Webシラバス群のデータ形式統合に関する考察

伊東栄典 * 　竇ギョク峰 † 　廣川佐千男 *

∗ 九州大学情報基盤センター
† 九州大学大学院システム情報科学府
〒 812-8581 福岡市東区箱崎 6-10-1, (092)-642-4037
{itou@, douai@matu, hirokawa@}.cc.kyushu-u.ac.jp

A study for schema integration for Web syllabi

Eisuke Itoh* 　Yufeng Dou† 　Sachio Hirokawa*

∗ Computing and Communications Center, Kyushu University.
† Graduate School of Information Science and Electrical Engineering,
Kyushu University

1 はじめに

Web の広がりに伴い，Web から体系的あるいは特定の目的に合致する情報を収集・分類する Web マイニングの研究が進んでいる [13]。HTML を代表とする半構造化データから知識を抽出する研究や，特定のテーマに関する情報を収集する研究が行なわれている。

一方，情報技術を教育に応用する e-Leaning が様々な教育機関で普及しつつある [2, 7, 11]。その一貫として，Web を介してシラバスを公開する教育機関が増えている。シラバスは教育内容についての詳細な情報を持つため，各教育機関の Web シラバスを統合できれば，教科書の調査や，教育内容の比較等に利用する事ができる。

我々は，Web マイニングの応用として，Web に公開されているシラバスを対象とし，Web シラバスを自動的に統合し，検索などに利用する研究を進めている [3, 9, 16]。シラバス Web ページから科目の各属性の内容を抽出し，それを DB に格納する事で，さまざまな知識獲得に利用できる。現在 Web 上に公開されているシラバスは，各組織が個別に作成したものであり，書式は統一されていない。ため，検索など系統的に利用するには，非均質な HTML シラバスファイルを標準的な書式に変換する必要がある。

本論文では，独自の書式で公開されている非均質な Web シラバスファイル群を，標準 XML スキーマに沿った XML ファイルへ変換する手法を提案する。変換後に用いる記述形式として，大学評価・学位授与機構 (以下，NIAD) が開発しているシラバス XML スキーマを用いる [4, 10]。

NIAD シラバス XML スキーマは 41 個の要素名を持ち，シラバスや授業概要について詳細な記述が可能である。次に，HTML で記述された Web シラバスを NIAD シラバス XML スキーマに変換するシステムを試作した。実際のデータをこのシステムに適用し，変換手法の提案評価も行なった。

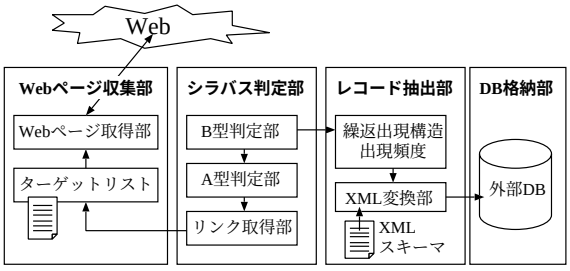

図 1: システム概要

2 Webシラバス統合

我々が開発している Web シラバス統合システムの概要を述べる。図 1 にシステムの概要図を示す。

まず，Web シラバス統合のためには，情報を記述するためのメタデータ形式が必要である。シラバスデータを記述するためのメタデータ形式については，NIAD で開発された XML スキーマが存在する。これを利用することにしている。

次に，Web からシラバスページを収集する必要がある。現在，松永らは Web から特定トピックのシリーズ型ページを集中的に収集するクローラーを開発している [9, 16]。国内高等教育機関のシラバスを対象にクローラーを動かし，約 5 万ページのシラバスを収集している。

三番目に，集めたページ群から個々の科目の情報をレコードとして抽出する必要がある。シリーズ型のページ群からレコードを抽出するために，様々な方法が考案されている [8]。例えば，ラッパーの自動生成や，テンプレートの抽出，XML への自動変換などが研究されている。

集めた科目情報を系統的に利用するためには，独自形式で記述されたデータを，統合的な記述形式への変換する必要がある。本研究は，この変換を対象としている。詳細は後述する。

最後に，統合したシラバス情報から，何らかの知識を提供す

るシステムが必要となる。知識の提供には，その分野の専門家がどのような知識を求めているかについて分析する必要がある。現在の所，具体的な知識提供方法については考察していない。様々な研究の調査と，色々な分野の意見を集めて，知識提供システムを構築していく予定である。

3 関連研究

Web からの特定トピックに関する情報の収集および統合に関して，様々な研究が行われている。幾つかの関連研究について述べる。

3.1 ラベル割当て

Wang ら [15] は，DB 情報を提示する Web サイトのページテキストから，元のデータを推定する方法について研究している。Wang らの手法は，関係データベース (以後，RDB と記述) としてのスキーマを明示的に持つサイトについての情報統合である。また，一つのサイト内だけを対象としている。他にも，Arasu[1] ら他多数の研究者がデータ抽出およびラベル割り当てについて研究している。

本研究では，一つのサイトではなく，多数サイト (教育機関) を対象としている。また Web 上に提供されるシラバスは多くの場合，RDB としてのスキーマが明示的に定義されていないため，属性名と属性値の対応を取るラベル割り当ての方法が必要となる。

3.2 スキーマ・マッチング

独自形式で記述されたデータを，別の形式に変換する問題は，スキーママッチングとして従来から様々な手法が研究されてきている [12]。

本研究では，独自形式 (スキーマ) で記述されたシラバスを，統合に用いるある一つの形式に変換することを目的としている。元の記述形式が完全に判別しているならば，スキーマ・マッチング問題の一例として扱う事ができる。

しかしながら，本研究が対象としているデータは HTML で記述された Web 上のデータであるため，元データの記述形式を推定する作業が必要となる。そのため単なるスキーマ・マッチング問題とは異なる。

4 Web シラバスの変換

Web シラバスデータを，標準的な形式 (NIAD シラバス XML スキーマ) に変換する手法について述べる。

4.1 シリーズ型 Web 文書

本論文の前提となるシリーズ型 Web 文書について述べる。Web 内には，同一の書式を持つ同種の文書が固まって存在する場合がある。新聞記事ページ，料理レシピ一覧，賃貸住宅一覧などが，その例である。このような Web 文書は，組織や個人が統一的な意思を以って作成された文書であるか，あるいは統一的スキーマをもデータベースから自動生成された文書である。

これら一連の文書を，岩沼らは「シリーズ型 HTML 文書」と呼び [14]，シリーズ型の文書から，レコードを抽出し，XML に変換する手法について研究している。シリーズ型の文書群は，統一的な書式として作成されるものであるため，記述書式の情報を与えることで，HTML 文書群から元のレコード情報を抽出することが可能になる。

Web シラバスは，同一サイトでは同一の書式に沿っている場合が多く，シリーズ型の文書である。シリーズ型文書の書式 (テンプレート) 抽出ができれば，レコードとなるデータ部分を扱う事ができる。シラバスには多くの場合，科目名，講義担当者名，概要などの，属性名となる文字列が記述されている。サイト (教育組織) の違いにより，シラバスの記述内容 (属性名) に揺らぎはあるものの，おおよそ意味的に類似した構造を持っている。属性名・属性値が連続するという場合は，比較的簡単にラベル割り当てが簡単に実現できる場合もある。しかしながら，属性名となる文字列が存在しない場合も多く，その場合，属性値となるべき文字列の特徴から，属性名を推定する必要がある。

4.2 NIAD シラバス XML スキーマ

Web シラバスから，科目内容をレコードとして抽出し，それを DB に格納する事で，さまざまな知識獲得に利用できる。統合して DB に格納する場合，格納するための形式が必要である。

教育情報コンテンツでは，内容を説明するためのメタデータ形式 (LOM:Learning Objcet Metadata) が考案されている IEEE1484 LTSC の提案する LOM[?] や，教育情報ナショナルセンター (NICER) の提案する LOM[11] などが提案・利用されている。しかしながら，これらは教育用コンテンツ (電子教材) を説明するための形式で，授業内容の説明には向いていない。

大学評価・学位授与機構 (NIAD) ではシラバスを記述するための XML スキーマを開発し，公開している [4, 10]。NIAD シラバス XML スキーマは 40 以上の要素名を持ち，シラバスや授業概要について詳細な記述が可能である。

各組織が公開している Web シラバスを，NIAD シラバス XML に変換できれば，標準的かつ統一的な書式でデータを扱うことが可能になる。本研究では，各組織の独自書式で記述された HTML Web シラバスを，NIAD シラバス XML スキーマに沿った形に変換する事を目指す。

4.3 変換手法

本論文では，HTML で書かれた Web シラバスから，レコードとなる科目情報は抽出されたものとして，抽出された科目データを，統一形式に統合する方法について述べている。

入力は，Web シラバスから科目情報をレコードとして抽出した CSV 形式のデータとし，出力は，各レコード (科目情報) を格納した NIAD XML シラバス形式の XML ファイルとする。図 2 に示すように，入力データは，ラベル部 L と，レコード部 R からなる。

NIAD シラバス XML スキーマの要素名は 41 個ある。その中で，科目情報の記述につかう要素名は，次の 30 個であり，それを集合 E で表す。

E = { code, title, eTitle, year, termSystem, term, day, time, requiredSelective, credit, classType, room, abstract, prerequisteCompetences, prerequisteCources, corequisteCources, courseObjectives, evaluation, textbook, references,

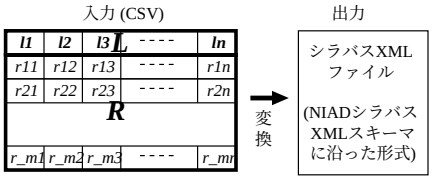

図 2: 変換の入出力

`remark, name, tel, e-mail, website, office, officeHour, preparation, topics, assignment`}.

統合形式への変換は，CSV で書かれた表型のデータのラベル部 L と，NIAD シラバス XML スキーマの要素集合 E との，対応に集約される。ただし，ラベル文字列が存在しない場合もあるため，その場合はレコード部から対応要素を推定する事になる。

そこで，E の各要素と対応するラベル文字列の候補の集合 (N) と，その要素のデータとなる文字列に特徴的なパターン (正規表現) の集合 (X) を用意し，それから，対応を取る事を考えた。表 1 にラベル文字列候補集合 N を，表 2 に E のデータ文字列パターン集合 X を示す。

表 1: E の各要素に対応するラベル文字列候補集合 N

要素名	対応候補文字列集合 N
code	授業コード, 授業番号, 科目番号
title	科目名, 講義名
eTitle	英文科目名
year	開講対象学年
termSystem	学期制, 半期, 四半期, 通年
term	開講学期
day	曜日
time	時限
requiredSelective	必修, 選択, 必選
credit	単位数, 単位
classType	授業形式, 講義, 演習, 実習, 実験
room	教室
abstract	授業概要
：	：
evaluation	成績評価方法
textbook	教科書
references	参考書
remark	その他
name	教官名
：	：

変換のアルゴリズムを以下に示す。

1. L の各文字列と，E に対応する候補文字列 (N) をマッチ。対応が取れる場合，ラベル l_i の列は，要素 e_j のデータとする。
   (if l_i =~ each n_j then $l_i \leftrightarrow e_j$)
2. 1. で対応が取れない場合，ラベル文字列の特徴を表すパターン集合 R の各要素とマッチ。
   (elsif r_ik =~ x_j then $l_i \leftrightarrow e_j$)

表 2: E のデータ文字列パターン集合 X

要素名	対応候補の正規表現
code	`/^\w\{2,5\}$/`
title	`/(.*論$)\|(論\w?$)|(.*研究$)\|(.*学$)|(.*演習$)\|(.*\Q ゼミナール\E)\|(.*学 (?!期)(..)?$)/`
eTitle	`/[a-zA-Z]{6,}/`
year	`/(年次)$\|年$/`
termSystem	`/^\d 学期|\^通年/`
term	`/^前 (..)?期|^後 (..)?期/`
day	`/曜/`
time	`/(.. 限)$\|(.. 時限)$/`
requiredSelective	`/^選|^必/`
credit	`/\d(単位)/`
classType	`/^(講義)$\|^(演習)$/`
e-mail	(RFC821.822 に準ずる形式)
website	(RFC1738 に準ずる形式)

5 実験

クローラー [9] により収集されたシラバスデータの中で，ランダムに 10 サイト (サイト毎に約 10 ページ) を選び，変換実験に用いた。表 3 に実験に用いたサイトを示す。

表 4: ラベルのマッチング率

データ No.	$\|L\|$	$\|L'\|$	$\|L' \cap E\|$
01	11	14	12
02	16	20	15
03	12	15	14
04	20	22	15
05	11	14	11
06	17	24	16
07	12	14	13
08	15	20	18
09	14	14	12
10	10	11	9

$|L|$ ：ファイルが実際に存在しているラベルの数
$|L'|$ ：意味的に正しく分割した場合のラベル数
$|L' \cap E|$ ：E の要素と一致したラベル数

表 4 と表 5 に，実験の結果を示す。全体の正解率は 89%になっており，ある程度高い正解率を得られた。ミスマッチおよびマッチ対象要素が無い部分に関する原因は，1 フィールド内に二つ以上の属性データが入っている場合である。現在の手法では，二つ以上のデータを持つフィールドを適切に分割する方法を持っていない。また，ただし，今回の手法は，N および X を元のデータを参考にしつつ人手で作成しているため，公平な評価とは言い難い。

6 おわりに

本研究では，Web 上に存在するシラバスを，標準的な形式に変換する手法に関して考察した。実際に変換システムを試作し，提案する変換手法についての評価も行った。比較的簡単な

表 3: 実験データのサイト

No	組織名	最終更新日	URL
01	大阪大学基礎工学部	2003/07/04	http://es6.es.osaka-u.ac.jp/sch/curri/syllabus/mechanical-science.html
02	大阪大学基礎工学部・大学院基礎工学研究科	2001/04/04	http://www-old.es.osaka-u.ac.jp/syllabus/gs13/ims.ms/
03	大阪大学基礎工学部・大学院基礎工学研究科	2002/04/30	http://www-old.es.osaka-u.ac.jp/syllabus/s14/be.htm
04	神戸大学国際文化学部	2002/04/12	http://ccs.cla.kobe-u.ac.jp/System/syllabus.html
05	神戸大学理学部化学科	2003/06/30	http://www.chem.sci.kobe-u.ac.jp/syllabus/
06	神戸大学大学院経済学研究科	2003/07/11	http://www.econ.kobe-u.ac.jp/sirabas/siragk/03gk1
07	広島大学高等教育研究センター	(なし)	http://rihe.hiroshima-u.ac.jp/
08	広島大学理学部	(なし)	http://eduinfo.sci.hiroshima-u.ac.jp/syllabus/science/H9/
09	広島大学理学部物理学科	2002/03/21	http://siss.hiroshima-u.ac.jp/japanese/2002/sci/64012.html
10	九州大学全学教育	(なし)	http://hesvr.rche.kyushu-u.ac.jp/syllabus/sbdepart.cgi?p=2000.0&i=00

表 5: レコード部分のマッチング

データ No.	フィールド数	プログラムがマッチした項目		
		A	$F1$	$F2$
01	12 × 13 = 165	117	39	0
02	14 × 10 = 150	137	3	0
03	14 × 11 = 154	132	0	22
04	15 × 10 = 150	140	0	10
05	11 × 11 = 121	119	0	2
06	16 × 11 = 176	143	0	33
07	13 × 9 = 117	103	14	0
08	18 × 10 = 180	150	0	30
09	12 × 11 = 132	132	0	0
10	9 × 10 = 90	90	0	0
合計	1416	1263	56	97
		正確率：89.19%		

A ：正しくマッチした数
$F1$：E の要素と実データが意味的に正しく合っていない
$F2$：実データ (R) に存在するが，E の要素とマッチしなかった

手法で，約 90%の正確率で変換することが可能であることがわかった。

今後の予定は，まずより多くのデータに対し本手法を適用し，実践的な評価を行う予定である。また，シラバス統合システム全体を構築し，様々な用途に提供していく予定である。

参考文献

[1] A. Arasu and H. Garcia-Molina : "*Extracting Structured Data from Web Pages*," Proc. of ACM SIGMOD/PODS 2003 Conf., pp.337-348, 2003.

[2] 情報処理振興事業協会, 先端学習基盤協会 : "*e* ラーニング白書 *2002/2003* 年版)", オーム社, 2002. (ISBN4-274-06480-8)

[3] 伊東栄典, 山田信太郎, 松永吉広, 廣川佐千男 : "国内 Web シラバスにおけるレコード抽出に関する一考察," 人工知能学会 研究会資料 SGI-KBS-A202, pp.59-64, Sep., 2002.

[4] 井田正明，宮崎和光，芳鐘冬樹，喜多一 : "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 (2A-6), pp.247-248,2003.

[5] 板井久美, 高須淳宏, 安達淳 : "HTML からの情報抽出と統合", NII Journal, No.6, pp.9-19, 2003.

[6] 小島秀一, 高須淳宏, 安達淳 : "Web ページ群の構造解析とグループ化", NII Journal, No.4, pp.23-35, 2002.

[7] IEEE1484 LTSC : "*IEEE1484 : IEEE Learning Technology Standards Committee*", http://ltsc.ieee.org/

[8] K. Lerman, C. Knoblock and S. Minton : "*Automatic Data Extraction from Lists and Tables in Web Sources*", http://www.cs.waikato.ac.nz/~ml/publications/1999/99SJC-GH-Innovative-apps.pdf

[9] Y. Matsunaga, S. Yamada, E. Ito, S. Hirokawa : "*A Web Syllabus Crawler and its Efficiency Evaluation*", International Symposium on Information Science and Electrical Engineering 2003 (ISEE 2003), pp.565-568, 2003.

[10] 大学評価・学位授与機構 : "Syllabus XML schema Ver.1.0", http://svrrd2.niad.ac.jp/syllabus/10/syllabus10.xsd, 2003.

[11] 教育情報ナショナルセンター : "NICER", http://www.nicer.go.jp/, 2002.

[12] Erhard Rahm and Philip A. Bernstein : "*A survey of approaches to automatic schema matching*", The VLDB Journal 10, pp.334-350, 2001.

[13] 坂本比呂志, 有村博紀 : "Web マイニング", 人工知能学会誌, 特集「テキストマイニング」, Vol.16, No.2, pp.233-238, 2001.

[14] 梅原雅之, 岩沼宏治, 永井宏和 : "事例に基づく HTML 文書から XML 文書への半自動変換– シリーズ型 HTML 文書における類似性の利用 –," 人工知能学会論文誌, Vol.16, No.5, pp.408-416, 2001.

[15] Jiying Wang, Frederick H. Lochovsky : "*Data Extraction and Label Assignment for Web Databases*", Proc.WWW2003, Budapest, Hungary, May., 2003.

[16] 山田信太郎, 松永吉広, 伊東栄典, 廣川佐千男 : "*Web* シラバス情報収集エージェントの試作," 電子情報通信学会和文論文誌 D-II, Vol.J86, No.8, pp.566-574, 2003.