

情報工学教材検索システムにおける教材選択のための情報提示

Information Presentation for the Learning Material Selection
in an Information Engineering Learning Material Search System太田 健吾†
Kengo Ota杉本 徹†
Toru Sugimoto

1. はじめに

近年、学生はわからない項目や興味を持った項目の学習のために、Web上に公開された教材を利用することが多い。その例として、大学が公開している講義資料がある。しかし、そのような教材は通常、組織化されることなく公開されているため、従来の検索技術では目的に合った資料を見つけることは難しい。

中森[1]は、Web上に存在する情報工学分野の大学講義資料を情報工学教材と捉え、ばらばらだった教材を学習項目オントロジーに基づき体系化することにより、ユーザがリンクをたどることで、教材にアクセスできるシステムを開発した。しかし、このシステムで検索結果として得られた多くの教材の中から、自分に合った教材を選ぶためには一つ一つ教材に目を通す必要があった。

本研究では、ユーザが最も学びたいと思う教材を効率よく選択できるようにするために、学習項目オントロジーに基づいて教材の内容に関する情報を抽出し、提示する手法を提案する。また、その成果を中森のシステムに反映させ、評価することで本研究の有用性を示すとともに、学生の学習のために必要となる情報について幅広く検討する。

教材の内容に関する情報を自動的に抽出する研究は、辻ら[2]によっても行われている。この研究では、教材を検索するための汎用的なメタデータである学習オブジェクトメタデータ（LOM）を自動的に生成することを目的に、Web上のシラバスから授業科目名と講義目的・内容の抽出を試みている。

また、情報を提示する手法に関しては林ら[3]の手法を参考にした。この研究は、検索結果ページのインタフェースとして各ページにおいて特徴的に出現する語句を特徴語群として提示する手法を提案している。

なお、本研究において教材ファイルとは、Web上に公開されたPDFファイルやPowerPointファイル形式の大学講義資料のことを指し、教材とは教材ファイルへのリンクを持つWebページ全体のことを指す。

2. 情報工学教材検索システム

2.1 システムの概要

図1に、先行研究[1]で提案されたシステムの概要を示す。このシステムは、ユーザが学びたい項目に合わせてトップページからリンクをたどり、教材にアクセスして閲覧することで学習することができる。本研究では、教材ページを改良して、教材の内容に関する情報をさらに詳しく提示することで、最も学びたいと思う教材を効率的に選ぶことができるようにする。

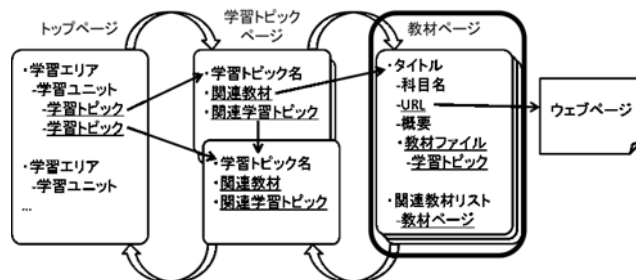


図1. 情報工学教材ポータルサイトの概要

2.2 学習項目オントロジー

学習項目オントロジーとは、情報工学分野において学ぶべき項目を体系化したオントロジーである。情報専門学科のカリキュラム標準J07における5つの領域のうち、コンピュータ科学領域における知識体系CS-BOK-J 2007[4]に基づいて作成された。この分野で学ぶべきことが学習エリア・学習ユニット・学習トピックの3層に分けて構造化されている。学習トピックは教材データベースにある教材と結びつけられている。表1は、学習エリアDSに属する学習ユニットDS1に関する情報の一部である。

CS-BOKとの違いは、学習ユニットの情報であったトピックを学習トピックとして独立させ、学習ユニットの下位構造に位置付けている点である。

表1 学習項目オントロジーの一部

DS	学習エリア名	離散構造
DS1	学習ユニット	関数, 関係, 集合
	学習トピック	集合
		ベン図
		補集合
		デカルト積
	学習成果	具体例を用い, 集合, 関数, 関係などに関する用語や記号を説明できる

3. 教材の内容に関する情報の提示

3.1 教材の内容に関する情報

図2は、教材に関する情報の例である。先行研究では学習トピックの抽出が行われた。本研究では教材の内容に関するその他の情報として、学習トピックより大きな学習分野に関する情報である特徴学習ユニットと、学習トピック以外の専門用語である関連技術用語を抽出し、提示する。

† 芝浦工業大学, Shibaura Institute of Technology

教材に関する情報		具体例
教材の内容に関する情報	教材名	コンパイラ理論
	学習トピック	字句解析, 構文解析, …
	特徴学習ユニット	言語翻訳システム, …
	関連技術用語	Ruby, UNIX, C言語, …
その他の情報	教材の種類	講義資料, 演習資料
	教材のサイズ	24個のファイル
	教材の年度	2010年度
	大学	芝浦工業大学

図 2. 教材に関する情報

3.2 処理の流れ

本研究における処理の流れを図3に示す。

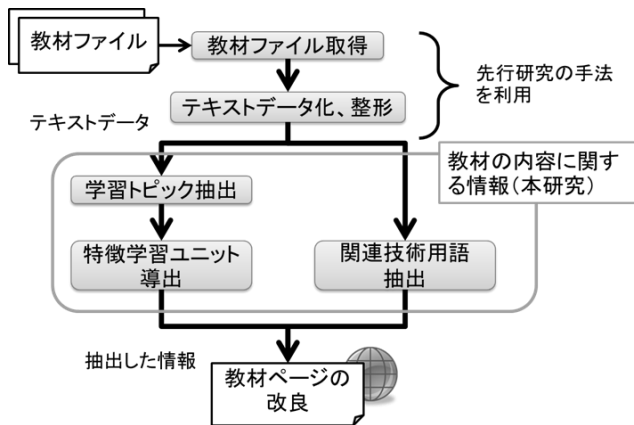


図 3. 処理の流れ

本研究ではまず、先行研究の手法を利用して教材ごとの教材ファイルを Web から取得した後、テキストデータ化と整形を行う。次に、前節で述べた教材の内容に関する情報を抽出する。最後に、抽出した情報をメタデータとしてまとめ、教材ページに反映させる。

3.3 特徴学習ユニットの導出

特徴学習ユニットは、学習項目オントロジーにおいてその教材と最も関連が強い学習ユニットの集合であり、おおむね大学における科目名に相当する情報を提示するものである。教材の科目名を直接抽出することは難しいが、その教材の学習内容を、学習ユニットの組み合わせによって表現することができる。

まず、学習トピックを MeCab[5]の単語辞書に登録し、教材ファイルから抽出したテキストデータを形態素解析することで学習トピックの抽出を行う。その後、図4のように学習項目オントロジーを利用して特徴学習ユニットの導出を行う。

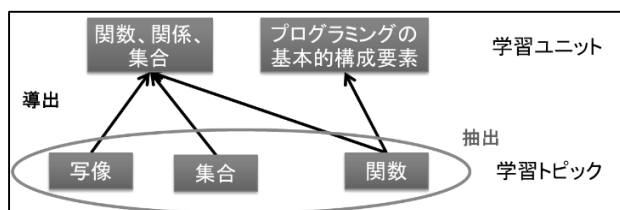


図 4. 特徴学習ユニットの導出

まず、すべての教材を対象とした TF-IDF 法に基づき教材 p における学習トピック t のスコア $\text{tfidf}(p, t)$ を求める。

$$\text{tfidf}(p, t) = \text{tf}(p, t) \times \text{idf}(t)$$

$$\text{idf}(t) = \log \left(\frac{N}{\text{df}(t)} \right)$$

$\text{tf}(p, t)$: 学習トピック t の出現頻度, $\text{df}(t)$: 学習トピック t を含む教材の個数, N : 教材数

このスコアをすべての学習トピックに対して求めて、それを学習ユニットごとに合計することで、教材 p における学習ユニット u のスコア $\text{score}(p, u)$ を導出する。

$$\text{score}(p, u) = \sum_{i=1}^n \text{tfidf}(p, t_i)$$

$t_i (1 \leq i \leq n)$: u を構成する学習トピック

スコアが大きい上位 5 個をこの教材の特徴学習ユニットとする。

3.4 関連技術用語の選出

教材ファイル内には、「C 言語」や「スマートフォン」のように、学習項目オントロジーにはない情報技術に関連した専門用語が数多く出現する。こうした用語は、同じ学習項目の教材でもより詳しい特徴を表し区別する役割を果たすため、抽出を行う。抽出には Wikipedia 日本語版[6]を利用する。また、ユーザーが技術用語について理解できるように Wikipedia に掲載されたカテゴリや概要文をあわせて提示できるようにする。

(1) 関連技術用語の候補の選出

まず、Wikipedia のカテゴリ階層を利用し、関連技術用語の候補の抽出を行う。カテゴリ「コンピュータ」を根として 6 階層までのカテゴリに含まれる語句を抽出した。この理由は、下位カテゴリに「ソフトウェア」や「コンピュータ言語」など、情報工学に関係ある技術のカテゴリが網羅されているためである。ただし、誤った抽出を防ぐため 3 文字以下の英単語、「DOC (ファイルフォーマット)」のように曖昧さ回避のため括弧が付いた単語、アニメやゲーム名など情報技術と関係が薄い用語、および「Performa」など一般英単語の一部として現れやすい単語を手作業で削除した。その結果、14,016 語を関連技術用語の候補として選定した。

また、図5のように Wikipedia における最初の説明文を概要文として抽出し、提示できるようにする。

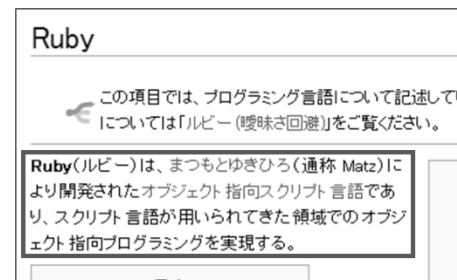


図 5. 概要文の抽出

(2) 教材からの関連技術用語の抽出

次に、選出した候補を学習トピックと同様に MeCab の単語辞書に登録し、形態素解析を行うことで、教材ファイルから候補となる語句を抽出した。

抽出された語句に対して、学習トピックと同様にすべての教材を対象とした TF-IDF 法によるスコアを求めて、そのスコアの上位 30 語を関連技術用語として選出した。

3.5 教材ページの改良

抽出した情報をまとめ、教材ページの改良を行う。表 2 は教材ページで提示する情報の例であり、図 6 はこれを実装した例である。提示された教材の内容に関する情報に基づいて、ユーザは学びたい教材を選ぶことができる。さらに、わからない関連技術用語があった場合は、各用語をクリックすることで Wikipedia から抽出した概要文が提示される。また、教材ファイルごとの学習トピックも提示されるので、実際に学んでいく際に大まかな内容を容易に把握することができる。

表 2 教材ページにおいて提示する情報の例

	提示する情報の例
教材名	2010年度コンパイラ理論の講義資料
教材のURL	http://...
特徴学習ユニット	PL3-言語翻訳入門、...
関連技術用語	Ruby, EBNF, 属性文法, 先読み, Python, ...
- 詳細情報	Rubyは、まつもとゆきひろにより開発されたオブジェクト指向スクリプト言語であり... [Ruby, オブジェクト指向言語, ...]
教材ファイル	07RubyIntro.pdf, ...
- 学習トピック	変数, 代入, 論理式, 有限オートマトン, ...
関連教材	2011年度プログラム言語論の講義資料, ...

2010年度コンパイラ理論の講義資料	
http://www.sakurai.comp.ae.keio.ac.jp/classes/2010Compiler.html	
この教材で学べる学習ユニット	
1. PL3 - 言語翻訳入門 (コア必修 2 時間)	
2. PL10 - 言語翻訳システム (選択)	
3. DS6 - オートマトンと正規表現 (コア必修 6 時間)	
4. PF3 - 基本データ構造 (コア必修 14 時間)	
5. DS7 - 計算論概論 (コア必修 4 時間)	
この教材に関連した技術用語	
Ruby EBNF 属性文法 先読み Python Schema 演算子 Perl 終端記号と非終端記号 識別子 コメント 左再帰 曖昧な文法 エラー Lex 文句 FooBar オートマトン 正規文法 Bison HelloWorld 入 Basename 数値型 文句解析器	
詳細情報 (Wikipediaより引用)	
Ruby (ルビー) は、まつもとゆきひろ (通称 Matz) により開発されたオブジェクト指向スクリプト言語であり、スクラッチ領域でのオブジェクト指向プログラミングを実現する。 [Ruby, オブジェクト指向言語, オープンソース, スクリプト言語, プログラミング言語]	
教材ファイル一覧	
07RubyIntro.pdf 学習トピック: 式, 変数, 代入, 論理式, 有限オートマトン, 関数, 配列, 文字列, 型	

図 6. 教材ページの実装例

4. 評価と考察

4.1 関連技術用語の適切さの評価

まず、選出した関連技術用語が実際に教材の中で取り上げられているかどうか評価した。5 つの教材を対象に、教材から抽出された関連技術用語の候補に人手で正解データを与え、選出個数ごとに適合率と再現率を計算した。この結果、図 7 のデータを得た。

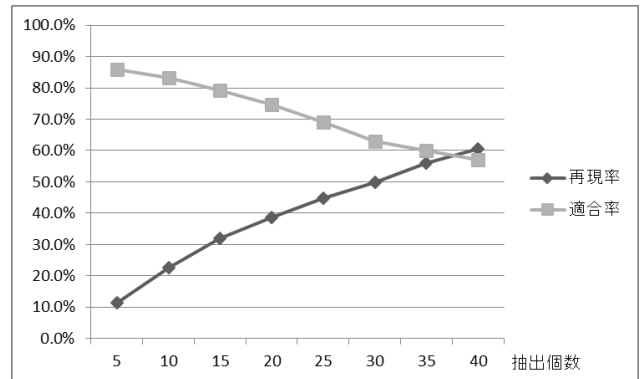


図 7. 関連技術用語の選出個数ごとの評価

関連技術用語として 30 個選出したときの再現率は 49.8%, 適合率は 62.9% となった。

表 3 は、コンパイラ関連の教材の 1 つから抽出された用語の一部である。正解を付けなかった単語の例として、「実装」や「導出」のように教材内で単独の技術用語としての言及がない語句があった。このほかにも、UNIX コマンドである「strings」が教材ファイルに含まれるソースコードから抽出されるなど、誤った抽出がいくつか見受けられた。今後、教材から語句を抽出する際に品詞や意味を考慮するなど改良を行い、より教材の内容に合った関連技術用語を選出できるようにしたい。

表 3 関連技術用語として選出された用語の例

順位	関連技術用語	正解	順位	関連技術用語	正解
1	型推論	○	6	オートマトン	○
2	Yacc	○	7	構文木	○
3	Lex	○	8	StandardML	○
4	実装	×	9	型付きラムダ計算	○
5	導出	×	10	計算機	○

4.2 教材ページの有用性の検証

次に、改良した教材ページが実際にユーザの役に立ったかどうかを評価する。3 つの学習トピック「テキスト処理」「SQL」「オブジェクト指向設計」についてそれぞれ 5 つの教材を用意し、各学習トピックについて次の順に提示する情報をもとに、学びたいと思う教材を順位付けするとともに、その順位を選んだ根拠を記述してもらった。

- 1) 教材名のみ
- 2) 教材名、特徴学習ユニット、関連技術用語を掲載した教材ページ
- 3) 実際の教材そのもの

この実験を学部 4 年生の被験者 5 人に行い得られた順位について、3 つの学習トピックそれぞれスピアマンの順位相関係数を計算し、平均した結果は以下の通りとなった。

表 4 順位相関係数の平均

比較対象	テキスト処理	SQL	オブジェクト指向設計	平均
教材名 → 教材ページ	0.18	0.48	0.20	0.29
教材ページ → 実際の教材	-0.08	0.38	0.28	0.19
教材名 → 実際の教材	-0.02	0.18	-0.06	0.03

結果より、教材ページを見てから実際の教材を見たときと教材を見てから実際の教材を見たときとを比べて、前者

の方が相関係数は大きい値となっていることがわかる。よって、教材ページは教材を選ぶために参考になっていると言える。しかし、その値が0.19小さいことから、教材ページと実際の教材との間に大きな隔たりがあることがわかり、課題を残す結果となった。

まず、教材名のみを見てもらったときにはタイトルから内容を判断できるかどうかを基準に順位付けをする人が多かった。続いて、教材ページを見て判断してもらった際には、特に提示された用語の中になじみがある単語の多い教材に対して高い順位を付けた人が多かった。例えば「『オブジェクト指向言語』『ハイパーテキスト』など聞いたことのある単語が書いてあるため、すんなり学べたと思った」などの記述がある。こうした点から、教材ページに提示された情報が参考になったことがわかる。

しかし、実際の教材を見てもらったときには、学習内容以外の情報が順位付けに影響することが多かった。例えば、「色使いが乏しい」など資料の見やすさや分量などの影響を受けたと考えられる記述があった。このため、学習内容以外の情報もユーザが学びたい教材を選ぶ基準として無視できないことがわかった。

4.3 システムの実用性に関する評価実験

さらに、改良した教材ページを実装したシステムを利用して教材を実際に選び、その教材を使って学習してもらうことで、システムの実用性について評価した。この評価により、本システムがわからないことや興味のあることについて詳しく調べるために有効であるかどうかを確認するとともに、今後の課題について明らかにする。

被験者9人（学部4年生7人、2年生2人）に対して、本システムを利用して学習項目オントロジーの学習ユニット「AL・アルゴリズム」の範囲の中で、被験者が学びたいことに基づいて自身で教材を検索し、選んだ教材を一定時間閲覧してもらった。この結果、「またこのシステムを利用して学習したいですか？」という問いに対して、最も低い1から最も高い4までの4段階評価で平均2.89のスコアを得ることができた。このため、本システムは興味のあることについて詳しく調べるために有効であったことがわかる。ここで、さらに詳しい結果について分析する。

まずは、教材を選ぶ上で決め手となった情報について記述してもらった。この結果、教材名と関連技術用語を参考にした人がそれぞれ3人ずつとなった。この点から、特に教材に関係する技術用語を提示することで教材の内容が具体的にいったと言える。

次に、教材を閲覧することで学習トピックについて十分に学ぶことができたか、また選んだ教材の難易度、分量、および読みやすさが適切であったかどうか4段階で評価し、教材を読んでよかった点と最も知れたかった点をそれぞれ記述してもらった（表5）。学習トピックについて十分に学ぶことができたという結果が得られていることがわかる。しかし、教材の分量や読みやすさの評価についてはあまり高い評価が得られなかった。

また、教材を読んでよかった点として「アルゴリズム等が図示されていてわかりやすい」など難易度や読みやすさについて一定の評価があったものの、「途中の説明がなく、わかりづらい」という意見や、「求める学習トピックについての記述がなかなか見つからない」など、教材として利用する上での読みやすさや分量に関して課題があることがわかった。

5. まとめと今後の展望

本研究では、情報工学教材を検索するシステムにおいて学習項目オントロジーに基づいて教材の内容をあらわす情報として特徴学習ユニットと関連技術用語の2種類の情報を抽出し、提示する手法を提案するとともに、今後の課題について検討した。

ユーザは教材ページを閲覧することで、教材から学習できる項目を大まかに判断できるようになった。また、情報工学の理論的な項目だけでなく、興味のあるハードウェアやソフトウェア、ネットワーク関連技術の項目に注目して教材を判断できるようになった。

今後の展望として、まず学習トピックと関連技術用語との関連づけを行うことが挙げられる。教材ファイルに限らずシラバスやWikipediaの本文文から情報を抽出することで、より精度の高い教材検索を実現していきたい。

また、システム全体を評価した結果をふまえて、今後はさらに学習項目について学生がより詳しく学習できるようにするために、学習項目オントロジーを利用して大学の講義資料に限らず、WikipediaやQ&Aサイトなども教材と一種として整理したい。

謝辞

教材ページの有用性の評価、システムの実用性に関する評価それぞれの評価実験に協力してくださった皆様に感謝します。

参考文献

- [1] 中森慎平, 杉本徹, “学習項目オントロジーに基づく情報工学教材の体系化”, 情報処理学会第74回全国大会, 2012.
- [2] 辻靖彦, 森本容介, “メタデータの自動生成を目的としたシラバス文書の情報抽出”, メディア教育研究, Vol.7, No.1, R1-8, 2010.
- [3] 林祐平, 品川徳秀, “特徴語付きウェブ検索インタフェースの提案”, 電子情報通信学会第19回データ工学ワークショップ(DEWS2008), No.B5-6, 2008.
- [4] 情報専門学科におけるカリキュラム標準 J07, <<http://www.ipsj.or.jp/12kyoiku/J07/J0720090407.html>>, (2013/06/24 取得)
- [5] MeCab, <<http://mecab.googlecode.com/>>
- [6] Wikipedia 日本語版, <<http://ja.wikipedia.org/>>

表5 選んだ教材についての4段階評価結果

項目	点数
十分に学べた	2.89
難易度	2.78
分量	2.33
読みやすさ	2.56