E-005

Ｗｅｂ上のシラバス情報の収集とＸＭＬ変換

Collection and XML conversion of syllabus information on Web

渡辺将尚† 絹川博之† 井田正明‡ 芳鐘冬樹‡ 野澤孝之‡ 喜多一††
Masanao Watanabe Hiroshi Kinukawa Masaaki Ida Fuyuki Yoshikane Takayuki Nozawa Hajime Kita

1. はじめに

大学のシラバスは，多くがWeb上で公開されている．

これらは大学教育についての価値ある情報源であるが，

(ｉ)公開場所がインターネット上で分散しているため収集に手間がかかること．

(ii)形式が各大学で独自のものとなっているためシラバス内容の横断的比較が難しいこと，

が利用の妨げとなっている．

これらの問題を解決するため，

(Ⅰ)Web上からシラバスを効率よく収集する方法．

(Ⅱ)収集してきたHTMLシラバス文書から，使用語句とその出現位置情報，使用されているタグ情報，に着目して必要な情報項目を抽出し，XMLスキーマで規定された共通の形式へ変換する方法．

を提案する．

2. Web上のシラバスデータの収集

2. 1 Web上からのシラバス収集

シラバスは大学の授業内容を示すものだが，授業内容の変更に応じて年次更新される．そのため手作業での収集は負担の大きな課題となるが，自動収集が可能なら手間が大幅に省ける．

2. 2 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，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．

大学のサイトは一般に，リンク構造になっているが，シラバスへのリンクの位置も複数あり，どの位置に存在しているのか予測も不可能である．

シラバスというものは，シラバスが1ページで存在するわけではなく，シラバスの一覧などを示すトップページが存在し，その下の階層にシラバスが複数存在する．

そこで収集の手間を省くためにもシラバスのトップページを検出できることを目的とする．

2. 3 シラバストップページ取得方法

シラバス収集としてまず最初に約800大学のURLを取得した．これはインターネット上にあった大学一覧というページから大学のURLだけを取得した．その後，得られたURLからリンクをたどり下の階層へ次々と行く．その際，抽出の対象となっている大学からリンクされた学外ページには飛ばない，一度得たURLは重複しないようにし，得るデータの種類は，拡張子が[. html][. txt][. jsp][. asp]であるものを対象とした．

シラバストップページの判定方法は次のようにした．

(1)リンクにシラバスと書かれているもの．

リンクにシラバスと書かれているなら，そのリンク先はシラバスと判定する．

(2)語句情報で判定する方法．

(a)ページにシラバス，もしくはシラバスの同義語が表示されていて，かつそれがリンクされていなかった場合，それ以降のリンクを調べ，そのリンク先にシラバス特有の表現が含まれているもの．

(b)(a)以外の場合の時．

リンク先のページより深い階層までのページを取得し，そこにシラバス特有の語句が存在したら，その大元となっている，シラバスのトップページが得られるもの．

2. 4 先行研究との違い

先行研究[2]ではgoogleAPIなどを利用し，シラバスを収集しているため，大学以外のシラバスも多く含むと予想される．

提案方式ではまず最初に大学のURL一覧を作り，そこから収集しているので，シラバスを探してくる範囲が狭まり，精度，効率が上がる．

3. シラバス文書からの情報項目抽出

3. 1 シラバス文書の特徴

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．

シラバスには特徴的な語句が多く含まれている．

(例1)[開講学年]：語尾に"年"，"学年"，"年次"など．

3. 2 必要な情報項目の抽出

シラバス項目として，以下の表のような14項目を選定し，大学評価・学位授与機構の提供しているXMLスキーマ[1]に定義されているタグ名をつけるものとする．

表1. XMLタグ形式

項目名	タグ名	項目名	タグ名
科目名	title	単位数	credit
科目英文名	e-title	教官名	name
開講学期	year	概要	abstract
開講学期	term	評価方法	evaluation
曜日	day	参考書	references
時限	time	教科書	textbooks
必修等	required_selective	目的	course-objectivse

3. 3 テンプレートを利用した抽出

以前の発表[4]では，[科目名]，[教官名]など，いくつかの基本的な項目について，特徴的な言語表現を利用した抽出手法を適用し，非常に高い精度・再現率で抽出が可能であることを示した．しかし，言語表現の特徴を特定しにくい[目的]，[概要]などの項目は抽出が困難であ

† 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

るためページの形式を利用して抽出する以外はないと考え，テンプレートを作成し，抽出する方法を提案する．

先行研究でのテンプレート生成は，タグを利用したテンプレート生成法が主流であるが[3]，今回の提案方式では，言語情報を利用してのテンプレート生成方法である．

(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のみを抽出し，一つのファイルにまとめておく．

(2) テンプレート生成．

(1)で生成した語句情報だけで，大学どうしを比較すると，記述順番や，形式は違うが，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ほぼ，共通なものが使われている．次に同一大学の中でのシラバスを比較する，記述順序，形式，項目を表す語句も共通している．そこで，大学内での各ページの差分を利用し，項目名を表す語を抽出し，各大学独自のテンプレートを生成する．その例を図１に表す．

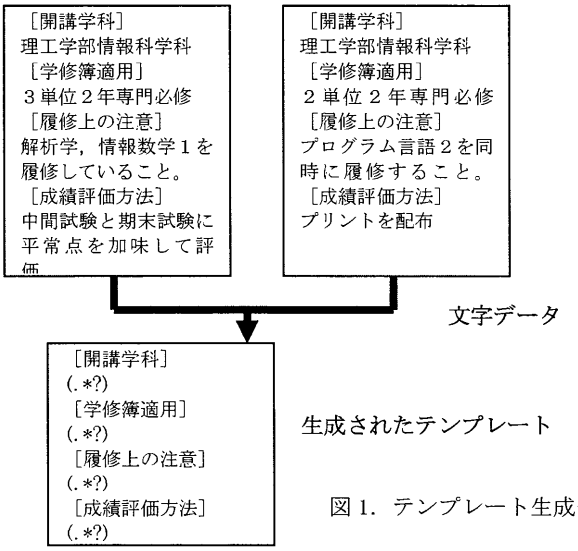

図 1．テンプレート生成例

4. XML 変換

抽出された結果を XML 形式に変換する．ここでのタグは表 1 に対応し，タグ付けする．ここでのタグ付けの際，例えば，教科書の場合，"教科書"，"教材"，"テキスト"などを同義語として同一タグに変換する．

5. 実験と結果

5. 1 シラバストップページ取得

日本の 15 大学の大学を対象とし，各大学につき 5 階層調べた．そして 571920 ページのファイルを収集した．この 15 大学は必ず大学全体のシラバスの元となるトップページが存在する．

そこで，シラバストップページが，当該大学のトップページから何階層目にあるかの分布を図 2 に示す．

5. 2 テンプレートを利用した項目抽出

日本の 13 大学の情報系学科のシラバスを対象とし，項目抽出実験評価した．対象としたシラバス HTML の記述形式は大学ごとに異なっている．今回は，言語表現の特徴に基づくパターンマッチングでは抽出が困難な[概要]，[評価方法]，[参考書]，[教科書]，[目的]の 5 つについての項目抽出結果を表 2 に示す．

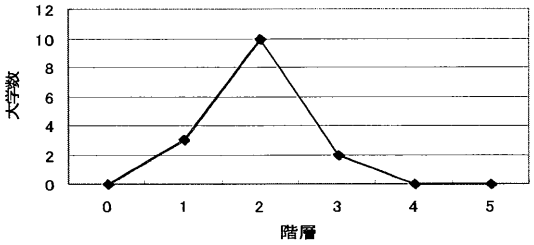

図 2．シラバストップページ分布図

表 2．　シラバス項目抽出の結果

項目名	出現数	抽出数	正解数	再現率	精度
概要	356	356	356	100%	100%
評価方法	685	685	685	100%	100%
参考書	639	512	512	80.3%	100%
教科書	288	247	247	85.8%	100%
目的	475	393	393	82.7%	100%

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

6. 考察

シラバストップページの取得という点については 3 もしくは 4 階層調べればよいと分かった．これは，あまり深い階層にトップページがあると，利用者にわかりづらいという点から来ていると考えられる．今後全国の大学のホームページについて適用し，処理方式の改善を図る必要がある．

項目抽出としては，大学単位でシラバスを比較すると必ず共通項目名が存在するので，それを利用して生成したテンプレートを用いて項目抽出することも可能となった．しかし，項目名を表す語句が存在しないで，項目の内容が出現する例外的なケースもまれにあったので，その際の抽出方法を以後検討する必要がある．

7. おわりに

今回の実験でシラバスの，収集，抽出，XML 変換を行った．収集では無駄なページを調べ時間がかかり過ぎたため，収集時間の短縮の必要がある．テンプレートを利用した抽出は精度は良いが再現率が十分ではないため，改良が必要である．今後，対象大学を増やし，柔軟かつ汎用的なプログラムを生成していきたい．

8. 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.

[2]山田，伊藤，庵川:"Web 上に公開されたシラバス情報の自動収集"，DICOMO2002.

[3]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.

[4]渡辺，絹川，井田，芳鐘，野澤，喜多:"シラバス HTML 文書からの情報抽出"，情報処理学会第 66 回全国大会講演論文集 p. 4/487-4/488.