 の文書を学習用データに，シラバスの文書を推定用データに用いる.	4.1
Wikipedia_0	文書と連結する Wikipedia 記事層だけを考慮する.	4.2(3)の a
Wikipedia_1	文書と連結する Wikipedia 記事層とそれらとさらに連結するカテゴリ層を考慮する．記事とカテゴリの数で辺の重みを正規化する.	4.2(3)の b
Wikipedia_2	文書と連結する Wikipedia 記事層とそれらとさらに連結するカテゴリ層を考慮する．カテゴリの文書への貢献度を TFIDF で求める.	4.2(3)の c

表 3 が示すように，シラバスの KA におけるカバレッジについてはどんな評価基準でも bag-of-words 手法の結果が Wikipedia 構造手法より少々上回っている．特に，上位三個の KA に着目する際に，関連する Wikipedia 記事概要文を CS2013 に加えた場合は精度 0.790 を達成した．表 4 が示すように，シラバスが KU におけるカバレッジについてはどんな評価基準でも Wikipedia 構造手法の結果が bag-of-words 手法のより少々上回っている.

表 3　KA におけるカバレッジの結果

	Bag-of-words 手法		
	LLDA	LLDA_snippet	LLDA_wikiabs
nDCG@3	0.777	0.777	**0.790**
cossim	0.725	0.748	**0.751**
	Wikipedia 構造手法		
	Wikipedia_0	Wikipedia_1	Wikipedia_2
nDCG@3	0.701	0.776	**0.785**
cossim	0.637	0.688	**0.709**

表 4　KU におけるカバレッジの結果

	Bag-of-words 手法		
	LLDA	LLDA_snippet	LLDA_wikiabs
nDCG@9	0.425	0.433	**0.441**
cossim	0.414	0.420	**0.425**
	Wikipedia 構造手法		
	Wikipedia_0	Wikipedia_1	Wikipedia_2
nDCG@9	0.261	**0.502**	0.460
cossim	0.326	**0.451**	0.446

このように，KA あるいは KU におけるカバレッジを推定する際，bag-of-words 手法では関連する Wikipedia 記事概要文を CS2013 に加える方法が優位性を示した．一方，Wikipedia 構造手法では，文書間の Wikipedia 構造におけるどの距離の求め方が最も精度が高くなるかは推定する知識カテゴリの粒度によって変わった.

5.4　考察

5.3 節の結果から見ると，bag-of-words 手法は粗い粒度の知識カテゴリのレベルで与えられた文書にラベルを付与するには安定したパフォーマンスを示した．すなわち，bag-of-words 手法で大雑把な文書のラベリングができる．しかし，さらに細かい知識のカテゴリで文書をラベリングするにあたって，ラベルの数の急増による各ラベルの特徴語の減少から，誤った知識カテゴリに割り振ってしまう欠点がある．それを解決するため，外部の関連文書をラベルの説明文に加える方法も試した．その結果，Wikipedia 記事の概要文のような良質な文章で元のカリキュラム基準を拡張したにもかかわらず精度の向上が著しくない．その理由は，カリキュラム基準とシラバスの文章の語彙レベルでの重複が少なく，あるとしても同形異義語のケースも多いことが考えられる．つまり，文脈情報の損失が bag-of-words 手法の限界をもたらしていると考えられる.

今回の実験では，Wikipedia 構造手法は KU におけるカバレッジを推定する場合しか優位性を示さなかった．一方，粗い知識カテゴリの確率を推定する際，文書と関連する Wikipedia 記事とカテゴリの一層だけを考慮すると，大半の知識カテゴリと関連する Wikipedia 記事とカテゴリとそれらを結びつける辺には大差が見られない恐れがある．一方，細かい知識カテゴリのカバレッジの推定には Wikipedia を用いることで文書の中の単語やフレーズの解読の精度が上がり，無関係な知識カテゴリに割り振ることが避けられる．要するに，Wikipedia 構造における頂点間の距離の計算の改善によりパフォーマンスの向上が期待できる.

6. むすび

本研究ではオンラインで公開された科目資源をより有効に利用するため，科目資源を専門分野知識との対応を付けることを目的として，bag-of-words 手法と Wikipedia 構造手法という二つの手法を提案した．実験

の結果，大きい知識カテゴリで科目の知識カバレッジを推定する場合，bag-of-words 手法が優位性を示した．一方，さらに細かい知識カテゴリで科目の知識カバレッジを推定する場合には，Wikipedia 構造手法がポテンシャルを示した．しかし，今回の文書間の Wikipedia 構造における距離の計算が単純であるため，より正確に文書間の関連度を推定できる計算方法を構築することが今後の課題である．また，bag-of-words 手法と Wikipedia 構造手法はそれぞれの長所を持つため，両手法を上手に組み合わせることで，より精度の高いモデルを構築できると考えられる．

参 考 文 献

[1] T. R. Dillahunt, B. Z. Wang, and S. Teasley, “Democratizing higher education: Exploring MOOC use among those who cannot afford a formal education,” *The International Review of Research in Open and Distributed Learning*, vol. 15, no. 5, Oct. 2014.

[2] F. M. Hollands, D. Tirthali, F. M. Hollands, and D. Tirthali, “Why Do Institutions Offer MOOCs?,” *Journal of Asynchronous Learning Networks*, vol. 18, no. 3, 2014.

[3] I. C. Society, “Computer Science Curricula 2013: Curriculum Guidelines for Undergraduate Degree Programs in Computer Science,” ACM, 2013.

[4] T. Sekiya, Y. Matsuda, and K. Yamaguchi, “Curriculum analysis of CS departments based on CS2013 by simplified, supervised LDA,” in *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, 2015, pp. 330–339.

[5] I. Kiyoshi and others, “Investigation on the Educational Contents among Informational Science and Engineering Departments by Using Syllabus (Intermediate Report),” Information Processing Society of Japan, Technical Report 6, 2010.

[6] D. Contractor, K. Popat, S. Ikbal, S. Negi, B. Sengupta, and M. Mohania, “Labeling Educational Content with Academic Learning Standards,” in *Proceedings of the 2015 SIAM International Conference on Data Mining*, 2015, pp. 136–144.

[7] “IEEE Standard for Learning Object Metadata,” *IEEE Std 1484.12.1-2002*, p. i-32, 2002.

[8] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, “Labeled LDA: A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 2009, pp. 248–256.

[9] Y. Dai, Y. Asano, and M. Yoshikawa, “Course Content Analysis: An Initiative Step toward Learning Object Recommendation Systems for MOOC Learners,” in *Proceedings of the 9th International Conference on Educational Data Mining*, 2016.

[10] J. Daiber, M. Jakob, C. Hokamp, and P. N. Mendes, “Improving Efficiency and Accuracy in Multilingual Entity Extraction,” in *Proceedings of the 9th International Conference on Semantic Systems*, New York, NY, USA, 2013, pp. 121–124.

[11] E. Gabrilovich and S. Markovitch, “Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis,” in *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, San Francisco, CA, USA, 2007, pp. 1606–1611.

[12] Y. Genc, Y. Sakamoto, and J. V. Nickerson, “Discovering Context: Classifying Tweets Through a Semantic Transform Based on Wikipedia,” in *Proceedings of the 6th International Conference on Foundations of Augmented Cognition: Directing the Future of Adaptive Systems*, Berlin, Heidelberg, 2011, pp. 484–492.

[13] S. P. Ponzetto and M. Strube, “Deriving a Large Scale Taxonomy from Wikipedia,” in *Proceedings of the 22Nd National Conference on Artificial Intelligence - Volume 2*, Vancouver, British Columbia, Canada, 2007, pp. 1440–1445.

[14] X. Hu, X. Zhang, C. Lu, E. K. Park, and X. Zhou, “Exploiting Wikipedia As External Knowledge for Document Clustering,” in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2009, pp. 389–396.

[15] V. Nastase and M. Strube, “Decoding Wikipedia Categories for Knowledge Acquisition,” in *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 2*, Chicago, Illinois, 2008, pp. 1219–1224.

[16] G. Salton and C. Buckley, “Term-weighting Approaches in Automatic Text Retrieval,” *Inf. Process. Manage.*, vol. 24, no. 5, pp. 513–523, Aug. 1988.