国内 Web シラバスにおけるレコード抽出に関する一考察

伊東栄典 † 山田信太郎 * 松永吉広 * 廣川佐千男 †

{itou@cc, yamashin@matu.cc, matsunaga@matu.cc, hirokawa@cc}.kyushu-u.ac.jp

† 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

概要

教育の情報化が進むにつれ，講義内容を紹介するシラバス情報を Web ページとして提示する教育組織が増えている。本研究では，各組織が独自に公開している Web 上のシラバス情報の抽出・統合を行い，ある分野に関する知識を獲得するシステムの実現を目指している。そのためには，シラバスページ収集，レコード抽出，知識提示といった機能を実現する必要がある。本稿では，国内の Web シラバスページから，シラバスの具体的レコードを抽出する方法について考察した。

同一サイトのページは同一の構造で書かれている事が多く，共通部分 (テンプレート) とそれぞれの科目ごとに異なる可変部分に分ける事ができる。HTML で記述されたページに共通に出現するタグの並びを抽出することで，テンプレートの抽出を行なった。また，そこからレコードおよびフィールドを抽出を行う方法を考案し，実装した。

A study of record extraction from Japanese syllabus web pages

Eisuke Itoh† Shintaro Yamada* Yoshihiro Matsunaga* Sachio Hirokawa†

{itou@cc, yamashin@matu.cc, matsunaga@matu.cc, hirokawa@cc}.kyushu-u.ac.jp

†Computing and Communications Center,

∗ Graduate School of Information Science and Electrical Engineering,

Kyushu University.

Hakozaki 6-10-1, Higashi-ku, Fukuoka, 812-8581, Japan.

Abstraction

E-Learning is becomming an important current issue in pedagogical field. A lot of syllabus pages are available as web pages in many educational organizations. The authors consider these syllabli as resources for Web Mining. The syllabus has an abstract information of each lecture. By integrating syllabli of a specific field, they can be a knowledge resource for sellecting good textbooks and for comparison of feature of educational organizations. Integration of syllabus data requires the following phases: (1) the construction of meta data scheme, (2) the collection of syllabus pages, (3) the extraction of records and fields from syllabus pages, and (4) the knowledge extraction. In this paper, we study the third phase.

1 はじめに

情報技術の発達と情報通信基盤の普及に伴い教育の電子化も進展している [5]。国内でも，教材やシラバスといった教育関連情報を電子化し，情報ネットワーク，特に Web を介してネットワーク上に公開する大学等の高等教育機関も増加している。本研究では Web 上に公開されているシラバス情報の収集，抽出および統合を行い，その情報を利用して何らかの知識を提供するシステムの開発を目指している。これにより，単位交換などの大学交流に役立てるための各組織の授業内容の提示および比較や，ある科目についての全国的な講義内容比較，あるいは自分の知らない特定分野 (科目) に関する情報の調査の支援が可能になる。

現在 Web 上に公開されているシラバスページは，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ので，系統的な利用は困難である。一方，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 [11] や，インターネット内に存在する特定テーマに関する情報を収集分類するシステムについての研究 [9] が行なわ

れている。また，Web データを自動収集するクローラーについも，目的に合致したページだけを効率よく収集する研究がある [1]。Web 上のシラバスはその質と量の両面において，Web マイニングの重要な課題である。

Web 上に公開されているシラバスからの知識獲得を行なう方式として，我々は以下のようなフェーズに分解し研究を進めている。

(1) シラバス統合用メタデータの作成
(2) Web からのシラバスページ収集
(3) シラバスページからのレコードおよびフィールド抽出
(4) レコードおよびフィールドの整理統合格納
(5) 格納されたデータからの知識提供

現在までに，(1) および (2) についてはすでに，実験及び考察を行なってきた [13, 14]。本論文では，(3) のシラバスページからのレコードおよびフィールド抽出について考察する。また，実際にレコード抽出に関して実験した結果についても報告する。

2 シラバス統合用メタデータおよびシラバスページ収集

2.1 シラバス統合用メタデータ

シラバス統合のために，シラバス項目を表現するメタデータを作成し収集の観点から評価を行なった [13, 14]。

公開されているシラバスページは，多くの場合，一つの科目の説明記述は表の形式になっており，その中の個々の内容は，項目名および項目値のペアになっている。しかし各組織ごとに表の構造も項目名の使い方も異なっている。そこで，項目名の差異を吸収するため，同じ意味を表す複数の項目名をある一つの項目名で代表するメタデータを作成した。このデータを**「共通計画表」**と呼ぶ (表 1)。

実際のシラバスページに記述されている項目名・項目値を共通計画表の形式に当てはめることで，シラバスデータの統合利用が可能になり，検索や抽出などに項目名を利用する事ができる。

表 1: 共通計画表

共通項目名	対応項目名
担当教官	担当教官，担当，担当者，教官名，担当教員
授業科目名	授業科目名，授業科目，テーマ，研究主題，講義科目，科目名
概要	概要，内容，授業目的，概要と目標，計画，講義の狙い
教材	教材，教科書，参考図書，テキスト，関連ホームページ
関連科目	関連科目，予備知識，必要知識，受講条件，履修しておくべき科目，先履条件
キーワード	キーワード，キー
授業コード	授業コード，コード番号，ID
授業学期	授業学期，開講学期，学期
単位数	単位数，単位
曜日と時間	日時，開講日
評価方法	評価方法，評価，成績

2.2 シラバスページ収集

シラバスページの収集は，Web 検索システムを利用して幅広くページを集める段階と，ノイズを除去する二段階からなる。

自動収集の一次段階には，Web サーチエンジン Google を使ったキーワード検索と，その結果からリンクを再帰的に辿るページ収集方法を用いた。保存するページは TEXT と HTML(content-type が text または html) に限定している。再帰的にリンクを辿る際，同一サイト内へのリンクを辿るようにしている。これは，一般には一覧表示するリンク集ページと，個々の科目を説明するページが固まって存在するためである。

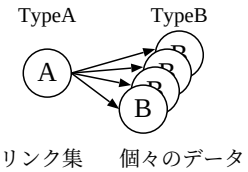

図 1: リンク構造

科目を一覧するリンク集ページを A 型，個々の科目を説明するページを B 型とすると，図 1 に示すリンク構造をもっていることが多い [7]。このことから，A 型もしくは B 型のページを発見できれば，リンク構造を利用してその他のシラバスデータも発見できる。

しかし，この方法で収集したページには，シラバスに関係ないページ (ノイズ) が多数含まれてしまう。そこで，シラバスに関連する単語の出現頻度を用いて，各ページの特徴抽出を行ない，シラバスか否かの自動的に推定する方法についても検討している [14]。

3 レコード抽出

個々の科目のシラバスを記述したページは，一般に図 2 に示す構造を持っている。図 2 のシラバスページは，個々の科目についての情報を格納した，B 型のページである。ここで，一つのファイルを「**ページ**」，一つの科目を説明する部分を「**レコード**」，レコード内の個々の項目を「**フィールド**」と呼ぶ。

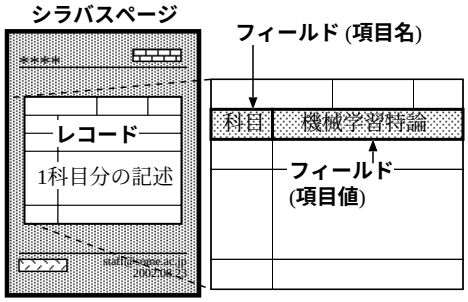

図 2: ページ・レコード・フィールド

一つのページに複数のシラバスが書いてある可能性も考えられるが，具体的な例については，一つのページには一つのシラバスという例が大半であった。その意味で「ページ = レコード」と考えられる。

図 1 で示したように，一つのサイトには複数の B 型のページが存在する場合が多く，かつそれらのページは共通の構造で書かれていることが多い。そこで，同一サイト内にある複数の B 型ページを集め，それらが共通して持つ構造をテンプレートとして抽出し，それを用いてレコードおよびフィールドを抽出する方法を考えた。つまり,「同一サイト内にある B 型シラバスページは，同じ構造を持つ」との仮定に基づいたテンプレート抽出である。なお，ここでは HTML で記述された B 型のページだけを対象としている。

3.1 タグパターンによるテンプレート抽出

テンプレート抽出には，シラバス・ページ群の HTML ファイルの共通的な木構造も考えられるが，本稿ではタグパターンの類似性を用いる。これは，[6, 10] で導入されたもので，抽出したいデータは特徴的なタグ列として表現され，レコードのフィールド部分はタグに挟まれたテキストとして表現されるという見方に基づく。

まず，ページを記述する HTML ソースから，HTML のタグだけに注目し，タグの並びの列 (これを**タグ列**と呼ぶ) を抜き出す。レコード部分が同じ構造で記述されているならば，その構造を作るタグ列の出現頻度が高くなる筈である。そこで，文字列のパターン検索と同様に，複数のページの HTML タグ列から，頻出するタグパターンを検索することで，シラバスページ群の持つ構造 (テンプレート) を抽出する。

ただし，タグパターン検索を行なう対象として，`<H1>,<DL>,<TABLE>`などのページの構造を表すブロックタグだけに注目し，文字飾りなどに用いるインラインタグは無視する。`<B>,<I>,<FONT>`などのインラインタグは，構造ではなく文字飾りを表す場合が多く，構造抽出に適さない場合が多いためである。また開始タグ内に記述されている attribute 部分も無視する。

簡単な例でタグパターンの抽出方法を示す。図 3 に示すようなページ P1,P2 があるとし，その HTML 記述が図 4 だとする。

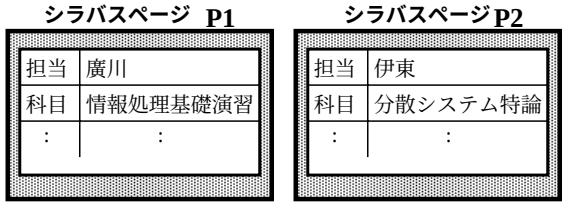

図 3: シラバスページ例

このとき，P1,P2のタグ列はどちらも図5の様になる。この後，複数のページのタグ列が，タグの並びが一致する部分はレコード部分を表すテンプレート構造であろうと推定できる。この例では，完全にタグの並びが同じであるため，タグ列 table.tr.td.*/td.td.*/td./tr.tr.td.*./td.td.*./td./tr./table がテンプレートとして抽出される。

```
<TABLE>
 <TR>
  <TD>担当</TD><TD>廣川</TD>
 </TR>
 <TR>
  <TD>科目</TD><TD>情報基礎演習</TD>
 </TR>
        :
</TABLE>

<TABLE>
 <TR>
  <TD>担当</TD><TD>伊東</TD></TR>
 <TR>
  <TD>科目</TD><TD>分散システム特論</TD>
 </TR>
        :
</TABLE>
```

図 4: HTML ソース例

```
<TABLE><TR><TD></TD><TD></TD></TR>
       <TR><TD></TD><TD></TD></TR>
              :
                                  </TABLE>
```

図 5: タグ列

3.2 フィールドの切り分け

タグパターンに出現するタグは，入れ子構造 (あるいは木構造) になっている。入れ子構造の一番深い部分 (木構造の葉に当たる部分) のタグで囲まれた所に，シラバスの内容となる文字列が存在する場合，このタグで囲まれた部分を一つのフィールドであると判定する。

図 5 の例の場合，<TD>と</TD>で囲まれた部分がフィールドとなる。フィールドの内容を出現順に f_1, f_2, f_3, f_4 とすると，P1 のフィールド値は次のようになる。

f_1: “担当”, f_2: “廣川”,
f_3: “科目”, f_4: “情報基礎演習”

3.3 項目名推定

複数のシラバスページにおいて，n 番目のフィールドの値が全て一致していることがある。これは，そのフィールドがレコードの一つの属性名を表していると考えることができる。逆にページ毎にフィールドの文字列が異なる場合，そのフィールドはレコードの一つの属性値を表していると考えることができる。

図 3 の例で考えると，シラバスページ P1 と P2 において，1 番目のフィールド値はどちらも「担当」であり，3 番目のフィールド値はどちらも「科目」である。そこで，f_1 と f_1 の文字列は，項目名を表しているものと推定できる。これに対し，f_2 と f_4 の文字列はページ毎に異なっているため，f_2 と f_4 は項目値だと推定する。また，共通計画表の文字列がフィールドに出現する場合はシラバスの項目名と推定することも妥当であろう。

4 実験

実験対象となるページを，2001 年 12 月 25 日〜27 日に収集した。「シラバス」を検索語とした Google の検索結果として得られた 649 個の URL を始点にして再帰的に収集を行い，452 サイト，80446 個のファイルを収集した。本稿では，その中の「www.a」で始まる全 20 サイト，4272 個の HTML ファイル (A 型ファイル 241, B 型ファイル 2738) に対しレコード抽出の実験を行なった。

レコード抽出の例として，九州大学大学院総合理工学府先端エネルギー理工学専攻のサイトを示す。このサイトにある A 型のページ (www.aees.kyushu-u.ac.jp/jyugyo01.html) からリンクされるファイルについてタグ列を求めた結果，22 個のファイルについてタグ列が同一であった。その結果，フィールド数が 23 個のテンプレートを抽出した。ページの例と，フィールド (および番号) を図 6 に示す。

22 個のファイルの中で，20%以上 (5 個以上) のファイルで，同じ場所に現れていた共通計画表の項目名と考えられるものを表 2 に示す。1 列

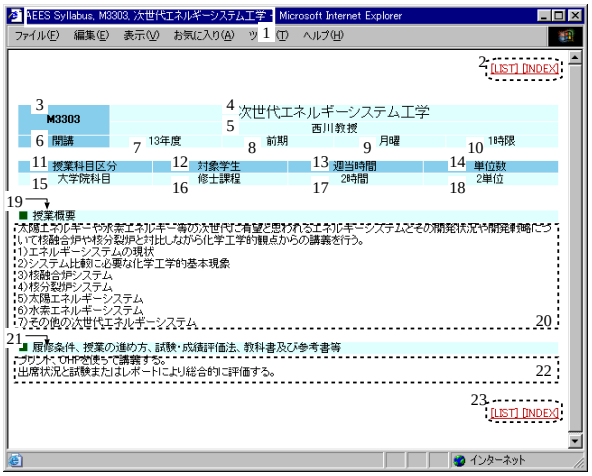

図 6: テンプレート抽出例

目は出現回数，2 列目は何番目のフィールドだったかを表し，3 列目が具体的な項目である。

サイト名と分析結果を表 3 に示す。表 3 で,「総数」欄は実験のために収集したファイル数,「B 型」欄は B 型 (科目を記述したレコードを含むページ) のファイル数,「A 型」欄は A 型ファイル数である。「A*欄」の数は，リンクを辿り集めたファイル群をクラスタリングした結果 5 個以上の要素からなるクラスタがあり，かつ共通テンプレートの同じフィールドに「共通項目名」と考えられる同じ単語がクラスタの要素の 20%以上において現れていたものの個数を表す。一つの A 型から，二つ以上のテンプレートが抽出されたり，抽出されたテンプレートが必ずしも B 型のシラバスページを表しているとは限らない。例えば www.affrs.tuis.ac.jp (東京情報大学) では，一つの A 型からのリンクは，半数がシラバス，半数が教官紹介になっていた。「A 欄」と「A*欄」の差は，A 型に含まれクラスタの要素数が少ない場合，すなわち少数のシラバスしか含んでいなかった場合や，そもそも収集 B 型ファイルが収集できていなかったりする場合であった。

5 関連研究

複数の大学から収集されたシラバス・データを統合するためには，シラバス・データの各フィールド内容の推定精度が高くなければならない。異なるサイトで異なるパターンで記述されたシラバス群に対し，フィールドの内容を推定することにより，同じ内容の部分を統合することができる。フィールド内容を推定し，複数の Web データを統合することは Web マイニングにおいて重要なテーマとなっている。本稿で考察した手法はシラバス以外のデータにも適用可能と考える。各々のフィールド内容を個別には推定できていなくても，フィールド間の従属性に着目して，フィールド内容の推定は可能と考えられる。例えば，[12] におけるデータ従属性を利用したデータベース合成の方法を適用することも検討したい。

本稿では，同一サイトで一つの共通ファイルからリンクされた HTML ファイル群をその大学，学部，学科のシラバス・ファイル候補として，それらの HTMl ファイル群の共通構造の分析とレコード抽出について考察した。これは半

表 2: フィールド抽出例

ファイル数	順番	内容
22	2	[LIST] [INDEX]
22	6	開講
14	7	13 年度
8	7	毎年度
10	8	前期
12	8	後期
6	9	水曜
5	9	火曜
11	10	2 時限
10	10	1 時限
22	11	授業科目区分
22	12	対象学生
22	13	週当時間
22	14	単位数
22	15	大学院科目
22	16	修士課程
22	17	2 時間
21	18	2 単位
22	19	■ 授業概要
22	21	■ 履修条件，授業の進め方，試験・成績評価法，教科書及び参考書等
22	23	[LIST] [INDEX]

表 3: 実験結果

ドメイン	総数	B 型	A 型	A*
www.ads.fukushima-u.ac.jp	504	429	4	2
www.aees.kyushu-u.ac.jp	154	109	3	2
www.affrs.tuis.ac.jp	273	118	74	36
www.age.ne.jp	2	2	0	-
www.agr.kyushu-u.ac.jp	515	493	21	0
www.agr.niigata-u.ac.jp	164	31	13	0
www.aj3.yamanashi.ac.jp	27	2	0	-
www.akeihou-u.ac.jp	114	104	7	3
www.akjim.yamanashi.ac.jp	1	0	0	-
www.ams.osakafu-u.ac.jp	2	2	0	-
www.anan-nct.ac.jp	443	321	10	8
www.anna.iwate-pu.ac.jp	35	3	0	-
www.aomori-akenohoshi.ac.jp	192	191	1	1
www.aomoricgu.ac.jp	228	100	11	4
www.apc.titech.ac.jp	108	0	1	0
www.arc.ynu.ac.jp	3	2	0	-
www.asa.hokkyodai.ac.jp	1148	718	89	34
www.asafas.kyoto-u.ac.jp	354	109	6	2
www.asahi-net.or.jp	5	3	1	0
www.asl.kuee.kyoto-u.ac.jp	1	1	0	-
合計	4273	2738	241	92

構造データの構造類似性の検出 [2] や最頻出パターンの抽出 [8, 3] という一般的な問題の具体例といえる。

6 おわりに

本稿では，自動収集したシラバスページ群から，レコードおよびフィールド，項目を抽出する方法について考察した。また，実際に抽出を行なうプログラムを実装し，収集したデータの一部に対してプログラムを適用し，レコードなどを抽出する実験を行った。本稿で考察した手法を用いると，図 1 の構造をもつページ群を自動的に収集・統合することができる。シラバス以外の例としては，新聞記事やグルメ情報，観光情報などがあげられる。今後はこのような情報を統合利用するシステムも構築したい。

参考文献

[1] S. Chakrabarti, K Punera and M. Subramanyam : "*Accelerated Focused Crawling through Online Relevance Feedback*", Proc. WWW2002, 2002.

[2] I. F. Cruz, S. Borisov, M. A. Marks and T. R. Webb : "*Measuring Structural Similarity Among Web Documents: Preliminary Results*", Springer LNCS 1375, pp.513–524, 1998.

[3] 福田賢治, 石野明, 竹田正幸, 松尾文碩 : "極大共通生垣を用いた情報抽出手法の提案", 情報処理学会研究報告 情報学基礎 66-20, pp.151–158, 2002.

[4] J. Han, J. Pei and Y. Yin : "*Mining Frequent Patterns without Candidate Generation*", Proc. ACM SIGMOD Intl. Conf. Management of Data, pp.1–12, 2000.

[5] 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", オーム社, 2001. (ISBN4-274-064190)

[6] 古賀康則, 田口剛史, 廣川佐千男 : "検索サイト統合のためのラッパー生成法", 第 12 回データ工学ワークショップ (CD-ROM), 2001.

[7] 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.

[8] T. Miyahara, Y. Suzuki, T. Shoudai, T. Uchida, K. Takahashi and H. Ueda : "*Discovery of Frequent Tag Tree Patterns in Semistructured Web Documents*", Springer LNAI 2336, pp.341–355, 2002.

[9] 大槻洋輔, 佐藤理史 : "地域情報ウェブディレクトリの自動編集", 情報処理学会論文誌, 42(9), pp.2310–2318, 2001.

[10] T. Taguchi, Y. Koga and S. Hirokawa : "*Integration of Search Sites of the World Wide Web*", Proc. CUM Vol2, pp.25–32, 2000.

[11] 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.

[12] 高須淳宏, 桂英史, 原正一郎, 相澤彰子 : "データ従属性に基づくデータベースの合成", 学術情報センター紀要, 第 4 号, 1991.

[13] 山田信太郎, 伊東栄典, 廣川佐千男 : "WEB 上に公開されたシラバスからの知識獲得", 情報処理学会第 63 回全国大会 講演論文集 (3), pp.45-46, 2001.

[14] 山田信太郎, 伊東栄典, 廣川佐千男 : "Web 上に公開されたシラバス情報の自動収集", マルチメディア, 分散, 協調とモバイル (DICOMO2002) シンポジウム論文集, pp.137-140, 2002.