4P-4

シラバス文書からの情報抽出支援システムの試作

渡辺将尚† 絹川博之† 芳鐘冬樹‡ 井田正明‡ 野澤孝之‡ 喜多 一††

東京電機大学大学院 工学研究科† 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

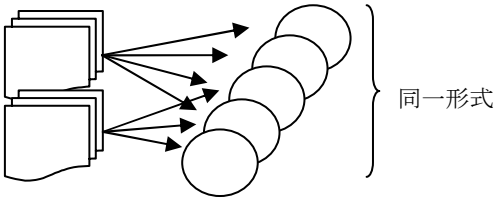
近年，ネットワーク環境の普及，情報技術の発達により，多くの大学で電子化が進んでおり，シラバスも Web を通じて公開されるようになっている．それらのシラバス文書から適切に情報を抽出し構造化できれば教育内容に関する検索，分析や教育課程の比較，設計など学生，教員等にとって有効な活用が期待される．発表者らは，文書の構造情報と言語情報を利用した自動抽出手法をベースとし，手動のフェーズを加え，ユーザの判断を反映させることで，より高いパフォーマンスで抽出を行うことが可能なシステムを試作した．本発表では，試作した情報抽出支援システムについて報告する．

2. シラバス HTML 文書の特徴

2. 1 Web 上の大学ページの特徴

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，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が，多くは個々の科目の内容を詳細に記述した科目別シラバスページと，科目名の一覧を記述した目次となるシラバス一覧（リンク先がシラバスページ）から構成されている（図 1）．



シラバス一覧　　科目別シラバスページ

図 1　大学シラバス構造

このとき，科目内容を記述した科目別シラバスページは，同一大学内では記述形式(書式) が類似している場合が多い．また，シラバス一覧のページは，大学内に複数存在する場合もある[2].

2. 2 科目別シラバスページの特徴

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 を超える項目が考えられるが，大学により記述されている項目とその内容は様々である．

しかしながら，ほとんどの場合，共通の特徴として次のような傾向が見られる．

(1)項目名を表す言葉の後に項目の内容を表す言葉がある．

(2)特徴的な言語表現を含む項目がある．

(例 1)[開講学年]：通常，語尾に「年」，「学年」，「年次」などを含む．

(例 2)[科目名]：多くの科目は，語尾に「論」，「実験」，「概論」，「演習」など科目名独特の表現を含む．

3. 情報抽出支援システムの構成

3. 1 システムの概要説明

今回試作したシステムは，抽出結果から繰り返し人手を加えることにより，高いパフォーマンスで抽出可能となるシステムである．

抽出支援システム構成の概要を図 2 に示す．

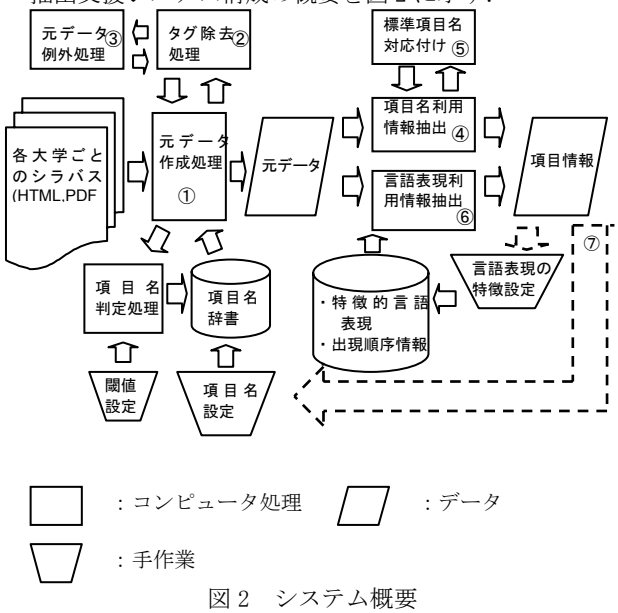


図 2　システム概要

3. 2 元データ作成処理

HTML ファイルからタグを除去し，<項目名><項目内容>…<項目名><項目内容>の順になるように語句情報のみを抽出する．このデータを元データとする．（図 2 ①）

(1) 項目名の判定方法

項目名を表す語はどの科目にもほぼ共通なものが使われている．この特徴を利用し，出現頻度が上位の語を項目名と判定する．(図 2 ②)

項目名と判定する出現頻度の閾値は，事例により適切な値にばらつきがあるので，大学ごとに閾値を設定する．ある大学のシラバスページ（総数 51）に出現する語を頻度（出現ページ数）順に並べたものを表 1 に示す．例えば，閾値をページ総数×0.6 とすると，出現頻度「31」以上の 14 語が選ばれる．そのうち「理工学部情報化学科」「2 単位」などの項目名となりえない語は，あらか

An information extraction support system from syllabus documents

Masanao Watanabe† Hiroshi Kinukawa† Fuyuki Yoshikane‡ Masaaki Ida‡ Takayuki Nozawa‡ Hajime Kita††

Graduate School of Engineering, Tokyo Denki University†

Faculty of University Evaluation and Research NIAD-UE (National Institution for Academic Degrees and University Evaluation)‡

Academic Center for Computing and Media Studies, Kyoto University††

じめ辞書に登録しておき除外するものとする．
以上の処理により，「成績評価方法」，「授業計画」，「教科書」，「授業内容」，「目標」，「履修上の注意」，「開講学年」，「教員名」，「開講学部」，「科目種別」，「時間数」，「単位数」の 12 語を項目名と判定する．

表 1　シラバス出現頻度の上位語

出現数	出現内容	出現数	出現内容
51	成績評価方法	51	開講学部
51	授業計画	51	科目種別
51	教科書	51	時間数
51	授業内容	51	単位数
51	目標	31	2 単位
51	履修上の注意	20	必修
51	理工学部情報科	12	選択
	学科	11	武田正之
51	開講学年	10	榎本進
51	教員名	10	1 組

(2) <項目名><項目内容>の順に抽出する方法

シラバスの多くは<項目名><項目内容>…<項目名><項目内容>の順に記述されているので，その記述順を保てば良い．ただし記述順が<項目名><項目内容>の繰り返しと異なる，テーブルタグを利用した表形式のケースがあり，これは以下のように処理する．(図 2 ③)

次の例を用いて説明する．

```
<tr><th>開講学部</th><th>開講学年</th><th>時間数</th><th>
単位数</th><th>科目種別</th></tr>
<tr><td>工学部ＡＢコース</td><td>２年３学期</td><td>２
</td><td>２</td><td>選</td></tr>
```

単にテーブルタグだけを消去すると

「開講学部」，「開講学年」，「時間数」，「単位数」，「科目種別」，「工学部ＡＢコース」，「２年３学期」，「２」，「２」，「選」となり，<項目名><項目内容>…の順にならない．

このような表では，一般に下記のように表頭の項目名と同順に項目内容が並んでいるので，

```
<tr><th>開講学部</th><th>開講学年</th><th>時間数
</th><th>単位数</th><th>科目種別</th></tr>

<tr><td>工学部ＡＢコース</td><td>２年３学期</td><td>２
</td><td>２</td><td>選</td></tr>
```

「開講学部」：「工学部ＡＢコース」，「開講学年」：「２年３学期」，「時間数」：「２」，「単位数」：「２」，「科目種別」：「選」と対応付け，<項目名>に対応する<項目内容>を抽出することができる．

3. 3　項目名を利用した情報抽出

3.2 節により元データは<項目名><項目内容>となっているので，項目名さえ判定できれば，言語表現の特徴だけでは特定しにくい「目的」，「概要」などの項目も容易に抽出できる．(図 2 ④)

3.2 節 (1) で選定した項目名の表記は，大学ごとにゆれがあるので，これまでのシラバス分析の経験[3]にもとづき作成した項目名辞書との照会により，各項目名が対応する NIAD-UE の標準項目名[1]を決定する．それに従って，直後の情報がその標準項目名に対応する項目内容として抽出される．(図 2 ⑤)

3. 4　特徴的な言語表現を元にした情報抽出

シラバスの記述に項目名が存在しないもののうち抽出項目の内容が予測できるものについては，あらかじめ項目内容の特徴的な言語表現や出現順序を学習させておき，それらの言語的・順序的条件を満たす情報を項目内容として抽出する方法を提案する．(図 2 ⑥)

3. 5　ユーザの判断を反映

今までのシステムでは例外的な語句が存在すると抽出が困難であったが，今回提案するシステムでは，結果を参照し，システムを数回訂正することにより．より高いパフォーマンスが期待でき，訂正はシステムに反映されるため，抽出する大学を増やすにつれ，効率的に抽出が行える（システムの学習）という特徴もある．
（図 2 ⑦破線部）

3. 6　出力結果

試作システムを用いて，NIAD-UE の XML スキーマ[1]で定義されている項目のうち 14 項目を選定し，大学シラバスから抽出した．選定項目は，「科目名」，「英文科目名」，「単位数」，「教員名」，「開講学期」，「時限」，「概要」，「目標」，「評価方法」，「教科書」，「参考書」，「授業計画」，「授業形式」，「その他」である．システムの出力結果を図 3 にて示す．

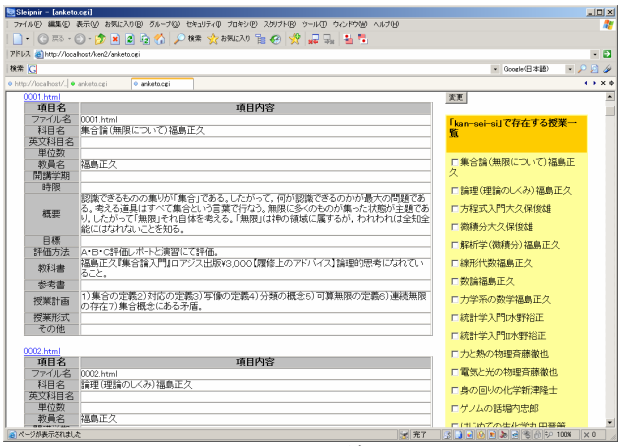


図 3　出力結果

4. おわりに

大学シラバスから半自動的に情報抽出を行う支援システムを試作した．本システムでは，ユーザの判断を反映させるため，①例外的な項目名への対応と，②閾値の適切な設定が可能となり，シラバス文書の特徴に応じて情報抽出が行える。元データファイルを語句情報のみにした理由は，シラバス文書が HTML ファイルだけでなく PDF ファイルの場合があり，この違いを統一するためである．よって，今後は PDF ファイルからの情報抽出も視野に入れシステムの完成を目指す．また，抽出対象とする項目数を増やし，より実用的なシステムを作る予定である．

参考文献

[1]井田，宮崎，芳鐘，喜多:"シラバス XML データベースシステム構築に関する考察"，情報処理学会第 65 回全国大会講演論文集，p. 4/247-4/248 (2003)

[2]山田，伊藤，庵川:"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，p. 5/1-5-7 (2002)

[3]渡辺，絹川，井田，芳鐘，野澤，喜多:"Web 上のシラバス情報の収集と XML 変換"，FIT2004 第 3 回情報学術フォーラム，p. 2/121-2/122 (2004)