

学習項目オントロジーとの対応付けを用いた情報工学教材の検索

田古島 太郎 西尾 太佑 杉本 徹

芝浦工業大学 工学部 情報工学科

1. はじめに

近年、インターネットの普及により OCW[1]のように Web 上での大学講義資料の無償公開の活動が進んでおり、大学関係者以外の一般ユーザも自由に閲覧することが可能となっている。しかし、公開された講義資料の多くは組織化されることなくばらばらに存在するため、従来の検索エンジンではユーザの目的に合った資料を見つけることが困難となっている。

本研究では、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、ユーザが求める教材が存在する Web ページを検索、閲覧できるポータルサイトを構築することを目的とする。対象ユーザとしては、情報工学専攻の学生だけでなく情報工学に興味のあるすべてのユーザを想定している。

本研究の特徴は、ユーザの入力を直接教材に結びつけるのではなく、情報工学の分野の学習内容を体系的に捉えるために整理した学習項目オントロジーを間に介することである。その結果、ユーザが学びたい事柄を分野全体の中で位置付けたり関連する項目を知ることができる。

2. システムの構成

システムの処理の流れを図 1 に示す。

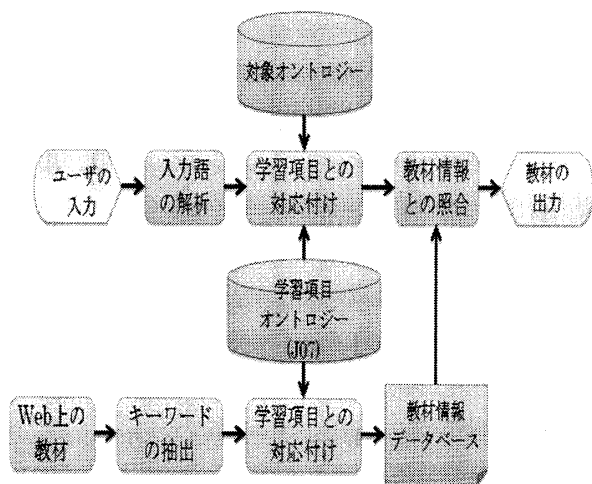


図 1: 処理の流れ

ユーザが学びたい事柄を入力すると、まずその入力語を形態素解析する。そして、その結果を使って学習項目オントロジーと対応付ける。また、あらかじめ学習項目オントロジーと結び付けられている教材情報データベースと照合し、入力語に適合した教材が存在する Web ページを選んでその URL を出力する。

3. 研究概要

3.1 学習項目オントロジー

学習項目オントロジーは、情報工学分野の学習内容を体系的に捉えるために、情報専門学科のための標準的カリキュラムである J07 におけるコンピュータ科学知識体系 CS-BOK-J 2007[2]に基づき作成する。この体系は 15 エリア(DS,PF…)、138 学習項目(DS1,DS2,…,PF1,PF2…)の階層構造になっているため、教材の分類、およびその検索を容易にすることができる。

表 1: 学習項目オントロジー

DS	エリア名	情報の基礎となる数学など
DS1	学習項目名	関数、関係、集合
	トピックス	集合、ベン図、補集合、デカルト積、べき集合
	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。

3.2 対象オントロジー

対象オントロジーは、学習内容の説明に出てくる用語に関する知識であり、IT 用語辞典[3]を用いて作成する。IT 用語辞典は tree 構造となっていて、「クイックソート」の上位概念は「アルゴリズム」となり、「アルゴリズム」の上位概念は「プログラミング」となる。対象オントロジーを用いることにより、子ノードの用語が学習項目オントロジーにないとき、親ノードの語で検索することが可能になる。

3.3 教材情報データベース

Web 上にある、情報工学分野の教材(pdf やパワーポイントファイル)がリンクされているページの URL を検索エンジンを用いて収集する。そして、収集したページにある教材からキーワードとなるような言葉を抽出する。そのキーワードと学習項目オントロジーに含まれる言葉を照合することにより、各 Web ページに対してそのページ上の教材で学習できるとされる学習項目を対応付ける。

Retrieval of Information Engineering Course Materials
Based on Associations with Learning Unit Ontology
Taro Takoshima, Taisuke Nishio, Toru Sugimoto
Department of Information Science and Engineering,
Shibaura Institute of Technology

3.3.1 キーワードの抽出と重み付け

Web ページにリンクされているすべての教材から名詞のみを取り出し、それぞれの名詞の出現回数を数える。それをすべての教材がリンクされている Web ページに対して行う。そして、各名詞に対して TF-IDF 法により重み付けを行う。

$$tfidf(p, w) = tf(p, w) \cdot idf(w)$$
$$idf(w) = \log \left(\frac{N}{df(w)} \right)$$

p: Web ページ w: 対象となる単語 tf: 単語の出現頻度
df: 単語を含む Web ページの個数 N: 総 Web ページ数

3.3.2 教材と学習項目の対応付け

教材がリンクされている Web ページから抽出したキーワードを学習項目オントロジー中の各学習項目に含まれる単語と比較し、該当したら学習項目に以下のような得点を与える。

Web ページ p における学習項目 u の得点

$$score(p, u) = \sum_{i=1}^n a_i \cdot tfidf(p, w_i)$$

ここで、 $w_i (1 \leq i \leq n)$ は Web ページ p から抽出したキーワード、 a_i は w_i が学習項目 u の学習項目名に等しい場合は 2 点、学習項目名ではなく、トピックス、または学習成果にある場合は 1 点とする。また、エリア名に含まれる場合は 1 点加算する。

score をすべての学習項目に対して求めて、得点が高い学習項目をこの Web ページと対応付ける。

3.4 学びたい事柄に基づく教材検索

ユーザが学びたい事柄を自然言語の文や語句で入力すると教材情報データベースを利用し、適合する教材を出力する。そのために、ユーザの入力を形態素解析し、その結果を学習項目と対応付ける。この対応付けには、対象オントロジーも活用する。そして、得られた学習項目を教材情報データベースと照合することにより適合する教材がリンクされたページを選んでユーザに対して出力する。

3.4.1 ユーザの入力と学習項目の対応付け

入力に含まれる単語を学習項目オントロジー中の各学習項目に含まれる単語と比較し一致したら、その学習項目に得点を与える。得点の与え方は、教材と学習項目の対応付けと同じである(ただし $tfidf(p, w)$ の部分を除く)。

また、入力に含まれる語が学習項目オントロジーに存在しない場合、対象オントロジーを使用して上位概念の単語に置き換えて学習項目オントロジーの単語と比較し、一致したら、その学習項目に半分の得点を与える。

3.4.2 教材 URL の出力

ユーザの入力と対応付けられた各学習項目に対して、教材がリンクされているページを教材情報データベース

を参照して選び、その URL を出力する。

4. 実験結果と評価

4.1 教材取得結果

検索エンジンを用いて教材を集めた結果、教材がリンクされている Web ページを 164 個取得した。その中で以下のような教材として不適当なページを除くと 70 個となった。

- ・ 1 つの Web ページ上に複数の科目の教材がある。
- ・ 教材ファイルのリンクが切れている。
- ・ 試験問題だけが載っている。

4.2 教材と学習項目の対応付け

各教材に対して前述の方法で対応付けた学習項目 (score の良い順に 1~3 個選ぶ) と、人手によって対応付けたものを比較してみた結果、以下ようになった。

表 2: 教材と学習項目との対応付けの比較

	上位 1 個	上位 2 個	上位 3 個
適合率	67%	59%	51%
再現率	18%	34%	46%

適合率が低い理由として、「語」「よう」「説明」などの一般的な名詞や、「の」「さ」「一」のような一文の語もキーワードとしてしまったことが挙げられる。また、再現率が低い理由として、1 つの教材に当てはまる学習項目が多く、そのすべてを対応付けることができなかったことが挙げられる。

4.3 学びたい事柄に基づく教材検索

ユーザが「整数を整列するプログラムを知りたい」と入力した場合、形態素解析が行われ「整数」「整列」「プログラム」の 3 つの名詞が抽出される。それらを学習項目オントロジーの単語と比較した結果、学習項目の上位 5 件「AL3(アルゴリズム設計例)」「AL1(アルゴリズムの解析の基礎)」「AL2(アルゴリズム設計手法)」「AL5 (高度なアルゴリズムの設計)」「AL4(アルゴリズムの高度な解析)」が対応付けられ、それらに関連した「データ構造とアルゴリズム」「アルゴリズム特論」のような教材が出力された。

5. おわりに

キーワードの抽出方法を見直すことにより、教材と学習項目の対応付けの精度を改善していきたい。また、J07 の各学習項目に割り振られたキーワードや学習項目間の関係を拡充することにより、各学習項目の内容をより的確に捉えた対応付けと応用を実現していきたい。

参考文献

- [1]OCW: http://www.jocw.jp/index_j.htm
- [2]情報処理学会: コンピュータ科学教育委員会 公開文献資料
<http://www.sb.cs.meiji.ac.jp/~hikita/csj2007/>
- [3]Yahoo!家電ナビ IT用語辞典:
<http://kaden.yahoo.co.jp/dict/>