シラバス HTML 文書からの情報抽出

渡辺将尚† 絹川博之† 井田正明‡ 芳鐘冬樹‡ 野澤孝之‡ 喜多 一††

東京電機大学 工学部† 大学評価・学位授与機構 評価研究部‡
京都大学 学術情報メディアセンター††

1. はじめに

近年,ネットワーク環境の普及,情報技術の発達により,多くの大学でさまざまな点で,電子化が進んでおり,シラバスがWeb を通じて公開されるようになっている.

公開されているシラバスの内容や HTML 記述形式は各大学独自のものとなっており,様々な大学のシラバスの比較検討が困難である.この問題点を解決することを目的に,シラバスに含まれている情報を抽出し,Web 上に公開されている各大学のシラバス HTML 文書から必要に応じた情報を抽出することを試みた.

2. シラバス HTML 文書の特徴

Web 上で公開されているシラバスの多くは HTML で記述され,種々の大学で独自形式である.大学によっては,学部ごとに形式が異なっていたり,さらに,学部の教科ごとに自由に作成されたりしているケースもある.

Web に公開されている各大学のシラバスには以下の特徴がある。多くの場合 1 科目に対してひとつの HTML ファイル（以下ではページと記す）が存在する.

(1)シラバスの構成要素

シラバスには,[科目名],[開講学年],[曜日],[開講学期],[単位数],[教官名],[目標],[講義内容]などといった 30 を超える項目が記述されているが,大学により記述されている項目とその内容はまちまちである.

(2)シラバスの項目の出現順序の傾向

大きく分けて３つの部分から成り立っている.

(a)ページの先頭部に[科目名]や[科目英文名]が現れる.

(b)ページの中間部に[開講学年],[曜日],[時限],[単位数],[教員名]などといった科目の基本事項が記述される.

(c)ページの末尾部に,[授業計画],[教科書],[成績評価方法]などが記述される.

ただし, (a),(b),(c)の各部分の中では項目の順序は大学によって異なる.

(3)項目ごとの特徴

(a)項目名を表す言葉の後に項目の内容を表す言葉がある.

(b)シラバスにおける特徴的な言語表現を含む.

(例 1)[開講学年]：語尾に”年”,”学年”,”年次”など.

(c)特徴的な言語表現を含む場合と含まない場合がある.

(例 2)[科目名]：語尾に”論”,”実験”,”概論”,”演習”など科目名独特の表現を含む場合がある.

(d)人名

(例 3)[教員名].

(e)特定の HTML のタグが繰り返し使用されている.

(例 4)[講義内容]：講義は通常 10～15 回構成であり,内容を記述する際にタグのパターンが同じ場合が多い.

3. 情報抽出方法

3.1 先行研究概観

(1)構造情報を利用した抽出方法.

(a)手作業でテンプレートを作成.

人手によって学習用記事から抽出情報を抜き出し,正解データを作成し,文書中の抽出情報を正解データと照らしあわせ正規表現記法「(.*?)」と置き換える.これによって一つ一つのページに適合した抽出用テンプレートを作成し,項目を抽出する方法[4].

(b)自動的にテンプレートを作成.

個々の HTML ページを行単位で比較し,差分を求めることにより,その HTML ファイルに対応している抽出テンプレートを自動的に作成し,項目を抽出する方法[1].

(c)ページの形式を利用.

HTML 文書のタグ情報を利用して,項目名,タグのキーワードとなるものを見つけ出し,位置関係を考慮し,抽出する方法[5].

(2)言語表現の特徴を利用した抽出方法.

(a)項目の特徴と出現順序を機械学習.

あらかじめ項目名の特徴的な言語表現や出現順序を学習させておき,それと類似した項目値を抽出する方法[3].

(b)内容部分（非タグ文字列）の特徴的な言語表現を利用[1],[2].

(3)「(1),(2)」の組み合わせた抽出方法.

例えば構造情報を利用して抽出した後で, 言語表現の特徴を利用して、情報を抽出する[5],[6].

3.2 本研究で提案するシラバスからの情報抽出方法

(1)タグを除去し残った文字列のみを抽出する.

HTML の全ての行を一行にした後に,タグに囲まれているものを抽出し,一つのファイルにまとめておき,ここで作成されたものを元データとする.

(2)特徴的な言語表現をもとに抽出する.

「2.(3)(b)」を抽出する際には,特徴的な言語表現を一つのテーブルにまとめ,そのテーブルと照合し,一致したものを抽出する.

(3)「2.(3)(c)」については,(2)と同様の方式で抽出できた抽出項目の,HTML 文書中の相対出現位置を記憶しておき, 推定抽出する.

(4)「2.(3)(d)」については辞書を利用して抽出する.

人名を抽出する祭には,まず最初に茶筌で形態素解析をする.その結果で,＜固有名詞―姓＞の次に＜固有名詞―名＞の順番で出現するものは,人名だと特定できる.

人名と特定できたら,ページの中の相対位置を調べ,出現予測位置とする.茶筌の辞書に収録されていない人名については,人名辞書を利用して取り出す.

Information Extraction from Syllabus HTML Documents
Masanao Watanabe† Hiroshi Kinukawa† Masaaki Ida‡
Fuyuki Yoshikane‡ Takayuki Nozawa‡ Hajime Kita††
School of Engineering, Tokyo Denki University†
Faculty of University Evaluation and Research
National Institution for Academic Degrees and University Evaluation‡
Academic Center for Computing and Media Studies, Kyoto University††

(5)項目名を表す言葉を基に項目値抽出する.

どの大学もシラバスとして書くべき内容はほぼ同じであり,項目名を表す語はどの科目にも共通なものが使われている.例えば[授業内容],[目標],[教科書]など.そこで語の,大学内の各ページでの共通出現頻度を求め,ある数値以上複数ページで出現した語を抽出する.この抽出した語を項目名と推定する.その項目名の後に項目値が来ることを利用し項目値を抽出する.

(6)HTML のタグを利用して抽出する.

現在大学の HTML はホームページソフトを利用して作成されていることが多いので,同じタグが繰り返されることが多い,例をあげると

(例 5)<TDWIDTH="40%"VALIGN="TOP"BGCOLOR="White"><FONT SIZE="-1">収束点列</FONT></TD>

のように大学独自の特定形式のタグが繰り返し現れる.「2.(3)(e)」で書いたように,同じタグのものは関連性があると考えられるので,抽出の手がかりとして利用できる.

3.3 先行研究との比較

先行研究では,抽出するページ形式がほぼ同じものを対象としているが,本研究ではさまざまな場合に対応可能なプログラムを作成することが目的であるため,一つの方法では抽出することが難しい.よって上に挙げたような様々な方法を組み合わせて抽出する.

4. 実験、結果

13 大学の情報系学科のシラバスを対象とし,情報抽出の実験を行った.対象としたシラバス HTML の記述形式は大学ごとにまちまちである.

今回の実験では,特徴的な言語表現(「3.2(1) ～(4))の方法. ただし今回の実験では人名抽出は茶筌のみで抽出した.)を利用し,抽出する項目を,[科目名],[科目英文名],[開講学年],[開講学期],[曜日],[時限],[必修等],[単位数],[教員名]の９つの基本項目に絞って抽出している.各項目の抽出結果を表１に示す.

表 1 実験結果

項目名	出現数	抽出数	正解数	再現率	精度
科目名	867	867	867	100%	100%
科目英文名	767	758	758	98.8%	100%
開講学年	817	809	809	99.0%	100%
開講学期	693	667	665	96.0%	99.7%
曜日	305	287	287	94.1%	100%
時限	349	331	325	93.1%	98.2%
必須等	231	225	225	97.4%	100%
単位数	760	679	677	89.9%	99.7%
教官名	1208	937	709	58.7%	75.7%

$$\text{再現率}=\frac{\text{正解数}}{\text{出現数}}(\%) \qquad \text{精度}=\frac{\text{正解数}}{\text{抽出数}}(\%)$$

5. 考察

(1)今回の実験では,データ観察に基づき,手がかりとする言語表現を人手で記述したため,ある程度柔軟な対応が可能であり,[教官名]以外は再現率 89.9～100%,精度 99.7～100%,[教官名]の再現率は 58.7%,精度は 75.7%となる結果が得られた.しかし,大学内においても項目名を表す語句が統一されてない場合など例外的なケースでは,必ずしも抽出が成功するとは断言できない.

(2)[時限]の再現率が 93.1%の理由としては,例外的な語句(例 6)が出てくるためである.

(例 6)"集中講義","夏季休業中","不定期","隔週"など.

今回は例のような語句を抽出の対象としていなかったため,今後取り出せるよう改良したい.

(3)[教官名]の再現率 58.7%,精度 75.7%との理由としては,やはり茶筌のみで抽出しているため茶筌に組み込まれている辞書にない名前や外国人の名前などは抽出できないという欠点がある.今後は人名辞書を利用して精度の向上を図る予定である.

(4)この実験では初めタグに囲まれた一つの語句を一つの項目に関するものとみなしている.[単位数]や,[開講学年]などは,数字,大学独自の表記しか入っていない場合もあり,誤りを発生させている.

(例 7)C2,A1 など.

(5)今回の実験は[科目名],[科目英文名]以外は語句情報のみで抽出したのでこのような結果になったが,全ての抽出項目に対して,出現位置を考慮し抽出する範囲を狭めることによって,より良い精度,再現率が期待できる.

6. おわりに

本研究の成果として,実験で対象としたシラバスでの比較的形が決まっている項目については,語句情報,出現順序だけも,かなり良い精度,再現率が得られる.今後は,より多くの大学の HTML データについて実験し,精度と再現率を高めたい.また,今回のプログラムでは抽出する項目が例外的語句だった場合取り出すことができなかった.今回は利用しなかった,項目名を表す言葉を基にした抽出方法,HTML タグを利用した抽出方法を組み込み,より柔軟でかつ汎用的なシラバス情報抽出プログラムを作成したいと思う.

7. 参考文献

[1]伊東, 松永, 山田, 廣川：Web シラバスからの DB 構成, 第 17 回人工知能学会全国大会, 1D4-08, 2003

[2]伊藤,山田,廣川："Web シラバス統合のためのレコード解析", 人工知能学会 研究会資料 SIG-SWO-A201,pp.(05-1)-(05-7),2002

[3]板井,高須,安達：HTML からの情報抽出と統合,NII Journal No.6(2003.3)P9～P19

[4]井出,藤吉,長井,中村,野村：テンプレートを用いた新聞記事からの製品情報抽出システム情報処理学会 NL 研究会, 96 NL 115-12, pp.83-90, 1996.

[5]富田,手塚,山本,長岡：HTML 文書からの商品情報抽出方式の提案,電子情報通信学会技術研究報告,KBSE97-27,p.15-22

[6]原田,風間,佐藤：参照 HTML テキストからの Web サイト紹介文抽出,情報処理学会第 63 回（平成 13 年度後期）pp(3-39)-(3-40)

[7] 松本,北内,山下,平野,松田,浅原: 日本語形態素システム『茶筌』version 2.0 使用説明書第 2 版, NAIST Technical Report,NAIST-IS-TR99012,奈良先端科学技術大学院大学,1999